RECOGNITION AND BINDING PROCESSES IN HIV-1 PROTEASE

by

Asuman Nevra Özer

B.S., Chemical Engineering, Boğaziçi University, 1997

M.S., Chemical Engineering, Boğaziçi University, 1999

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

Graduate Program in Chemical Engineering

Boğaziçi University

2008

ACKNOWLEDGEMENTS

The work presented in this thesis was conducted in the Polymer Research Center of Boğaziçi University and it was supported by the NIH: AIDS-FIRCA grant RO3TW006875-01.

First of all, I would like to express my gratitude to my thesis supervisor, Prof. Türkan Haliloğlu, for her guidance, friendly encouragement and patience throughout my PhD study.

I am grateful to Prof. Pemra Doruker for her interest and precious comments on my work at our PRC meetings. My sincere thanks also go to my committee members Prof. Işıl Bozma, Assoc. Prof. Işıl Aksan Kurnaz and Assoc. Prof. Uğur Sezerman for the time they devoted to reading and commenting on my thesis.

I feel lucky to have worked with Dr. Celia A. Schiffer for fruitful collaborations. In addition to sharing her valuable knowledge in scientific discussions, I will always appreciate her kindness and hospitality. I also give special thanks to my friends in her group at UMASS Medical School, for making life in Worcester bearable and enjoyable.

Heartfelt thanks are due to all my friends in PRC, including Canan Dedeoğlu, with whom I shared both the most difficult and also the most pleasant times of my everlasting PhD, for their interest and support especially during my crazy thesis writing and defense phases. I would also like to thank Banu, Berna, Elif, Özlem, for being my motivating academic friends, and Çiğdem, Pınar, Tuna, for making sure that they will always be there for me through good or bad.

Finally, this thesis is dedicated to my dear parents and my dear brother, whom I owe everything. I could not make it for today without their unconditional and endless love, support and encouragement.

ABSTRACT

RECOGNITION AND BINDING PROCESSES IN HIV-1 PROTEASE

HIV-1 protease is a drug target against AIDS and understanding its molecular recognition processes is important in development of drugs. Here, the combined computational methodologies used put three different perspectives together to study the recognition and binding processes in HIV-1 protease complex structures. To investigate the substrate specificity, a biased sequence search threading (BSST) technique is introduced. The potential sequence space is efficiently explored by a low resolution knowledge-based scoring function and potential substrate sequences are predicted, which are correlated with the natural substrates. The change in the molecular recognition events, which lead to drug resistance via mutations and/or co-evolution between protease and substrate, is studied by analyzing the collective dynamics of ligand bound protease structures using the Anisotropic Network Model (ANM). The analysis of the dynamic fluctuations imply that substrate and inhibitor complex structures fall into two groups, which differ by the direction of the fluctuations of some mechanistically crucial sites that determine the main rotational axes in the cooperative modes of motion. The network of key interactions within the protease complex structures is also examined by the communication pathways generated using both topological features reflected by the Gaussian Network Model (GNM) and residue-specific interactions estimated by a modeled van der Waals potential. The hinge regions with minimum fluctuation in the most cooperative modes, i.e. dimerization, active site, flap and substrate cleft regions of the protease, act as messengers in the communication. The short pathways between the substrate and protease active site defines the core regions that either function in ligand recognition or interact with the residues that confer drug resistance as the key interacting regions. Moreover, the examination of structural properties of mutant structures indicates a higher correlation of the wild-type complex with the co-evolved

structure than the other mutant structures with respect to both dynamic fluctuations and ensemble of short pathways. Overall, this study adds a further structural and dynamic view to the understanding of the HIV-1 protease system with respect to its interactions to the substrates and drugs, and further to drug resistance.

ÖZET

HIV-1 PROTEAZDA PEPTİT TANIMA VE BAĞLANMA MEKANİZMALARI

HIV-1 proteazın moleküler tanıma mekanizmasının anlaşılması, AIDS ilaçlarının geliştirilmesi için önemlidir. Bu çalışmada kullanılan hesapsal yöntemler, HIV-1 proteaz bileşik yapılarındaki tanıma ve bağlanma süreçlerini üç farklı perspektiften araştırmaktadır. Sübstrat belirginliği incelenmesi için geliştirilen yanlı sekans aramalı giydirme tekniği, düşük rezolüsyonlu bilgi bazlı puanlama fonksiyonu kullanarak sekans uzayını başarılı bir şekilde tarar ve doğal sübstratlarla yüksek ilintili potansiyel sübstrat sekansı tahmini yapar. Mutasyonlar ve/veya proteaz-sübstrat arasındaki eşevrim ile ilaç rezistansına sebep olan tanıma mekanizmasındaki değişiklik, eşyönsüz elastik ağyapı modeli ile bileşik yapıların kollektif hareketleri incelenerek araştırılmıştır. Dinamik dalgalanmalara ve kooperatif hareket modlarında asal dayanak eksenlerini belirleyen önemli bölgelerin dalgalanma yönlerindeki değişikliklere göre, sübstrat ve ilaç bileşik yapılarının ikişer gruba ayrıldığı gözlemlenmiştir. Gaussian ağyapı modelinin yansıttığı topolojik özellikler ve modellenmiş van der Waals potansiyeli ile çıkarılan rezidülere özgü etkileşimler kullanılarak oluşturulan haberleşme yolları ile, bileşik yapılardaki etkileşim ağı incelenmiştir. Haberleşmede etkili bölgelerin, en kooperatif modlarda en düşük frekansta dalgalanan, ikizleşme, aktif bölge, kanat ve sübstrat çevresi gibi dayanak bölgeleri olduğu belirlenmiştir. Sübstrat ve proteaz aktif bölgesi arasındaki en kısa haberleşme yolları, peptit tanınmasında rol oynayan veya ilaç rezistansı gösteren rezidülerle etkileşen çekirdek bölgelerin anahtar etkileşim bölgeleri olduğunu göstermiştir. Ayrıca, mutantların dinamik dalgalanmalar ve kısa haberleşme yolları açısından analizi ile, yaban tipi ve eşevrim geçiren yapılar arasında diğer mutantlara göre daha yüksek ilinti bulunmuştur. Bu çalışma, HIV-1 proteaz sisteminde sübstrat ve ilaçlarla etkileşim ve ilaç rezistansı konusunda yeni yapısal ve dinamik bakış açıları geliştirmektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

c_{ij}	Cross-correlation coefficient between sites i and j
C	Covariance matrix
C^{α}	Alpha Carbon atom
d	Distance
E	Energy
F	Force
h	Heavy side step function
H	Hessian matrix, Shannon entropy
k_B	Boltzmann constant
k_l	Bond stretching constant
k_θ	Angle bending constant
l_i	Bond length between residues i-1 and i
l_i^s	Sidechain bond length between residues i-1 and i
m	Mass
M	Mutual information
N	Number of residues
ns	nanosecond
P	Probability
$P_A(x)$	Probability of finding residue type A at state x
P^0	Background probability
q	Partial atomic charge
r_c	Cutoff separation
R	Gas constant
R_i	Position vector of the i^{th} atom
ΔR_i	Fluctuation of residue i about its mean position
t	Time
T	Temperature
u	Eigenvector of Kirchoff matrix
U	Orthogonal matrix of eigenvectors

V	Potential function
V_n	Bond rotation constant
Å	Angstrom
Δ	Diagonal matrix
ϵ	Well depth
ϵ_o	Dielectric constant
γ	Harmonic force constant, Phase factor
λ	Eigenvalue of Kirchoff matrix
Λ	Diagonal matrix of eigenvalues
π	Pi number
ϕ_i	Torsional angle of bond i
ϕ_i^s	Torsional angle for the sidechain bond i
Φ	Protein conformation
σ	Collision diameter
Γ	Kirchoff (connectivity) matrix
θ_i	Bond angle between backbone bonds i and $i-1$
θ_i^s	Bond angle for the sidechain of residue i
Ψ	Rotational bond angle between N and C atoms in backbone
ω	Torsion angle

A	Ala, Alanine
AIDS	Acquired Immunodeficiency Syndrome
ANM	Anisotropic Network Model
apv	Amprenavir
atv	Atazanavir
BB	Backbone site
BSST	Biased Sequence Search Threading
C	Cys, Cysteine
ca	Capsid
D	Asp, Aspartic acid
drv	Darunavir

DNA	Deoxyribonucleic acid
E	Glu, Glutamic acid
Env	Envelope
F	Phe, Phenylalanine
FDA	Food and Drug Administration
G	Gly, Glycine
Gag	Group antigen
GNM	Gaussian Network Model
HAART	Highly Active Anti-retroviral Therapy
HIV	Human Immunodeficiency Virus
H	His, Histidine
I	Ile, Isoleucine
idv	Indinavir
in	Integrase
K	Lys, Lysine
L	Leu, Leucine
lpv	Lopinavir
LR	Long range
M	Met, Methionine
ma	Matrix
MC	Monte Carlo
MD	Molecular Dynamics
N	Asn, Asparagine
nc	Nucleocapsid
nfv	Nelfinavir
NMR	Nuclear Magnetic Resonance
mRNA	messenger ribonucleic acid
msf	Mean squared fluctuations
P	Pro, Proline
PC	Principal Component
PCA	Principal Component Analysis

PDB	Protein Data Bank
PI	Protease Inhibitor
Pol	Polymerase
pr	Protease
psu	CARB-AD37 inhibitor
psv	CARB-KB45 inhibitor
Q	Gln, Glutamine
R	Arg, Arginine
rh	Ribonuclease H
RMSD	Root mean squared deviation
RMSIP	Root mean square inner product
RNA	Ribonucleic acid
ro1	RO1 inhibitor
rt	Reverse transcriptase
rtv	Ritonavir
S	Ser, Serine
SS	Sidechain and backbone sites
sqv	Saquinavir
SR	Short range
SS	Sidechain site
SU	Surface glycoprotein
T	Thr, Threonine
tf	Transframe
TM	Transmembrane glycoprotein
tmc	TMC-114 inhibitor (Darunavir)
tpv	Tipranavir
W	Trp, Tryptophan
wt	Wild-type
Y	Tyr, Tyrosine
V	Val, Valine

1. INTRODUCTION

1.1. Background and Significance

Acquired Immune Deficiency Syndrome (AIDS) is a set of infections resulting from the damage of the human immunodeficiency virus (HIV) to the human immune system (Weiss, 1993). AIDS is a pandemic now. In 2007, the number of people living with the disease worldwide is estimated to be 33.2 million, and the number of deaths is estimated as 2.1 million people, including 330,000 children (UNAIDS, 2007). Although treatments can slow the course of the disease, no vaccine or cure is currently available. Thus, suppression of viral replication and maintaining it at low levels have become critical objectives in the HIV-1 research field. To this end, highly active antiretroviral therapy (HAART), which is a strategy to improve the length and quality of life of infected individuals, has become successful (Hoggs et al., 1998). Many patients have had complete response to HAART. However, reports of failure, partial response, and/or breakthrough with antiretroviral treatment, have compromised the future of HIV-1 treatment (Scott and Schiffer, 2000).

1.1.1. Human Immunodeficiency Virus (HIV)

HIV was first identified as the agent that causes AIDS in 1983 (Barre-Sinnoussi et al, 1983). HIV-1 is a member of the retrovirus family (Figure 1.1), which are small envelope viruses that contain a diploid, single-stranded RNA genome. The retroviruses are highly prone to mutations. The viral nucleic acids (RNA) and the enzymes required for early replication events (PR and RT) are found in the inner core of the virus particle which is surrounded by capsid proteins. The capsid is surrounded by a lipid membrane and a virus matrix protein is inserted into the inner surface of the membrane. The envelope glycoprotein protrudes through the membrane and forms the outer surface of the virus particle.

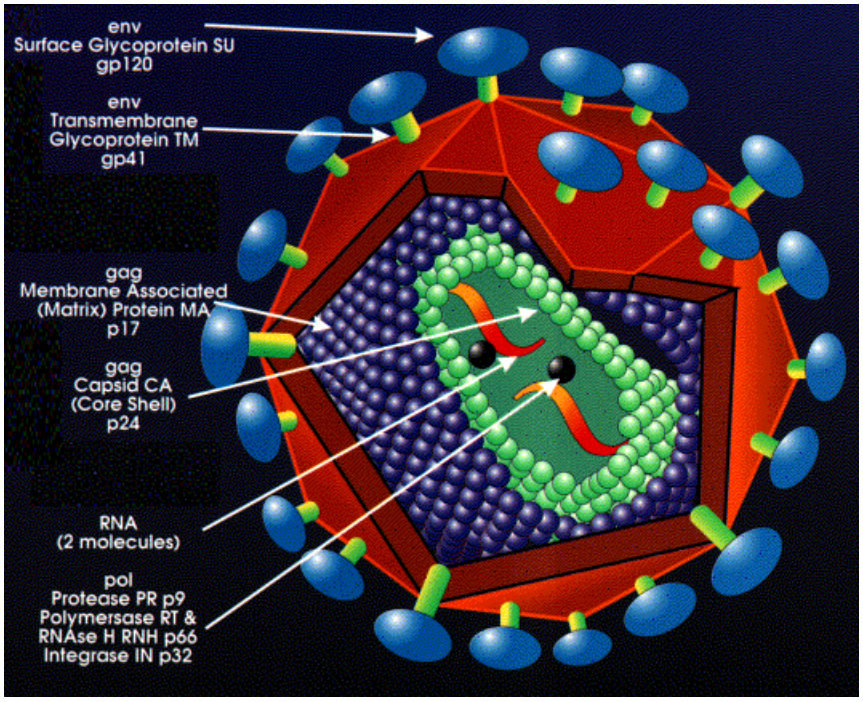


Figure 1.1. The Human Immunodeficiency Virus (HIV)

HIV uses the enzyme reverse transcriptase to make a DNA copy of its RNA genome for replication. A double-stranded DNA intermediate is then produced by a complementary copy of this DNA. The DNA intermediate inserts into the host cell chromosomes. The HIV proviral DNA is then activated and transcribed into HIV genomic RNA and HIV mRNA. The viral mRNA is translated into viral proteins at the host cell's ribosomes. HIV uses a HIV encoded enzyme, namely the protease, in order to cleave a large gag-pol polyprotein and gag polyprotein into functional proteins. These proteins are essential to the structure of HIV and to its RNA packaging. Viral maturation occurs by the binding of the active site of the HIV protease to the polyproteins and cleaving them into functional proteins (Figure 1.2).

The HIV-1 genome has three reading frames: *gag*, *pol*, and *env*, which code for several proteins that are essential for virus assembly and replication. Of these genes, *gag* codes for proteins that make up the viral core, *pol* codes for reverse transcriptase, protease, and integrase, and *env* encodes proteins that make up the viral envelope (Figure 1.3). The reverse transcriptase is found within the virus particle and copies the retroviral RNA sequence into single-stranded DNA when a host cell is infected. A complementary strand of DNA copy of the retroviral genome integrates into the DNA of

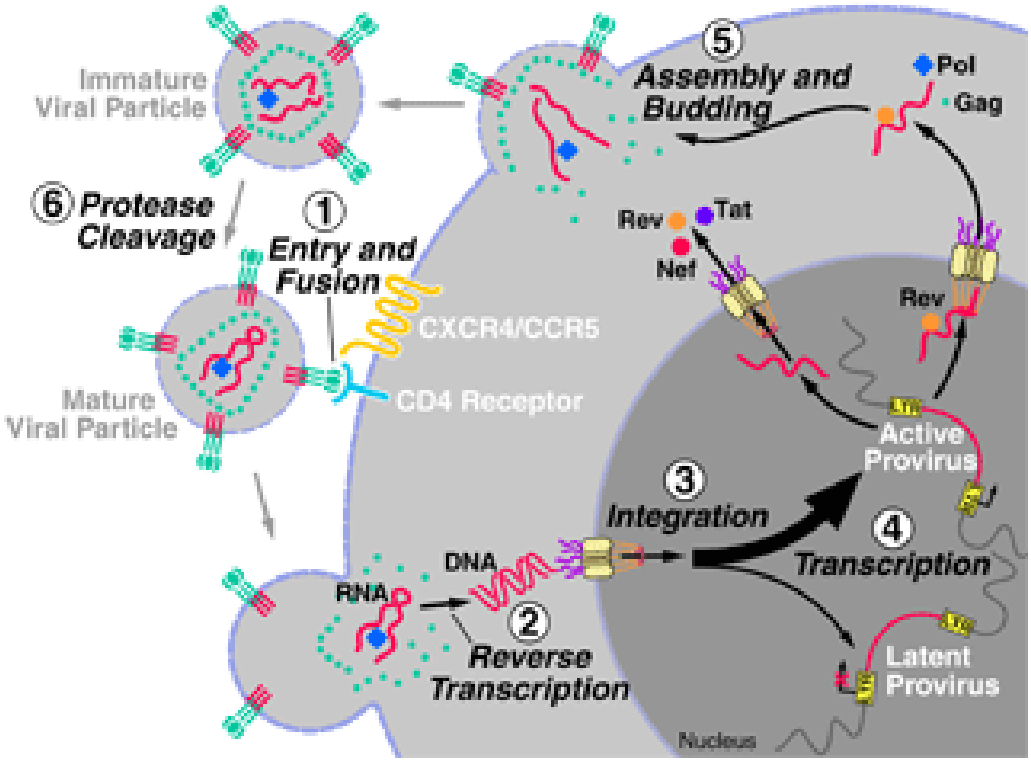


Figure 1.2. Life cycle of HIV

the host cell. The integrated proviral genome can stay in this state for a long time until the development of AIDS symptoms due to destruction of helper T-cells. New copies of the virus are formed by the expression of the retroviral genes by the host cell. The *gag*, *pol* and *env* reading frames are expressed as polyproteins. These polyproteins have to be separated in order eventually for the individual protein molecules to function. HIV proteases then cleave the polyproteins into functional proteins, MA (matrix antigen; p17), CA (capsid antigen; p24), NA (nucleocapsid antigen), PR (protease), RT (reverse transcriptase), and IN (integrase). Likewise, the env gene is transcribed and translated into a polyprotein that is cleaved by proteases into SU (surface glycoprotein; gp120) and TM (transmembrane glycoprotein; gp41).

1.1.2. HIV-1 Protease

HIV-1 protease is essential for the life-cycle of HIV (Weber and Harrison, 1999). The aspartic protease cleaves newly synthesized polyproteins and creates the mature

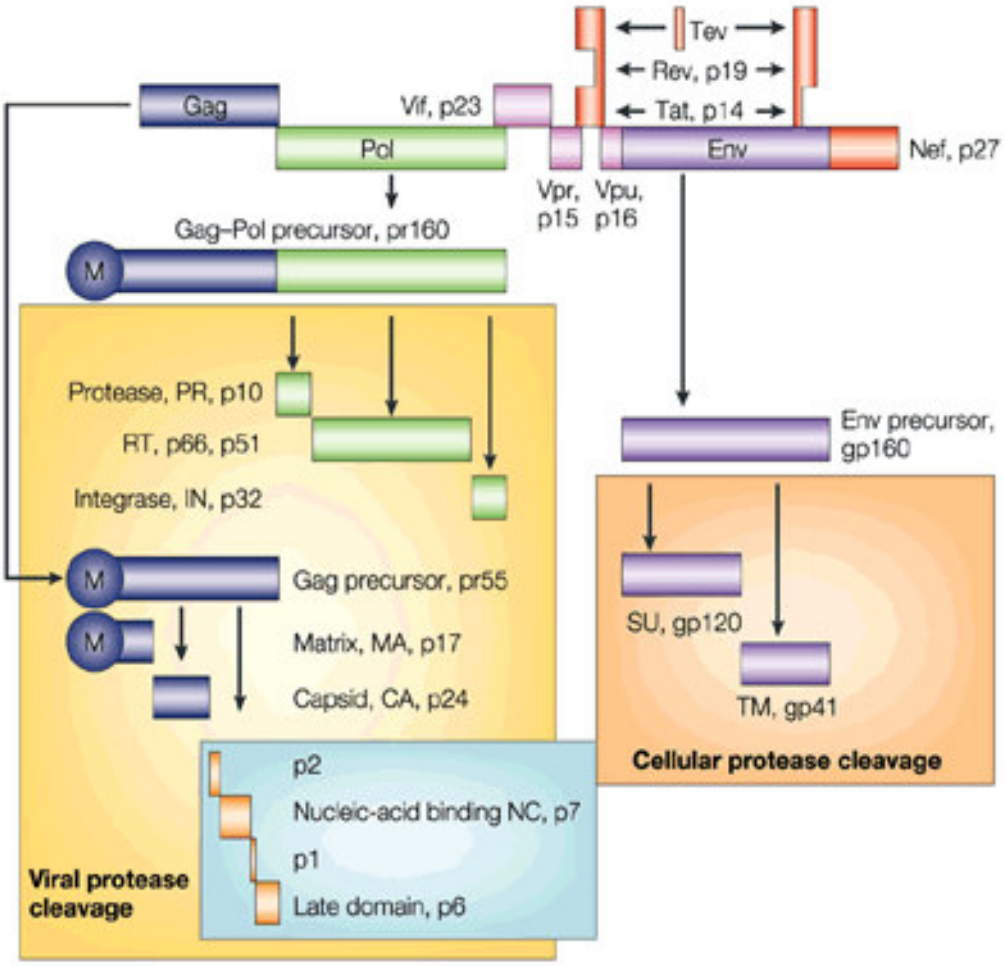


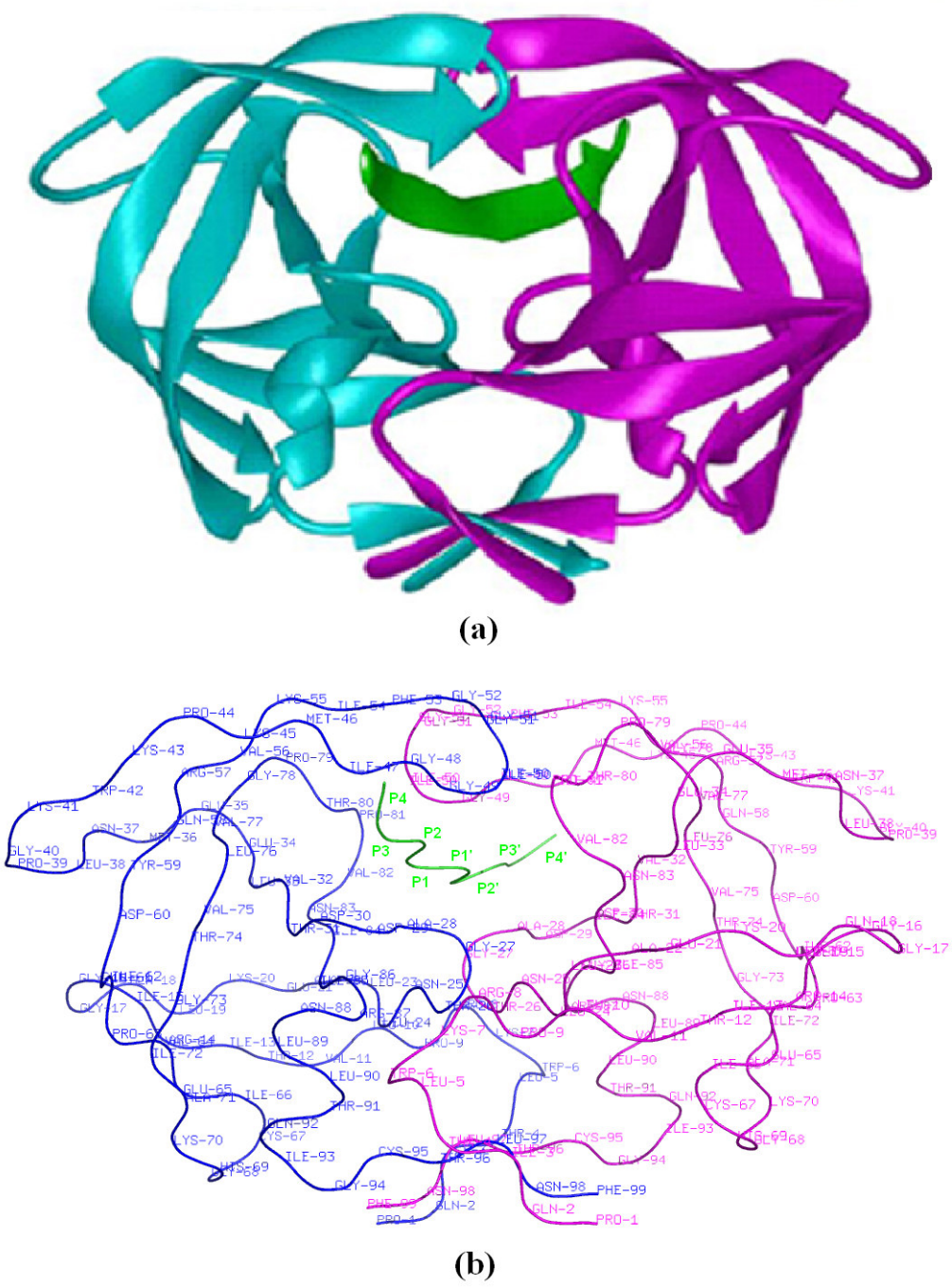
Figure 1.3. The HIV genome. Gag (group antigen; codes for matrix antigen p17, capsid antigen p24, and nucleocapsid antigen); Pol (polymerase; codes for reverse transcriptase, protease, and integrase); Env (envelope; codes for surface glycoprotein gp120 and transmembrane glycoprotein gp41); Tat (transactivating protein; regulates transcription of integrated DNA of HIV); Rev (regulator of viral expression; passage of RNA transcripts out of the nucleus); Nef (negative factor; needed for full pathogenecity of HIV); Vif (viral protein R; aids transport of uncoated nucleoprotein to the nucleus); Vpu (blocks transport of CD4 to the host cell surface).

protein components of an infectious HIV virion. Because of its sensitive and essential function, HIV-1 protease is an excellent target for drug therapy (Goodsell, 2000). The HIV protease exists as a homodimer, with each subunit made up of 99 amino acids (Figure 1.4) and it allows viral maturation by processing the Gag and GagPol polyproteins (Henderson et al., 1988; Chou, 1996). The protease has a single active site which is formed by the dimer interface and capped by two flexible flaps. The active site has the Asp25-Thr26-Gly27 sequence where the two Asp25 residues (one from each chain) act as the catalytic residues (Wlodawer and Erickson, 1993). The flap region includes two solvent-accessible loops (residues 33-43 of each chain) followed by two flexible flaps (residues 44-62 of each chain) and is important for ligand-binding interactions. The terminal region (residues 1-4 and 95-99 of each chain) is important for dimerization and stabilization of the active protease. A large conformational change occurs during ligand binding, which involves the opening and closing of the flaps over the binding site (Yang et al., 2008).

1.1.3. Substrates

The peptide bond hydrolyzed by the protease is referred to as the scissile bond. The hydrolysis of the peptide bond is catalyzed by the conserved D25 residue of the protease by activating a nucleophilic attack by a water molecule on the carbonyl of the scissile amide bond (Moore and Dreyer, 1993). The P1 position is the amino acid immediately upstream of the scissile bond, and the P1' position is the amino acid immediately downstream of the scissile bond (Figure 1.5). Flanking amino acids towards the N-terminus are referred to as P1, P2, P3, P4 and those towards the C-terminus are named P1', P2', P3', P4'. The corresponding pockets in the protease are referred to as S1, S1', S2, S2', etc.

Despite the symmetry conferred on its active site by being a homodimer, HIV-1 protease recognizes ten non-homologous octameric substrate sites (Table 1.1) within the Gag and GagPol polyproteins that are asymmetric. The asymmetry of these substrates in both shape and charge distribution can be observed by their amino acid sequences around the cleavage sites (Prabu-Jeyabalan et al., 2000; Prabu-Jeyabalan et



(a)

(b)

Figure 1.4. Substrate bound HIV-1 protease complex structure.

(a) Cartoon representation, (b) α-Carbon trace with residue labels identified.

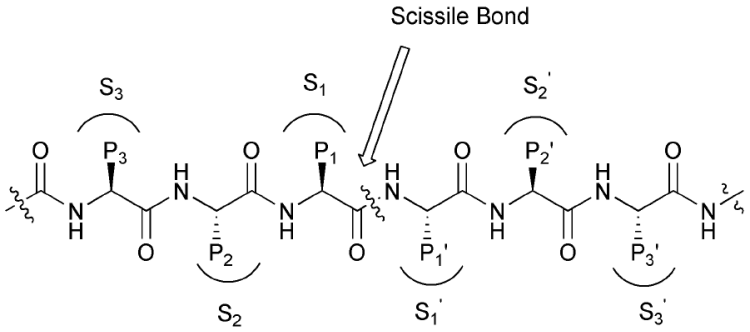


Figure 1.5. The schematic diagram of a substrate bound to protease subsites. The scissile bond is indicated by an arrow.

al., 2002). The crystal structures of complexes of inactive variants of wild type HIV-1 protease with substrates peptides have been determined (Prabu-Jeyabalan et al., 2000; Prabu-Jeyabalan et al., 2002; Prabu-Jeyabalan et al., 2004) (Figure 1.6). Despite the fact that the substrate sites are asymmetric, the currently prescribed inhibitors are relatively symmetric around the cleavage site, permitting a single mutation to impact the inhibitor binding twice, while possibly impacting substrate binding to a lesser extent. Substrate recognition in HIV-1 protease is based on a conserved shape rather than a particular sequence (Prabu-Jeyabalan et al., 2000). This theory implies that there is an interdependency between the different protease substrate subsites in order for a particular sequence to be a substrate. This interdependency likely results from the fact that not all positions within the substrate sites are able to tolerate mutations as can be seen in the variation within the substrate sequences of different subtypes. A systematic study of substrate variation with patient therapy has not been performed as has been done for the protease (Wu et al., 2003). Substrate variation is being investigated both in vivo (Mammano et al., 1998) and in vitro (Lin et al., 2000). However, computational techniques, which utilize the three dimensional structures of the substrate complexes, may be useful to predict which substrate sites are most likely to be susceptible to compensatory mutations.

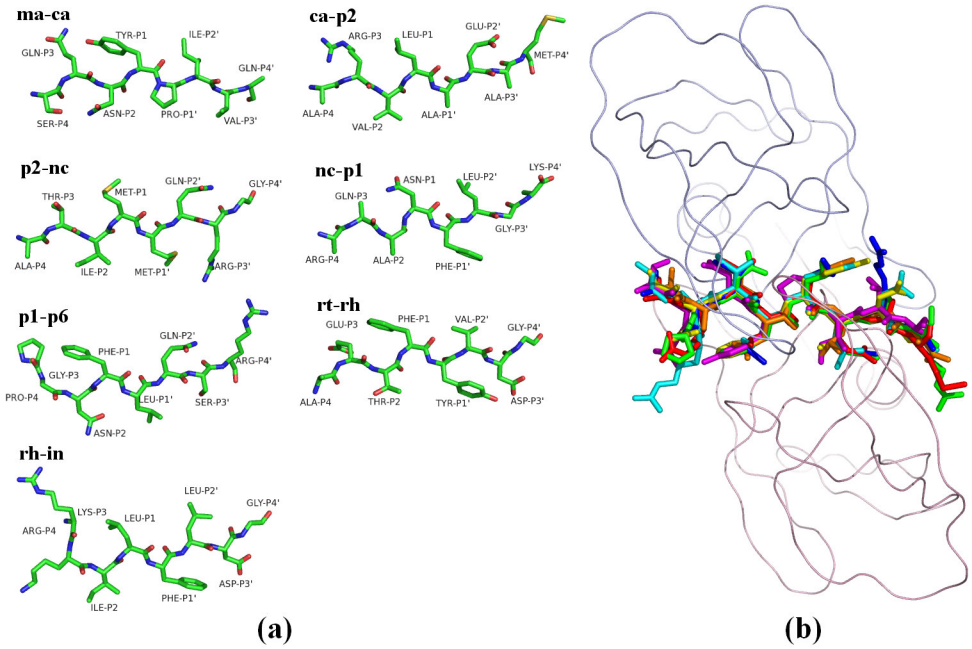


Figure 1.6. Structures of the natural substrates of HIV-1 protease. (a) Conformation of seven natural substrate peptides as observed in complexes with an inactive variant of HIV-1 protease, D25N. The peptides are colored by atom type. (b) Superimposed structures of the natural substrate peptides. The colors of the peptides are: magenta, matrix-capsid; red, ca-p2; blue, p2-nc; orange, nc-p1; green, p1-p6; yellow, rt-rh; and cyan, rh-in. The α-carbon trace of the protease is of the ca-p2 substrate peptide complex. The figures are made with the graphics program PYMOL (Delano, 2002).

Table 1.1. Amino acid sequences of the natural substrate cleavage sites of HIV-1 protease. The natural substrates with available crystal structures are highlighted in bold.

	P4	P3	P2	P1		P1'	P2'	P3'	P4'
Substrate sites in the *Gag* polyprotein									
matrix-capsid	S	Q	N	Y	*	P	I	V	Q
capsid-p2	A	R	V	L	*	A	E	A	M
p2-nucleocapsid	A	T	I	M	*	M	Q	R	G
nucleocapsid-p1	R	Q	A	N	*	F	L	G	K
p1-p6	P	G	N	F	*	L	Q	S	R
Substrate sites in the *Pol* polyprotein									
transframe-protease	S	F	N	F	*	P	Q	I	T
protease-reverse transcriptase	T	L	N	F	*	P	I	S	P
reverse transcriptase-RNaseH	A	E	T	F	*	Y	V	D	G
RNase-integrase	R	K	I	L	*	F	L	D	G

1.1.4. Inhibitors

The development of the HIV-1 protease inhibitors is regarded as a major success of structure-based drug design. Indeed, the protease inhibitors are considered the most potent drugs currently available for the treatment of AIDS (Prabu-Jeyabalan et al., 2002). These drugs are essential components of most HAART therapies (Flexner, 1998). Nine FDA-approved HIV-1 protease inhibitors, indinavir (IDV), nelfinavir (NFV), amprenavir (APV), saquinavir (SQV), ritonavir (RTV), lopinavir (LPV), atazanavir (ATV), tipranavir (TPV) and most recently darunavir (DRV or TMC) are all competitive inhibitors (King et al., 2004), binding at the active site by mimicking the tetrahedral intermediate of its substrate and essentially becoming "stuck", disabling the enzyme. Therefore, they compete directly with the enzyme's ability to recognize substrates (Prabu-Jeyabalan et al., 2000; Prabu-Jeyabalan et al., 2002). The protease inhibitors are peptidomimetics that resulted from structure-based drug design efforts of both academia and the pharmaceutical industry. 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 three-dimensional shape and electrostatic character of these drugs are fairly similar (Figure 1.7), therefore a small set of mutations can result in a protease variant with multi-drug resistance. This evolution of drug resistance in HIV-1 protease presents a new challenge to future structure-based drug design efforts.

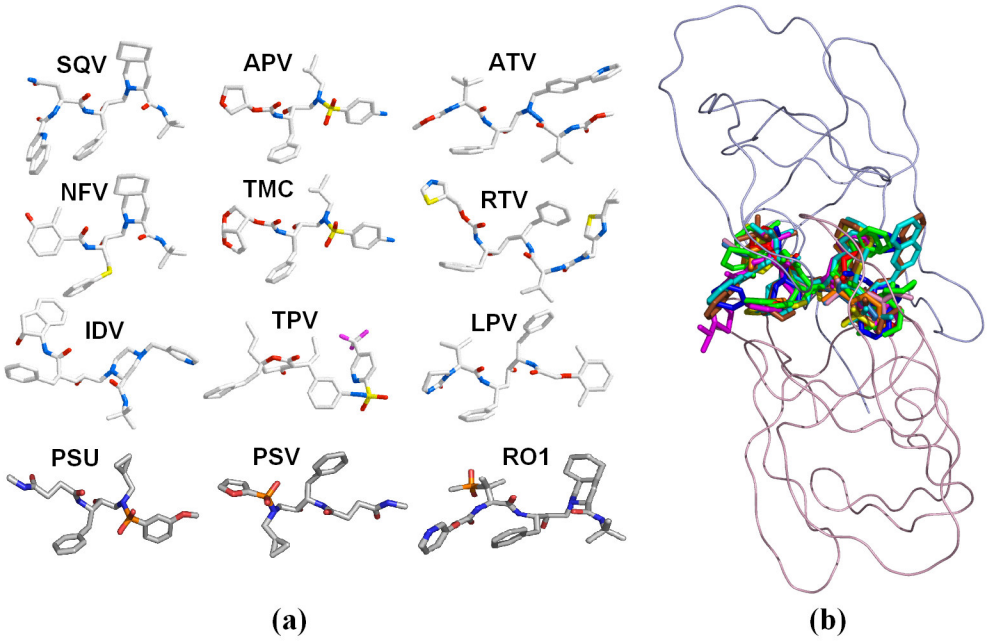


Figure 1.7. Structures of the HIV-1 protease inhibitors.

(a) The HIV-1 protease inhibitors used in this study. The structures are colored by atom type. (b) Superimposed structures of the inhibitors that are used in this study. The colors of the peptides are: red, APV; blue, IDV; green, LPV; yellow, NFV; pink, PSU; skyblue, PSV; brown, RO1; magneta, RTV; cyan, SQV; orange, TMC. The figures are made with the graphics program PYMOL (Delano, 2002).

1.1.5. Drug Resistance and Co-evolution

Because HIV is a retrovirus with a high rate of replication, it exists as a quasispecies or swarm of viral variants in pseudoequilibrium, where potential drug resistant mutants are likely to preexist prior to therapy. Drug resistance is a subtle change in the balance of recognition events from the relative affinity of HIV-1 protease to bind inhibitors to its ability to bind and cleave substrates. Mutations in HIV-1 protease that alter inhibitor binding and cause drug resistance can also affect substrate specificity. The virus will be under selective pressure to co-evolve the substrate sequence, thereby allowing the protease to retain activity (King et al., 2004; Prabu-Jeyabalan et al., 2004; Kolli et al., 2006). Earlier studies have shown that substrate specificity of the protease is based on the shape adopted by the substrate sequences, defined as "the substrate envelope" (Prabu-Jeyabalan et al., 2002; King et al., 2004). 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 they are the ones which co-evolve with the protease (Kolli et al., 2006). An example of this evolutionary communication between substrate (or inhibitor) and enzyme is the apparent co-evolution of the nc-p1 cleavage site with the V82A mutation in the protease (Prabu-Jeyabalan et al., 2004). The most frequently observed case change occurs at P2, where alanine mutates to valine in viral sequences that also contain the V82A drug resistant protease mutation. Selection for a valine at this site makes sense as valine is the wild-type residue at P2 another substrate sequence, the ca-p2 cleavage site. The p1-p6 is another cleavage site that undergoes co-evolution with HIV protease (Kolli et al., 2006). Mutations in the substrate cleavage-site p1-p6 covary with the D30N/N88D protease mutations. Asp30 is important both to binding of NFV and also likely to recognition of the p1-p6 cleavage site. Structural analysis shows that both NFV and p1-p6 have atoms that protrude beyond the substrate envelope and contact Asp30 of different monomers respectively. Thus, both the inhibitor and the p1-p6 substrate are likely to be affected by D30N mutation. This likely explains the particular co-evolution of the p1-p6 cleavage site with the D30N-resistant mutation and also why no other co-evolution with any of the other substrates occurs (Kolli et al., 2006).

1.2. Plan of Attack

HIV protease has been a key and effective target in the treatment of patients infected with HIV. Understanding the subtle balance of molecular recognition events that confer drug resistance in HIV-1 protease is crucial to the development of next generation of drugs for the treatment of HIV infection. Besides experimental efforts, computational means are essential for the rational design of new drugs. The complexity and the multi-disciplinary nature of drug resistance require that evolutionary routes, the dynamics of the target protein and many other aspects are to be addressed for drug design. The computational search for mutational plasticity in the HIV-1 protease complex structures by the analysis of energetic interactions, investigation of dynamic fluctuations of residues, and identification of the pathways of communication as a network of interacting residues in the structure should add a further structural view to the understanding of the drug resistance behavior of HIV-1 protease. The study in this thesis might contribute both to the overall understanding of the plasticity of the ensemble of HIV-1 protease conformations and sequences and to the technology of drug design. The findings might have potential applications in protein engineering, rational protein design, and structure-based drug discovery.

1.2.1. Substrate Specificity by a Biased Sequence Search

A protein's behavior is a function of its sequence within a defined environment. The main purpose of any computational approach to protein design or structure prediction is to solve the problem of determining the fitness (effective energy) of a particular sequence in a given conformation or state. Two conflicting requirements for the energy function used are physical accuracy and computational efficiency. The size of the sequence space compatible with a given protein fold is very large. Nevertheless, it is still small compared with the full space of a protein sequence, whose size is 20^N, where N is the number of residues of the protein. The protein topology appears to be determinant (Koehl and Levitt, 1999a; Koehl and Levitt, 1999b) for the space of the allowed sequences in a given fold. There are two major classes of fitness functions: statistical effective energy functions (Lazaridis and Karplus, 2000) and physical effective energy

functions. Statistical potentials are derived from databases of proteins with known structures (Russ and Ranganathan, 2002). The advantage of these potentials lies in their computational efficiency, mathematical simplicity, and their ability to implicitly capture effects such as desolvation, loss of entropy, and the hydrophobic effect, which are hard to account for explicitly. The disadvantage of statistical potentials is that the accuracy and physical interpretability are compromised. Physical based potentials use atomic-level representations to capture underlying physical phenomena and approximate the free energy of the studied system (Lazaridis and Karplus, 2000; Gordon et al., 1999; Pokala and Handel, 2001). The advantage of these potentials is that they have the potential to provide a more comprehensive understanding of the observed phenomena; however they are computationally more expensive. An optimal energy function would have the simplicity and computational efficiency offered by statistical potentials while retaining the theoretical rigor and physical interpretability of physical based potentials.

A novel threading approach based on knowledge-based potentials to search for the substrate specificity and sequence volume that has fitness to wild type HIV-1 protease is developed in this work. A biased sequence search threading (BSST) methodology is introduced, in which both short-range and distance-dependent knowledge-based potentials are employed to score the sequences threaded on the template structures of substrate bound HIV-1 protease. This technique should probe the preferences of substrate sites and the interdependence between them and help to suggest which sites within the substrate sequences are more likely to be tolerant to change and which are not. The difficulty is in rigorously testing the large number of possible sequences; 20^8 for an eight residue substrate. However, a biased search utilizing the Metropolis criterion (Metropolis et al., 1953) can focus the search only on those sequences that are more likely to bind. Different template structures provide a structure space as well as a base for the differences between the behavior of the substrates, which adjust themselves within the consensus volume and which protrude outside of the consensus volume. The assumption is that the patterns that account for specificity are encoded within the particular conformations adopted by the HIV-1 protease and its interactions with its substrates. To be effective, this methodology must efficiently, via a

biased search with a low-resolution knowledge-based scoring function, explore the potential substrate sequence space. Determining which substrates are less adaptable, by verifying the technique's ability to predict sequence variability, will help elucidate the plasticity of the active site and may be useful in future inhibitor design.

1.2.2. Dynamic Fluctuations

Fluctuations of biomolecular complexes around their native states are important for functional analysis in molecular biophysics. Several features such as entropy changes upon binding, possible drug binding sites, or the overall stability and function can be deduced from the detailed analysis of these fluctuations (Keskin et al., 2002; Micheletti et al., 2002). There exists significant correlation between cooperative motions of the structure and its biological function (Bahar et al., 1998; Bahar, 1999). There are several computational methods that could be used to identify these dominant correlated motions. The common approach is to decompose the dynamics into a collection of modes of motion focusing on a few low frequency/large amplitude modes which are expected to be relevant to function (Bahar et al., 1998; Bahar, 1999). The process of extracting the dominant collective modes from fluctuations in molecular dynamics (MD) trajectories, also called principal component analysis (PCA), is now an established computational method of studying protein dynamics. The major disadvantage of this method is the sampling inefficiency of MD simulations, especially in large size molecular systems (Doruker et al., 2000). Also, MD force fields are not optimal for low-energy fluctuations around the native state (Hamacher and McCammon, 2006). Alternatively, there are a variety of studies where the cooperative motions could be studied by normal mode analysis (Cui and Bahar, 2005). Recently, elastic network models have gained considerable attention in studying the large scale motion of protein structures (Chennubhotla, 2005). Among these approaches, the Gaussian (Bahar et al., 1997a; Haliloglu et al., 1997) and the anisotropic (Atilgan et al., 2001) network models (GNM and ANM) applied to the HIV-1 protease system have produced results that are highly in accord with those of both experimental and MD simulations, despite their simplicity (Bahar et al., 1998; Kurt et al., 2003b; Micheletti et al., 2004; Perryman et al., 2004; Trylska et al., 2007; Yang et al., 2008).

Complex mutational patterns are required for HIV-1 protease inhibitor resistance and structural factors appear to be responsible for the covariation among many of the protease residues (Wu et al., 2003). To this end, the structural fluctuations of HIV-1 protease in interaction with the substrates and the inhibitors should be elaborated to enhance the understanding of the dynamics of HIV-1 protease in relation to its function. This is also to predict possible binding sites for allosteric inhibitors to regulate HIV-1 protease dynamics and aid in the evasion of the drug resistance that continually develops (Perryman et al., 2004).

The crystal structures of both substrate and inhibitor liganded protease are analyzed by the Anisotropic Network Model (ANM) (Atilgan et al., 2001) in this work. Additional to the wild type protese complex structures, the mutant and coevolved HIV-1 protease-subtrate structures are included in the analysis. The elastic network model here is constructed by incorporating all the atoms of the structure for the dynamic analysis. The size and orientation of motion of protease and peptide positions, emphasizing on specific regions of protease such as dimerization, active site, flap and substrate cleft regions, are elaborated by comparative analysis between different natural substrate and inhibitor complex structures. This analysis together with the examination of the structural and dynamic properties of wild-type, mutant and co-evolved structures could contribute to the understanding of the binding as well as the drug resistance mechanism of HIV-1 protease.

1.2.3. Pathways of Communication

Protein topology has been shown (Mirny and Shakhnovich, 2001; Levy et al., 2004) to play an important role in the determination of protein function and folding kinetics. A key feature of many complex systems is their robustness, which is expected to be embedded in the protein topology. Proteins evolve toward a robust design that can tolerate mutations and environmental changes. On the other hand, they are vulnerable to perturbations at key positions or to drastic changes in the environment. If protein structures are information processing networks, mutations of amino acids that are crucial for network communications are expected to impair function. Information

communicated through these networks can be transmitted in a physical (or chemical) form. The residues that are presumed to receive and propagate the information should be central in the interaction network, lying on the shortest pathways (i.e., an ensemble of shorthest pathways) between most residue pairs in the protein. A number of theoretical results have suggested the crucial role of central residues in protein network communication. Examples include, the role of highly connected amino acids as nucleation centers in protein folding (Vendruscolo et al., 2002), the correlation of the most interconnected residues at protein-protein interfaces with residues that contribute the most to binding free energy (del Sol and O'Meara, 2004), and the role of central active site residues (Amitai et al., 2004; del Sol et al., 2006a) in transmitting information between protein residues.

Since there is a network of interactions within a structure's native topology, which is associated with its function and stability, there should be an extent of correlation between this network of interacting residues and drug induced mutation patterns in the HIV-1 protease complex structure. The communication pathways within the HIV-1 protease complex structure are studied by a new methodology in this work. An ensemble of pathways of communication is generated within complex structure starting from the each position in substrate sequence in a given template structure. In generation of the pathways for the scoring of the interaction between the two residues, two approaches are utilized: The intensity of the interactions based on the residue specific potential functions and the coupling between the fluctuations predicted by the Gaussian Network Model (Bahar et al., 1997a; Haliloglu et al., 1997), which reflects topological features of the structure with no specificity in interactions. The analysis of the most dominant and shortest pathway(s) generated provides information about the interdependency of substrate recognition by HIV-1 protease and the differences in the behavior of each of the two monomers with respect to these pathways of interactions.

1.2.4. Contribution

Combined computational methodologies used in this thesis puts three different perspectives together to study the recognition and binding processes in HIV-1 protease

complex structures within the paradigm of sequence, structure and dynamics. In the first part, a novel threading approach based on knowledge-based potentials, namely the biased sequence search threading (BSST) is introduced to search for the substrate specificity of HIV-1 protease. In the second part, the structural fluctuations of the ligand bound HIV-1 protease are analyzed to identify the functionally plausible dynamic motion comparatively between substrate and inhibitor complexes. In the third part, a methodology to generate communication pathways within the HIV-1 protease complex structure is developed to identify the residue interactions that are possibly crucial in the binding interactions.

2. MATERIALS AND METHODS

2.1. HIV-1 Protease Structures

The crystal structures of HIV-1 protease in complex with its seven natural substrates and ten inhibitors given in Table 2.1 are used in this study.

Table 2.1. Substrate and inhibitor bound structures of HIV-1 protease

	PDB code	Reference
Substrates		
capsid-p2 (ca-p2)	1F7A	Prabu-Jeyabalan et al., 2000
matrix-capsid (ma-ca)	1KJ4	Prabu-Jeyabalan et al., 2002
nucleocapsid-p1(nc-p1)	1TSU	Prabu-Jeyabalan et al., 2004
p1-p6	1KJF	Prabu-Jeyabalan et al., 2002
p2-nucleocapsid (p2-nc)	1KJ7	Prabu-Jeyabalan et al., 2002
RNase-integrase (rh-in)	1KJH	Prabu-Jeyabalan et al., 2002
rev.trans.-RNaseH (rt-rh)	1KJG	Prabu-Jeyabalan et al., 2002
Inhibitors		
amprenavir (apv)	1HPV	Kim et al., 1995
indinavir (idv)	1HSG	Chen et al., 1994
lopinavir (lpv)	1MUI	Stoll et al., 2002
nelfinavir (nfv)	1OHR	Kaldor et al., 1997
CARB-AD37 (psu)	2PSU	Chellappan et al. 2007
CARB-KB45 (psv)	2PSV	Chellappan et al. 2007
ro1	2F3K	Prabu-Jeyabalan et al., 2006
ritonavir (rtv)	1HXW	Kempf et al., 1995
saquinavir (sqv)	1HXB	Krohn et al., 1991
darunavir (tmc)	1T3R	Surleraux et al., 2005

2.2. Threading

Threading, known as fold recognition, is a method that may be used to suggest a general structure for a new protein (Jones et al, 1992; Jones and Thornton, 1993). In this method, the amino acid sequence is threaded through known three-dimensional structures and the energy of the structure is evaluated based on some form of potentials to score the structure. In designed algorithms, sequences that minimize the potential function are expected to have greatest likelihood of adopting the target structure. A variety of scoring functions have been used for threading (Bryant and Lawrence, 1993; Jernigan and Bahar, 1996; Jones and Thornton, 1996). The scoring functions are in general very simple because of the large number of possibilities to consider, which also reflects the low resolution nature of the problem. Many of the scoring functions used are the potentials of mean force that provide energy of interaction between two residues as a function of their separation. Use of such knowledge-based potentials was demonstrated in several studies (Sipply, 1990; Altuvia et al., 1995; Altuvia et al., 1997; Schueler-Furman et al., 2000).

In this work, both short-range and distance-dependent knowledge-based potentials are employed to score the structures using a conventional threading method in which the sequences are exhaustively sampled by a biased search technique introduced. With this, two features will be tackled, the number of sequences and the scoring function, which makes the search relatively complex for determining the rules of predicting amino acid sequences that will be potentially binding sequences.

2.2.1. Virtual Bond Model

In this simplified model, the backbone of the protein structure is represented by the virtual bond model originally proposed by Flory and collaborators (Flory, 1969). Each residue is represented by two interaction sites, one at its alpha-carbon atom and one at its sidechain centroid. Hence, the sidechain interaction center is selected on the basis of specific properties of the amino acid type (Bahar and Jernigan, 1996). A schematic representation of the model is given in Figure 2.1, where a protein segment

between backbone sites C^{α}_{i-2} and C^{α}_{i+1} is shown.

Accordingly, the conformation of the backbone is defined by a set of $3N-6$ generalized coordinates for a protein with N residues. These are $N-1$ backbone virtual bonds l_i connecting alpha-carbon atoms $i-1$ and i, $N-2$ bond angles θ_i, the angle between $l_i$and l_{i+1}, and $N-3$ bond torsional angles ϕ_i. Sidechain conformation, on the other hand, is conveniently expressed by the set (l^s_i, θ^s_i, ϕ^s_i), l^s_i being the bond length connecting backbone and sidechain interaction site, θ^s_i is the bond angle between l_i and l^s_i, and ϕ^s_i the torsion angle defined by l_{i-1}, l_i and l^s_i.

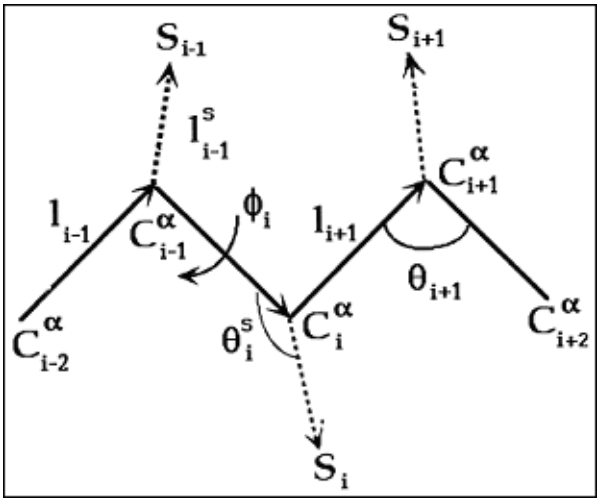


Figure 2.1. Schematic representation of the virtual bond model

2.2.2. Energy of the Protein Conformation

Statistical residue-specific knowledge-based potentials are used to calculate the energy of a given protein conformation with the virtual bond model. The total energy of the peptide is composed of both short and long range interactions: Short range interactions include backbone and side chain conformational energies of the peptide; long range interactions include intermolecular nonbonded interactions of the peptide with the protease and the intramolecular long range interactions of the peptide.

The long-range interaction energy of the peptide is calculated by employing distance-dependent inter-residue potentials (Bahar and Jernigan, 1997). Two effective

interaction sites per residue (its alpha-carbon atom for the backbone and a residue-specific side-chain site) are considered, and the energy of interaction between any two interaction sites are evaluated depending on the distance in-between and the type of amino acid that the sites belong to. The total interaction energy of the peptide is found by summation over all n peptide and N protease residues as

$$E_{LR}(\Phi) = \sum_{i=1}^{n}\sum_{j=1}^{N} E_{SS}(r_{ij}) + \sum_{i=1}^{n}\sum_{j=1}^{N} E_{SB}(r_{ij}) + \sum_{i=1}^{n}\sum_{j=1}^{N} E_{BB}(r_{ij}) + \sum_{i=1}^{n-3}\sum_{j=i+3}^{n} E_{SS}(r_{ij}) + \sum_{i=1}^{n-4}\sum_{j=i+4}^{n} E_{SB}(r_{ij}) + \sum_{i=1}^{n-5}\sum_{j=i+5}^{n} E_{BB}(r_{ij}) \quad (2.1)$$

where r_{ij} is the distance between sites i and j in conformation Φ. The terms account for potentials between side-chain sites (SS); side-chain and backbone sites (SB) and two backbone sites (BB) of residues i and j, respectively. The conformational energy of the backbone is calculated using the statistical potentials extracted from protein structures as based on the virtual bond model given by Bahar et al. (1997a) for bond angle and bond torsions.

$$E_{SR}(\Phi) = \sum_{i=2}^{n-1} E(\theta_i) + \sum_{i=3}^{n-1} \left[E(\phi_i^-)/2 + E(\phi_i^+)/2 + \Delta E(\phi_i^-, \phi_i^+) \right] + \sum_{i=3}^{n-1} \left[\Delta E(\theta_i, \phi_i^-) + \Delta E(\theta_i, \phi_i^+) \right] \quad (2.2)$$

Here, the first summation is the bending of backbone bond angles; the second is the torsion of bonds ϕ_i^- and ϕ_i^+ which are the rotational angles of the virtual backbone bonds preceding and following the i^{th} α-carbon, respectively. Terms are also included for the pairwise interdependence of the torsion and/or bond angle bending. For the side chains, the statistical potentials converted (Kurt et al., 2003) from the probability distributions for packing of side chains in low resolution models (Keskin and Bahar, 1998) are used. The energy associated with a side chain bond angle at state θ_i for a residue type A is

$$E_A(\theta_i) = -RT \ln[P_A(\theta)/P_A^0(\theta)] \quad (2.3)$$

where $P_A(\theta)$ is the statistical probability of finding that bond at angle θ and $\mathrm{P}^0_A(\theta)$ is the background probability assuming a uniform distribution. In the discrete state formalism adopted, the background probabilities are directly proportional to the mesh sizes. Analogous expressions are used for side chain bond lengths and torsions. The side chain conformational energy is summed up over all n side chains in the peptide as

$$E^s_{SR}(\Phi) = \sum_{i-1}^{n} E(l^s_i) + \sum_{i-1}^{n} E(\theta^s_i) + \sum_{i-1}^{n} E(\phi^s_i) \tag{2.4}$$

where $\mathrm{l}^s_i, \theta^s_i$ and ϕ^s_i are the bond length, bond angle and torsion angle of side chain i.

2.3. Elastic Network Models

2.3.1. Gaussian Network Model (GNM)

Gaussian Network Model (GNM) is a recently developed simple analytical model that is used to analyze the vibrational dynamics of globular proteins (Bahar et al., 1997a; Haliloglu et al., 1997). GNM has been applied to a number of different biomolecular systems including RNA complexes (Bahar and Jernigan, 1998), enzyme complexes (Bahar and Jernigan, 1999), substrate-binding proteins (Keskin et al., 2000) and monomeric proteins (Haliloglu et al., 1997), and it has been shown to effectively and efficiently predict X-ray crystallographic temperature factors (Bahar et al., 1997a), H/D exchange behavior (Bahar et al., 1998b) and order parameters from NMR-relaxation measurements (Haliloglu and Bahar, 1999). GNM uses the known topology of protein-protein contacts to model the protein as an elastic network with uniform single-parameter harmonic potentials between the alpha-carbons of residue pairs in contact.

Using GNM, the dynamics of a biomolecular system can be decomposed into a collection of internal modes at different frequencies with a procedure similar to normal mode analysis, yet computationally much more efficient. The slowest modes, with the lowest frequencies, represent the most cooperative motions involving the overall

structure. These dominant modes of motion give information about the molecular dynamics relevant to biological function occurring on a global scale (Amadei et al., 1993; Hinsen, 1998; deGroot et al., 1998; Bahar et al. 1998a). The first modes reflect localized motions involving high-frequency fluctuations of individual residues. These residues are generally in dense regions of the structure with high coordination numbers and they are potentially important for the structural stability.

GNM theory finds its roots in the elasticity theory of random polymer networks (Flory, 1976). In this theory, it is assumed that the native-state protein is equivalent to a three-dimensional elastic network. The junctions in this network are the alpha-carbon atoms, and the interactions between the neighboring residues are represented by harmonic potentials with a uniform spring constant. The residues i and j in the folded protein are assumed to undergo Gaussianly distributed fluctuations ΔR_{ij} about their mean positions in the separation $R_{ij} = |R_j - R_i|$.

According to the GNM theory (Bahar et al., 1997a), the potential energy of the network in terms of ΔR_i, using the harmonic potential approximation, is

$$V = (\gamma/2)\left[\sum_{i,j}^{N} \Gamma_{ij}(\Delta R_j - \Delta R_i)^2\right] = (\gamma/2)\left[\sum_{i,j}^{N} \Gamma_{ij}\langle \Delta R_i \cdot \Delta R_j\rangle\right] \qquad (2.5)$$

Here, the normalization constant γ is the counterpart of the single parameter of the Hookean pairwise potential originally proposed by Tirion (1996) and represents and represents the inter-residue interactions in the native state. The equilibrium correlation between the fluctuations ΔR_i and ΔR_j of residues i and j is given by

$$\langle \Delta R_i \cdot \Delta R_j\rangle = (3k_B T/\gamma)[\Gamma^{-1}]_{ij} \qquad (2.6)$$

where Γ is a symmetric matrix known as the Kirchoff or connectivity matrix, R_i is the position vector of the i^th alpha-carbon, k_B is the Boltzmann constant, T is the absolute temperature. The mean-square fluctuations of individual residues can be readily found from equation 2.6, taking $i = j$.

The elements of the Kirchoff matrix are defined as

$$\Gamma_{ij} = \left\{ \begin{array}{ccccc} -1 & if & i \neq j & and & R_{ij} \leq r_c \\ 0 & if & i \neq j & and & R_{ij} > r_c \\ -\sum_{i \neq j} \Gamma_{ij} & if & i = j & & \end{array} \right\} \tag{2.7}$$

Here r_c is the cutoff separation defining the range of interaction of non-bonded alpha-carbons. A reasonable cutoff distance including all residue pairs within a first interaction shell is 7 $\mathring{A}$ (Bahar and Jernigan, 1997). The i^{th} diagonal element of Γ characterizes the local packing density or the coordination number of residue i.

The inverse of Γ may be written as

$$\Gamma^{-1} = U(\Lambda^{-1})U^T \tag{2.8}$$

where U is an orthogonal matrix whose columns u_i are the eigenvectors of Γ, and Λ is the diagonal matrix of the eigenvalues (λ_i) of Γ, usually organized in ascending order. Mean-square fluctuations of the alpha-carbon atoms and the cross correlations between them are given by the respective diagonal and off-diagonal elements of Γ^{-1}.

It is possible to decompose Γ^{-1} into the sum of contributions from individual modes as

$$\Gamma^{-1} = \sum_{k=2}^{N} \lambda_k^{-1} u_k u_k^T = \sum_{k=2}^{N} A^{(k)} \tag{2.9}$$

Here $A(k)$ is the NxN matrix (for a protein of N residues) describing the contribution of the k^{th} vibrational mode to atomic fluctuations. The first eigenvalues of Γ, identically equal to zero, is not included in the above summation. The i^{th} eigenvalue represents the frequency of the i^{th} mode of motion, and the i^{th} eigenvector gives the shape of this mode as a function of residue index.

2.3.2. Anisotropic Network Model (ANM)

The anisotropic network model (ANM) (Atilgan et al., 2001) performs harmonic vibrational analysis of protein structures around their equilibrium states and predicts the directionalities of the collective motions in addition to their magnitudes. The elastic network is formed by connecting all neighboring atoms and the conformations that describe the fluctuations of residues from the average in the principal directions of motion are generated. The total potential energy for a system of N nodes is the summation over all harmonic interactions of close-neighboring (i, j) pairs. The cutoff distance is chosen as 9 $\mathring{A}$.

ANM is an extension of GNM where the fluctuations are anisotropic (dependent on direction), which incorporates the X, Y and Z components of the position vector, R_i, independently. Therefore, the overall potential for the ANM calculations, includes the fluctuations for all components. The harmonic potential can be expressed as

$$V = (\gamma/2)\left[\sum_{i,j}^{N} h(r_c - R_{ij})(\Delta R_j - \Delta R_i)^2\right] = (\gamma/2)\left[\sum_{i,j}^{N} H_{ij}(\Delta R_j - \Delta R_i)^2\right] \quad (2.10)$$

where γ is the harmonic force constant and R_{ij} is the distance between the sites i and j in the native structure of the protein. $h(x)$ is the heavy side step function which is 1 if $x \geqslant 0$ and zero otherwise.

In ANM, Γ of GNM is replaced by the Hessian matrix H of the second derivative of the intramolecular potential function in equation 2.10. H is a $3Nx3$ symmetric matrix and composed of NxN super elements H_{ij} each of size $3x3$, given by the second derivatives of V with respect to R_i and R_j of C^{α}-atoms of respective i^{th} and j^{th} residues. The correlation between ΔR_i and ΔR_i decomposed into $3N - 6$ modes of motions is then given by

$$\langle \Delta R_i \cdot \Delta R_j \rangle = (3k_B T/\gamma) tr[H^{-1}] = (3k_B T/\gamma) \sum_{k=1}^{N} tr[\lambda_k^{-1} u_k u_k^T]_{ij} \quad (2.11)$$

$tr[H^{-1}]_{ij}$ is the trace of the ij^{th} submatrix $[H^{-1}]_{ij}$ of H^{-1}, that is the sum of the diagonal elements of this $3x3$ matrix. It refers to three different components of ΔR_i and ΔR_j; whereas, when $i = j$, the self correlations between the components ΔR_i are obtained. Here the knowledge of fluctuation vectors permits us to construct and explicitly view pairs of alternative conformations sampled by the individual modes, simply by adding the fluctuation vectors $\pm\Delta R_i$ to the equilibrium position vectors in the respective modes.

2.4. Molecular Dynamics

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

$$\frac{d^2\mathrm{R_i}}{dt^2} = \frac{\mathrm{F_i}}{m_i} \tag{2.12}$$

In equation 2.12, 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.

$$V\left(\mathrm{R}_1,...,\mathrm{R}_N\right) = \sum_{bonds} \frac{k_l}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_\theta}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + cos(nw - \gamma)) + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 R_{ij}}\right) \quad (2.13)$$

Equation 2.13, 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. red(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

$$F_i = - \bigtriangledown V_i(R_1, ..., R_N) = -\frac{\partial V(R_1, ..., R_N)}{\partial R_i} \tag{2.14}$$

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 Equation 2.12.

2.4.2. Simulation Details

The AMBER (Case, 2004; Case, 2005) simulation package with the *ff03* forcefield is used in all the simulations. The protein is 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 are generated from a Maxwellian distribution and the temperature is gradually raised to 300 K. The temperature is maintained at 300 K and the pressure 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) is 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 $\mathring{A}$. A time step of 2 fs is employed in the Leapfrog integrator. The coordinates and energies are recorded every 0.4 ps in the 11 ns simulation, but data is extracted from the full trajectory at every 20 ps instead to decrease computation time. This generates a sample coordinate set of 550 frames. The potential energy, the root mean square deviation (RMSD) and the mean-square fluctuations are investigated for both the full trajectory and the sample set of 550 frames. The good correlation between the two data sets shows the reduced sample set represents the full trajectory, thus, the calculations are performed with the 550 representative frames.

2.4.3. Cluster Analysis

A representative set of conformations is selected among the large amount of conformations generated by MD simulations for subsequent analysis using cluster analysis. The similarity measure to group the MD simulated frames is root mean square devia-

tion (RMSD) in this study.

The cluster analysis procedure requires a similarity matrix in which each element represents the structural difference between a pair of structures. A similarity matrix is constructed by measuring the distance between frames using the root-mean-squared deviation (RMSD):

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}} \tag{2.15}$$

N is number of atoms over which RMSD is measured and d_i is the distance between coordinates of atom i in the two structures, when they are superimposed (Leach, 2001).

MMTSB Toolset's (Feig et al., 2004) *kclust* utility is used to perform conformational clustering. It uses k-means which is a high-performance clustering algorithm; it starts by randomly choosing a collection of frames from the trajectory, each of which is assigned to its own cluster. It is then iterated over all other frames. Each frame is assigned to the cluster whose centroid is closest; the centroid for this cluster is then recomputed. The iterative procedure continues until all frames are assigned to their clusters. The number of clusters is a parameter which depends on the cutoff value of RMSD (cluster radius); as RMSD cutoff increases, less number of clusters are found by the algorithm. The conformations generated from the MD trajectories of HIV-1 protease complex structures are clustered separately in order to group together "redundant" conformations and examine the unique conformers.

2.4.4. Principal Components Analysis (PCA)

The dimensionality of a data set is the number of variables that are used to describe each object. However, there are significant correlations between these variables. To eliminate these correlations, PCA is a commonly used method for reducing the dimensionality of a data set. In general, a principal component is a linear combination of the variables. The first principal component of a data set corresponds to that linear

combination of variables which gives the 'best fit' straight line through the data when they are plotted in the v-dimensional space. More specifically, the first principal component maximizes the variance in the data so that the data have their greatest 'spread' of values along the first principal component. The second and subsequent principal components account for the maximum variance in the data not already accounted for by previous principal components. Each principal component corresponds to an axis in a v-dimensional space, and each principal component is orthogonal to all the other principal components. There can clearly be as many principal components as there are dimensions in the original data, and indeed in order to explain all of the variation in the data it is usually necessary to include all the principal components. However, in many cases only a few principal components may be required to explain a significant proportion of the variation in the data (Leach, 2001).

PCA is performed for the structures taken from MD simulations, using the AMBER software *ptraj* utility (Case, 2004; Case, 2005).

The input is an n by p coordinate matrix, X, where n is the number of snapshots and p is three times the number of atoms. Each row in X represents the atom coordinates of each snapshot structure. The elements of the covariance matrix, C, is calculated as

$$c_{ij} =< x_i - < x_i >> \cdot < x_j - < x_j >> \tag{2.16}$$

where averages are over the n snapshots. The covariance matrix, C, can be decomposed as

$$C = P \Delta P^T \tag{2.17}$$

where the eigenvectors, P, represent the principal components (PCs) and the eigenvalues are the elements of the diagonal matrix, Δ. Each eigenvalue is directly proportional to the variance it captures in its corresponding PC (Alpaydin, 2004).

2.4.4.1. Overlaps between PCs and ANM Modes. The overlap between the motion spaces of the first I PCs and the first J low-frequency modes is defined by the root mean-square inner product (RMSIP) (Amadei et al., 1999) as

$$RMSIP(I,J) = (\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}(P_i \cdot M_j)^2)^{1/2} \tag{2.18}$$

where P_i is the i^{th} PC, and M_j is the j^{th} normal mode. This RMSIP indicates how well the motion space spanned by the first I PCs is represented by the first J modes.

3. SUBSTRATE SPECIFICITY IN HIV-1 PROTEASE BY A BIASED SEQUENCE SEARCH METHOD

A change in specificity of recognition implied by drug resistance may also imply a change in the substrate specificity of HIV-1 protease. Computational threading techniques can be utilized to predict substrate specificity by determining the relationship of the substrate sequence and three-dimensional structure of the protease. The goals here are 1) to ascertain if the energy functions used in the scoring of threaded sequences can identify the natural substrates, 2) to elucidate the roles of the specific residue combinations in interactions with the binding site and 3) to predict the sequences of yet unidentified potential substrates. From this data it may be possible to predict which substrate sequences are more likely to tolerate changes in HIV-1 protease due to drug resistant mutations and which are not.

3.1. Biased Sequence Search Threading (BSST)

A substrate sequence of eight residues - with 20 possible amino acids at each residue - requires an analysis of an ensemble of 20^8 sequences in complete enumeration. To make this large set of potential sequences manageable, a biased search can selectively search only those sequences which have lower energies rather than sampling all sequences. This allows for efficient sampling of a vast number of the potential sequences on various template substrate complexes of HIV-1; and the potential function which incorporated both the short and long range interactions of the peptide with the protease to score the sequences on the template structures. Here, the biasing sequence search technique using the Metropolis criteria is introduced (Ozer et al., 2006) to search towards regions of the sequence space with a higher likelihood of identifying members of the binding sequences (lower energy sequences). The energy window from the minimum for the sampling sequences can be adjusted with the temperature in the Metropolis criterion (Metropolis et al., 1953) that controls the acceptance ratio of the threaded sequences.

The method is performed as follows: Potential octameric substrate sequences, non-binding sequences (Chou, 1996) and some random sequences are threaded through the known protease-substrate complex structures (Table 2.1) and total energy of the peptide is evaluated using the statistical residue-specific knowledge-based potentials given in Section 2.2.1. Then, any one of the residues in the peptide is randomly changed with any of the other nineteen amino acids. Metropolis criterion (Metropolis et al., 1953) determines the acceptance of each sequence by the difference in the energies of the old and the new sequences to find sequences of lower energy. The combination of using different template structures and threading random sequences allows a thorough search in sequence space. To avoid initial sequence induced biased search, the procedure is repeated by starting from different sequences on each of the template structures.

Seven crystal structures of the protease complexed with its natural substrates (ma-ca, ca-p2, p2-nc, nc-p1, p1-p6, rt-rh, rh-in) were utilized to thread the sequences of the nine natural substrates (Table 1.1), nonbinding sequences (Chou, 1996) and random sequences with BSST. The potential functions were previously (Kurt et al., 2003) ranked by the energies of binding and nonbinding peptides based on the potential functions with a variety of threading techniques. The total energy difference was approximately 30 kT within the list of binding peptides. The same energy window, 30 kT, is utilized here for the sequences generated with BSST by Metropolis sampling. The total number of sequences in the pool is 206,847, excluding 30% of the high energy sequences did not alter the results. The preferences of the sites on the substrate for specific amino acids were calculated within these sequences and compared with that of the natural substrates and the cleavable sequences (Chou, 1996) in Chou's database. In these calculations, the probability of specific residues at each site was calculated on independent, pairwise dependent and triplewise dependent preferences of these sites. The calculations were repeated using only the peptide conformational energy and only the non-bonded interaction between the peptide and the protease allowing the relative contribution of each component in substrate recognition to be elucidated.

3.2. Amino Acid Sequence Preference at Particular Sites within The Substrate

To predict the preference of each substrate site for specific amino acids, the distribution of the amino acids sampled at each of the eight peptide positions by BSST was calculated. The preferences were compared with the sequences of natural substrates (Figure 3.1). For most of the peptide positions, the observations are consistent with the sequence variability in the natural substrates. For the P4 position, there is an Ala residue in three of the natural substrates, which is the most probable amino acid in our sequence pool as well. The same case is observed at the cleavage point involving P1 and P1' positions for Phe and Leu, at P2' position for Ile and at P4' position for Gly residues. However, some amino acids which occur often within the nine natural substrates are not in the sequences generated. Examples of these are: Arg at P4 which is a charged amino acid, Gln at P3 and P2', Asn at P2, Ser at P3' which are polar amino acids and Pro at P1'. Thus, some substrate positions, P1, P1' and P2' are predicted well by independent preferences, while others are not.

The independent preferences of the sites on the substrate for specific amino acids show that BSST preferentially selected for natural substrate sequences at a particular site within a substrate. There are only nine natural substrate sequences, yet amino acids found with the highest probabilities at any of the eight positions of these natural substrates are also picked by our method with the highest probabilities. The amino acids flanking the scissile bond are generally hydrophobic (Pettit et al., 2002) and the residues picked in highest probabilities by our method at both of these positions are hydrophobic amino acids such as Phe and Leu. The potential role of the P1 amino acid in regulating the rate of cleavage is explored in Pettit et al.'s work and the requirement of hydrophobic amino acids at P1 position is confirmed (Pettit et al., 2002). Our results with BSST reproduce this, finding Phe, Leu and Met residues, which are all hydrophobic amino acids, at P1 position. For P3, P2, P2' and P3' positions, the amino acids that are not picked by our method, although they are seen in high probabilities within the natural substrates, are generally polar amino acids such as Gln, Asn and Ser. This might be due to the lack of explicit solvent considerations in the energy potentials.

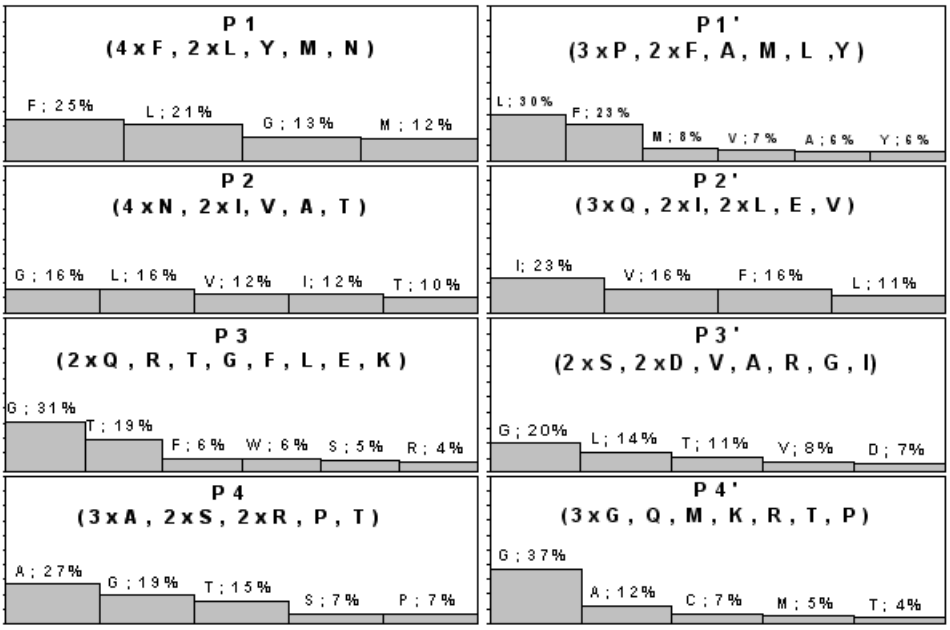


Figure 3.1. Distribution of the selected amino acid residues observed at each of the eight peptide positions (P4-P4') based on the independent preference of each substrate site. The amino acids seen in the nine natural substrates at each site are given in the title parentheses with the number of times each particular amino acid occurs.

3.3. Pairwise Amino Acid Sequence Preference within The Substrate

As BSST selects for particular positions within the natural substrates, the effect of the amino acid preference of substrate sites on each of the other substrate sites was also analyzed. The joint probabilities of amino acids for all possible 28 pairwise combinations of eight substrate positions are calculated and mutual information statistics was used as a measure of covariation between positions within the biased sequences.

3.3.1. Mutual Information Statistics

The average strength of the effective interaction between two variables is the mutual information, a measure of the interdependence between each variable (Crooks

et al., 2004). Mutual information, $M(x, y)$ for the positions x and y is defined as

$$M(x, y) = H(x) + H(y) - H(x, y) \tag{3.1}$$

$H(x)$, $H(y)$ and $H(x, y)$ are the Shannon entropies of positions x and y and the joint entropy of these two positions, respectively.

$$H(x) = -\sum_{i=1}^{m} P(x_i) \log P(x_i) \tag{3.2}$$

$$H(x, y) = -\sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, y_j) \log P(x_i, y_j) \tag{3.3}$$

Here, m and n are the numbers of different amino acids represented at positions x and y, respectively. $P(x_i)$ is the probability of amino acid i at position x, and $P(x_i, y_j)$ is the probability of each combination of amino acids x_i and y_j.

If the amino acids at the two positions vary independently, they will form many combinations and the mutual information will be low. If the positions covary, there will be fewer combinations and mutual information will remain relatively large, as the joint entropy is small (Hoffman et al., 2003).

The mutual information values for the 28 substrate pairs within an eight residue peptide can be seen on Figure 3.2. Most of these covariant pairs are at the positions on the primed side of the cleavage site. However, the P1-P1' pair at the cleavage site also covaries. In the pairs which are two, three or four residues apart, the P4' position is involved in the most prominent covariant pairs. Overall, the pairs with high mutual information values are mostly neighboring positions in either sequence or structure.

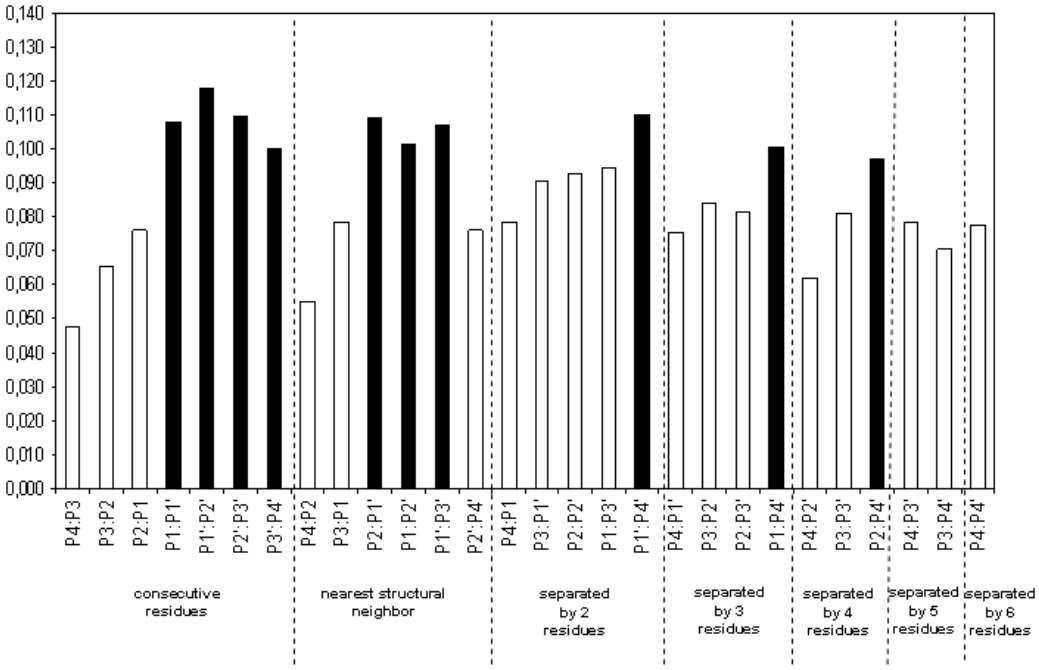


Figure 3.2. Mutual information values of the pairwise substrate interactions. The pairs are grouped in terms of the separation in amino acid sequence. The pairs with high mutual information values, i.e. the covarying pairs are shown in black bars.

3.3.2. Preferences of Pairs

To understand the characteristics of the highly probable pairs, the results were analyzed in more detail. Table 3.1 lists the most probable amino acid pairs in all 28 possible pairs on the peptide sequence. The prominent pairs which occur in any of the nine natural substrates or any of the 62 sequences from Chou's training database of cleavable peptides (Chou, 1996) are indicated in the last two columns. Among the total of 400 possible pairwise combinations of 20 amino acids, most of the pairs generated by BSST reproduce the residue pairs in two or more of the nine natural substrates. The amino acid pairs in which most of the natural substrates are reproduced are P1-P1', P4-P1, P1-P4', P2-P4'. Many pairs are also reproduced in Chou's database. However in three pairs, P3-P2, P3-P2' and P3-P3', none of the natural substrate pairs are found, and P3-P2' and P3-P3' are not reproduced in Chou's database either. Overall, the pairs which reproduce the largest number of natural substrate sequences are those involving the P1 site. The P1-P1' pair reproduces four natural substrates with high probabilities of Phe and Leu residues at each position of the pair. Also, the P1-P4' and P4-P1 pairs reproduce five and four natural pairs respectively. The most probable amino acid pairs except the ones reproducing the natural substrates are Ile-Gly and Gly-Gly at P1-P4' and Gly-Leu and Gly-Phe at P4-P1, which are all hydrophobic amino acids or Gly. Natural substrates involving Gly are picked by our method as well as the predicted pairs involving Gly residues. This overselection of Gly residues is expected as the BSST method searches towards the sequence space of lower energy and the term for the side chain conformational energy in the energy function is zero for Gly. The amino acid preferences at each peptide position in the pairs were also compared with the independent preferences at those positions. The additional amino acids that are picked in the natural substrate pairs although they are not highly probable independently are Ala at P3' on P1-P3', P2-P3', P4-P3' pairs of ca-p2 substrate and Arg at P4' on P1-P4' pair of p1-p6 substrate. Thus, overall, the pairwise prediction more accurately produces the sequence patterns seen in the natural substrates than does the independent prediction.

Table 3.1: The most probable pairs generated with BSST compared with natural substrates and peptides in Chou's database (Chou, 1996)

Pair	Significant pairs generated by BSST		%	Natural substrates reproduced	Number of reproduced peptides in Chou database
P4-P3	A	G	8.6		
	A	T	7.2	p2-nc	1
	G	G	5.7		2
	P	G	3.7	p1-p6	1
	T	T	3.7		
	T	G	3.3		
P3-P2	G	V	7.5		1
	G	L	7.1		
	T	G	5.9		
	G	G	3.7		
	G	I	3.7		
	T	T	3		1
P2-P1	L	F	7.1		1
	V	L	6.2	ca-p2	3
	I	F	3		2
	G	L	2.9		
	G	G	2.6		
	I	L	2.5	rh-in	4
	T	F	2.5	rt-rh	3
P1-P1'	F	L	11.1	p1-p6	1
	L	F	8	rh-in	2
	M	F	5.4		1
	G	L	4.5		
	F	Y	4.4	rt-rh	3

Table 3.1: continued

	L	L	3.7		1
	L	A	3.5	ca-p2	4
P1'-P2'	F	L	7.6	nc-p1, rh-in	3
	L	I	6.3		
	L	F	5.7		
	F	I	5.3		
	Y	V	5	rt-rh	2
	L	G	3.4		
	V	I	3.2		
	F	F	3.2		
P2'-P3'	V	L	6		
	G	G	3.8		
	I	V	3.6	ma-ca	1
	L	L	3.6		
	F	T	3.6		
	I	G	3.5		
	I	T	3.1		
P3'-P4'	L	G	9.2		
	G	G	4.4		1
	V	A	4.3		
	D	G	4.2	rt-rh, rh-in	2
	G	C	3.9		
	N	G	3.7		
P4-P2	A	V	7.6	ca-p2	1
	A	G	5.6		
	G	G	3.1		
	P	L	3.1		1
	G	L	3		
	A	I	2.9	p2-nc	1
P3-P1	G	F	9.1	p1-p6	1

Table 3.1: continued

	G	L	8		
	T	G	4.5		
	G	M	3.5		
	F	F	3.1	tf-pr	2
	T	L	2.8		2
	G	G	2.7		
P2-P1'	L	L	7.3		
	G	L	5.6		
	V	A	4.9	ca-p2	4
	G	F	4.7		
	I	F	3.9	rh-in	2
	I	L	3.8		1
	L	F	3.4		1
	T	L	3.4		1
P1-P2'	F	V	9.2	rt-rh	4
	L	L	5.4	rh-in	2
	L	I	5.2		
	F	F	4.8		
	M	I	3.2		
	F	I	3.2	pr-rt	2
	M	L	3.2		
P1'-P3'	L	G	7.5		
	L	T	6.6		
	F	L	5.6		
	F	D	3.1	rh-in	1
	F	N	2.9		
	Y	L	2.7		1
	M	G	2.6		
P2'-P4'	V	G	9.4	rt-rh	
	I	G	7.5		

Table 3.1: continued

	L	G	6.6	rh-in	1
	I	A	5.5		
	F	G	4.1		
	F	A	3.8		
P4-P1	A	F	7.3	rt-rh	4
	A	L	6.4	ca-p2	2
	G	L	5.2		4
	G	F	4.8		
	P	F	4.7	p1-p6	3
	A	I	4.1		
	A	M	3.2	p2-nc	3
	G	M	3.1		
	A	G	3		
	T	G	2.9		
P3-P1'	G	L	10.8	p1-p6	1
	T	L	8.3		1
	G	F	5.6		
	G	A	4		1
	T	F	3.7		
	G	V	2.5		
	R	F	2.4		
	T	M	2.1	p2-nc	1
P2-P2'	V	I	4.6		
	L	I	4.6		
	L	F	4.6		
	G	I	4.4		
	T	V	3.6	rt-rh	3
	I	I	3.2		
P1-P3'	F	T	6.6		
	F	L	5.1		

Table 3.1: continued

	G	G	4.5		
	L	G	3.9		
	L	L	3.7		
	V	G	3.7		
	L	A	2.5	ca-p2	3
P1'-P4'	F	G	12.8	rh-in	1
	L	G	9.6		
	Y	G	4.1	rt-rh	1
	V	A	3.7		
	L	C	3.3		
P4-P1'	G	F	6.5		1
	A	L	6.4		
	G	L	4.5		1
	P	L	4.5	p1-p6	1
	T	L	4.1		
	A	F	4		
	A	A	3.7	ca-p2	3
	A	Y	3.6	rt-rh	3
P3-P2'	G	I	9.2		
	G	F	7		
	T	I	4.8		
	F	V	3.5		
	G	L	3.3		
	T	G	2.9		
P2-P3'	L	T	4.8		1
	T	G	4.4		
	V	A	3.5	ca-p2	1
	G	G	3.4		
	V	G	2.7		
P1-P4'	F	G	9.9	rt-rh	4

Table 3.1: continued

	L	G	8.3	rh-in	3
	M	G	6.2	p2-nc	1
	I	G	4.2		
	G	G	4		
	W	A	3.8		
	Y	A	3.1		
	L	M	2.6	ca-p2	1
	F	R	2.4	p1-p6	1
P4-P2'	A	I	8.3		1
	A	V	7.3	rt-rh	2
	G	I	4.1		
	G	V	3.7		1
	G	L	3.7		
	A	F	3.7		
	T	I	3.5	pr-rt	1
P3-P3'	T	G	6.2		
	G	T	6.2		
	G	G	5.2		
	G	A	3		
	G	L	2.8		
	G	V	2.6		
P2-P4'	G	G	8.3		1
	I	G	5.4	p2-nc, rh-in	5
	L	G	4.1		
	L	A	4		
	V	M	3.6	ca-p2	1
	T	G	2.9	rt-rh	1
	C	G	2.7		
P4-P3'	T	G	5.3		
	A	L	4.9		

Table 3.1: continued

	G	L	4.2		
	A	G	3.6		
	P	T	2.9		1
	A	A	2.7	ca-p2	2
	T	V	2.6		1
P3-P4'	T	G	9.1	p2-nc	1
	G	G	8.2		
	G	A	3.7		
	W	A	3.6		
	F	G	3.1		
	G	M	3.1		
	T	C	2.8		
P4-P4'	A	G	14.3	p2-nc, rt-rh	2
	G	G	9.1		2
	T	A	3.9		
	T	G	3		
	S	A	3		1
	A	M	2.7	ca-p2	1

Neighboring positions in either sequence or structure vary in an interdependent manner and covariation is observed mostly at the primed side of the substrate sequence. This results from the covariation analysis within the biased sequences and the fewer combination of the pairs with high mutual information values. The pairwise combinations of the sequences generated by BSST reproduced both the natural substrates and the 62 cleavable peptides in Chou's training database taken from experimental data19 although the database contains other proteins than Gag and Pol. The P1-P1', P4-P1, P1-P4', P2-P4' pairs, which reproduce most of the natural substrates as well as a higher number of peptides in Chou's database compared to the other pairs, are probably less tolerant to mutations. On the other hand, the P3-P2 pair, in which none of the natural substrates are reproduced, and the P3-P2' and P3-P3' pairs, in which none of the

natural substrates nor the peptides in Chou's database are reproduced, can probably tolerate mutations. The high probabilities of Phe and Leu residues on the covariant P1-P1' pair, which also reproduce p1-p6 and rh-in substrates, support the requirement of the residues flanking the target scissile bond being generally hydrophobic as stated in Pettit et al.'s work.28 The requirement of hydrophobic amino acids at P1 position is fulfilled within the pairwise combinations except the Tyr and Trp residues, yet Tyr exists in ma-ca substrate. The pairs where most of the natural substrate sequences are reproduced involve the P1 position as well, which has a potential role in cleavage. The overselection of Gly residues is a result of searching towards the sequence space of lower energy as the term for the side chain conformational energy in the energy function is zero for Gly. The energy potentials could be modified to overcome this overselection, yet Gly residues exist in natural substrates. The preferences of each peptide position on the pairwise combinations are almost consistent with the independent preferences of these positions. There are also some amino acids that could be picked in the natural substrate pairs although they are not highly probable independently, such as Ala at P3' on the ca-p2 substrate and Arg at P4' on the p1-p6 substrate. This indicates that in some cases pairwise correlations represent the natural substrates better than independent preferences, that is, coupling is important for these amino acids.

3.4. Triplewise Amino Acid Sequence Preference within The Substrate

As our methodology selected both for particular positions and pairwise combinations within the natural substrates, the triplewise interactions in the low energy sequences generated were also analyzed. The interactions among three variables can be quantified by the triplet mutual information, $M^3(x, y, z)$. This is the average information carried by the triplewise interactions, in excess of the information carried by the pairwise interactions (Crooks et al., 2004). $M^3(x, y, z)$ is defined as

$$M^3(x, y, z) = -H(x) - H(y) - H(z) + H(x, y) + H(x, z) + H(y, z) - H(x, y, z) \quad (3.4)$$

where $H(x, y, z)$, the joint entropy of the three positions x, y and z, is

$$H(x, y, z) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{o} P(x_i, y_j, z_k) \log P(x_i, y_j, z_k) \tag{3.5}$$

Here, o is the number of different amino acids represented at position z and $P(x_i, y_j, z_k)$ is the probability of each combination of amino acids x_i, y_j and z_k.

The 56 triplewise interactions in the sequences generated by BSST were compared with the nine natural substrate sequences, while the total number of triplewise interactions within the 20 amino acids is 8000. Only five natural substrates are repeatedly seen in the triplets: ca-p2, p2-nc, p1-p6, rt-rh and rh-in. The P1-P1'-P4' triplet reproduces four natural substrates and the triplets P4-P2-P4', P4-P1-P1', P4-P1-P4', P4-P1'-P4', P2-P1-P4', P1-P1'-P3' reproduce three natural substrates. However, many other triplets that occur in the natural substrates are not observed; such as P4-P3-P2', P4-P3-P3', P3-P2-P1, P3-P2-P1', P3-P2-P2', P3-P2-P3', P3-P1-P2', P3-P1'-P2', P3-P2'-P3', P3-P2'-P4', P3-P3'-P4', P2-P1-P2', P2-P2'-P3'. Most of the triplets that reproduce two or more of the natural substrates involve at least one of the cleavage positions P1 and P1'. Of these, the predicted triplets have Phe, Met, Leu, Gly at P1 position and Tyr, Leu, Phe, Ala at P1' position as the most probable amino acids. The triplewise preferences of each peptide position were also compared with their independent and pairwise preferences. The amino acid that is picked in the triplets of natural substrates, although not highly probable neither independently nor in the pairs, is Ser at P3' on P3-P1-P3' and P3-P1'-P3' triplets of p1-p6 substrate. Therefore the triplewise preference also adds some additional data reproducing the sequence variability in the natural sequences.

The sequence variability within the natural substrates is represented better by the triplewise preferences of the positions in the peptide sequences generated by BSST than the variability in specific positions and pairwise combinations. The most probable amino acids of the 56 triplets reproduce five of the natural substrates, which are ca-p2, p2-nc, p1-p6, rt-rh and rh-in. The triplets which reproduce three or more natural

substrates and are probably more selective than the others, mostly involve the P1 and P1' positions. These positions are important as they surround the cleavage site and are probably less tolerant to mutations. Of the triplets involving the cleavage site, the highly probable hydrophobic amino acids at P1 and P1' are both consistent with the pairwise preferences and the results of Pettit et al.28 as well. In fact, most of the highest probable amino acids at each independent positions and pairs are also picked within the highest probable triplets. There are also a few amino acids that are picked in the triplets although they are neither as highly probable independently nor significant in pairwise combinations, such as Ser at P3' on the p1-p6 substrate.

3.5. Significance Assessment

The significance of particular pairwise and triplewise associations are assessed by permutation tests (Hoffman et al., 2003), in which 1,000 shuffles are randomly generated to form a reference distribution. As is commonly performed (Korber et al., 1993; Meller and Elber, 2001; Webber and Barton, 2001; Hoffman et al., 2003), the number of sequences in each shuffle is maintained to be the same as in the original sequence pool. $M(x, y)$ and $M^3(x, y, z)$ are recalculated for each shuffle. Then, P values describing the significance of M are calculated as the number of shuffles in which the M value of the permuted shuffle is greater than the M value of the original pool, divided by the total number of shuffles performed (Korber et al., 1993; Hoffman et al., 2003). As large values of M indicate fewer combinations as a result of covariation, getting low M values for random shuffles is an expected result because of many combinations. This strategy for calculating significance is analogous to significance assessment by Z-score which is a measure of the deviation from random distribution, where the distribution with scores that are far from random average value are more significant (Meller and Elber, 2001; Webber and Barton, 2001). For all the pairs and triplets within the sequences generated by BSST, no permuted value of M exceeds the original value, indicating that all original M values are significant with all P values being zero; therefore all the pairwise and triplewise preferences within the substrate sequences are analyzed.

3.6. Prediction of Potential Substrate Sequences

Any prediction method based on statistical theory is composed of an algorithm and a database (Chou, 1996). The database of the prediction method used here is the pool of lower energy sequences generated by the biased search. The techniques can be further utilized to calculate a combined probability for octameric sequences and therefore potential substrates. These sequences can then be compared with the natural substrates to access how accurately these prediction schemes are working. To predict cleavable substrate sequences, the probabilities of octameric sequences are calculated in three ways: using the independent probabilities, pairwise conditional probabilities and triplewise conditional probabilities of the peptide positions within the sequences in this database.

The probability for a specific octameric sequence using the independent probabilities, p of each the peptide positions through P4 to P4' is calculated by

$$P_{octamer} = p(\text{P4}) \cdot p(\text{P3}) \cdot p(\text{P2}) \cdot p(\text{P1}) \cdot p(\text{P1'}) \cdot p(\text{P2'}) \cdot p(\text{P3'}) \cdot p(\text{P4'}) \qquad (3.6)$$

For the probability calculation using the pairwise interdependences, the most convenient approach is via the conditional probabilities, q. The probability for an octameric sequence using the pairwise conditional probabilities of the peptide positions is calculated by

$$\begin{aligned} P_{octamer} = {} & p(\text{P4}) \cdot p(\text{P4,P3}) \cdot p(\text{P3,P2}) \cdot p(\text{P2,P1}) \cdot \\ & p(\text{P1,P1'}) \cdot p(\text{P1',P2'}) \cdot p(\text{P2',P3'}) \cdot p(\text{P3',P4'}) \end{aligned} \qquad (3.7)$$

Similarly, the probability for an octameric sequence using the triplewise condi-

tional probabilities is calculated by

$$P_{octamer} = p(\text{P4}) \cdot p(\text{P3,P2}) \cdot p(\text{P4,P3,P2}) \cdot p(\text{P3,P2,P1}) \cdot$$
$$p(\text{P2,P1,P1'}) \cdot p(\text{P1,P1',P2'}) \cdot p(\text{P1',P2',P3'}) \cdot p(\text{P2',P3',P4'}) \qquad (3.8)$$

The number of residue positions in the top 100 most probable predicted octameric sequences matching the residues of the natural substrates is assessed (Figure 3.3). The number of matching positions as well as the number of natural substrate sequences reproduced is higher in the sequences predicted using the triplewise conditional probabilities than using the pairwise or independent probabilities. The representative sequences predicted using the triplewise conditional probabilities are listed in Table 3.2. These potential substrate sequences reproduce five or more residues within each of five of the natural substrates, namely rt-rh, ca-p2, p1-p6, rh-in and p2-nc. Within these natural substrates, p2-nc, which is the least picked, is the most variable and rt-rh and ca-p2, which are reproduced mostly, are not variable in sequences when comparing the natural variation among the subtypes.

The triplewise preferences of the peptide positions generated by BSST can be further utilized to predict potential substrate sequences. Using the triplewise conditional probabilities to predict the potential substrate sequences produces the most accurate prediction when the sequences are compared with the natural substrates. The predicted potential substrate sequences have five or more residue positions matching with the residue positions of most of the natural substrates. The natural substrate which is the least picked is p2-nc, which is the most variable substrate and rt-rh and ca-p2, which are reproduced mostly, are not variable substrates. This implies that there is a complex interdependence between the different substrate residue positions.

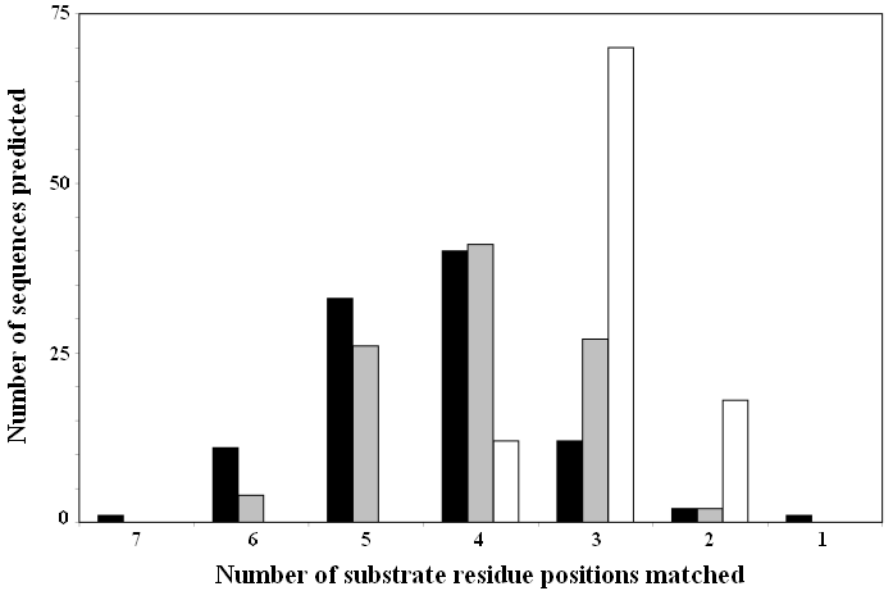


Figure 3.3. Histogram of the number of residues in the predicted octameric sequences that match residues in one of the nine natural substrate sequences. The sequences predicted using the triplewise conditional probabilities reproduced the natural sequences with highest fidelity and they are shown in black bars, the sequences predicted using the pairwise conditional probabilities are shown in gray bars and the sequences predicted using the independent probabilities are shown in open bars.

Table 3.2: Representative sequences within the top 100 sequences predicted using the triplewise conditional probabilities. The residue positions matching natural substrates are highlighted in bold.

Natural substrates	Number of matches	P4	P3	P2	P1	P1'	P2'	P3'	P4'
rt-rh	7	**A**	F	**T**	**F**	**Y**	**V**	**D**	**G**
ca-p2	6	**A**	G	**V**	**L**	**A**	I	**A**	**M**
ca-p2	6	**A**	G	**V**	**L**	**A**	F	**A**	**M**
rt-rh	6	**A**	F	**T**	**F**	**Y**	**V**	L	**G**
ca-p2	6	**A**	G	**V**	**L**	**A**	L	**A**	**M**
rt-rh	6	**A**	H	**T**	**F**	**Y**	**V**	L	**G**
ca-p2	6	**A**	G	**V**	**L**	**A**	H	**A**	**M**
rt-rh	6	**A**	M	**T**	**F**	**Y**	**V**	L	**G**
rt-rh	6	**A**	I	**T**	**F**	**Y**	**V**	L	**G**
rt-rh	6	**A**	G	V	**F**	**Y**	**V**	**D**	**G**
rt-rh	6	**A**	F	C	**F**	**Y**	**V**	**D**	**G**
rt-rh	6	**A**	Q	**T**	**F**	**Y**	**V**	L	**G**
ca-p2	5	**A**	G	**V**	**L**	**A**	I	**A**	Y
ca-p2	5	**A**	G	**V**	**L**	**A**	I	**A**	W
rt-rh	5	**A**	G	V	**F**	**Y**	**V**	L	**G**
rt-rh	5	**A**	F	C	**F**	**Y**	**V**	L	**G**
ca-p2	5	**A**	G	**V**	**L**	**A**	I	G	**M**
rt-rh	5	**A**	F	F	**F**	**Y**	**V**	L	**G**
ca-p2	5	**A**	G	**V**	**L**	**A**	F	**A**	Y
ca-p2	5	**A**	G	**V**	**L**	**A**	I	**A**	G
p1-p6	5	**P**	**G**	L	**F**	**L**	F	T	**R**
ca-p2	5	**A**	G	**V**	**L**	**A**	F	**A**	W
ca-p2	5	G	G	**V**	**L**	**A**	I	**A**	**M**
rt-rh	5	**A**	F	V	**F**	**Y**	**V**	L	**G**
rt-rh	5	**A**	F	S	**F**	**Y**	**V**	L	**G**
rt-rh	5	G	F	**T**	**F**	**Y**	**V**	L	**G**

Table 3.2: continued

ca-p2	5	**A**	G	**V**	**L**	**A**	I	**A**	I
rt-rh	5	**A**	F	**T**	**F**	W	**V**	L	**G**
rh-in	5	A	T	G	**L**	**F**	**L**	**D**	**G**
p1-p6	5	**P**	**G**	L	**F**	**L**	H	T	**R**
ca-p2	5	**A**	G	**V**	**L**	**A**	I	**A**	Q
rh-in	5	A	T	**I**	**L**	**F**	**L**	L	**G**
ca-p2	5	**A**	G	**V**	M	**A**	I	**A**	**M**
p2-nc	5	**A**	**T**	**I**	**M**	F	L	L	**G**
rt-rh	5	**A**	H	C	**F**	**Y**	**V**	L	**G**
ca-p2	5	G	G	**V**	**L**	**A**	F	**A**	**M**
rt-rh	5	T	F	**T**	**F**	**Y**	**V**	L	**G**
rt-rh	5	**A**	G	G	**F**	**Y**	**V**	L	**G**
rt-rh	5	**A**	H	V	**F**	**Y**	**V**	L	**G**
rt-rh	5	**A**	I	C	**F**	**Y**	**V**	L	**G**
rt-rh	5	T	T	**T**	**F**	**Y**	**V**	L	**G**
rt-rh	5	**A**	T	G	**F**	**Y**	**V**	L	**G**
ca-p2	5	**A**	G	**V**	**L**	**A**	I	**A**	V
ca-p2	5	**A**	G	**V**	**L**	**A**	I	**A**	E
rt-rh	5	**A**	M	C	**F**	**Y**	**V**	L	**G**

3.7. Contribution of Peptide Conformational Energy and Peptide-Protease Interaction Energy in Recognition

BSST was performed considering the total energy of the peptide, i.e. Metropolis criterion (Metropolis et al., 1953) was applied to the total energy which included both short and long range interactions. The calculations were then repeated twice after threading the peptide sequences onto the substrate positions. In one approach, the low energy sequences were generated without considering the protease, i.e. Metropolis criterion (Metropolis et al., 1953) was applied to the peptide conformational energy. In the second approach, the low energy sequences were generated only with considering the nonbonded interactions, i.e. Metropolis criterion (Metropolis et al., 1953) was applied to the long range interaction energy. The independent, pairwise and triplewise amino acid preferences of the peptide positions of the sequences generated when the lower energy sequence space was searched according to peptide conformational energy were consistent with the preferences of the sequences generated by BSST according to the total energy of the peptide. Although the highest probable amino acids preferred by the positions were mostly the same in the sequences generated by BSST considering peptide conformational energy, the probability values of the mostly preferred amino acids were lower and more uniform. The top 100 most probable potential cleavable sequences predicted using the triplewise conditional probabilities of these new sequences generated by BSST according to peptide conformational energy had five residues or less matching the residues of the natural substrates. The only natural substrate with five residues reproduced was ca-p2. Moreover, when the calculations were repeated with the sequences generated by BSST according to the energy of nonbonded interactions only, it was not possible to evaluate significant probability values for the amino acid preferences of the peptide positions and to predict cleavable substrate sequences. The long range interactions between the peptide and the protease were not as important as the near-neighbor interactions within the peptide, although the addition of the effect of long range interactions enhanced the recognition. Only when the top 100 most probable cleavable sequences were predicted using the sequences generated by BSST according to total peptide-protease energy were the sequences of five to seven residues of five natural substrates reproduced.

The repeated BSST calculations considering only the conformational energy of the peptide and only the energy due to the interactions between the peptide and the protease suggest the relative roles of each in recognition. Although the amino acid preferences of the peptide positions were similar, far fewer of the natural substrate residues were matched in the potential cleavable sequences predicted when BSST was carried out without considering the protease. The results of BSST utilizing only the energy of nonbonded interactions, was not able to either reproduce preferences of peptide positions or to make any predictions to match substrate sequences. Only utilizing the entire potential with the local protease substrate interactions with BSST was able to successfully reproduce substrate sequences.

4. DYNAMIC FLUCTUATIONS IN HIV-1 PROTEASE

The crystal structures of substrate and inhibitor liganded HIV-1 protease are analyzed by a structure-based method, namely the Anisotropic Network Model (ANM). The crystal structures of the protease in complex with its seven natural substrates and ten inhibitors (Table 2.1) are used. According to the inhibition constants of inhibitors reported previously (Wang and Kollman, 2001; Prabu-Jeyabalan et al., 2006; Chellappan et al., 2007; Hou et al., 2007), tmc and lpv are the strongest inhibitors whereas psu and psv are the weakest inhibitors among the ones used in this study.

The size and orientation of motion of residues in protease and peptide positions are elaborated comparatively between different complex structures. The conformations generated by the ANM are superimposed prior to this analysis. Conformations generated from MD trajectories of seven wild-type HIV-1 protease-natural substrate complex structures, the D30N mutant p1-p6 complex, the D30N-N88D mutant p1-p6 complex, and the D30N/N88D/LP1'F co-evolved p1p-p6 complex structures are also utilized here. A representative set of conformations is selected among the large amount of conformations generated by MD simulations for subsequent analysis using a clustering method. Further, the ANM analysis is performed for the representative members of the largest cluster of each structure. Principal component analysis (PCA) is also applied to molecular dynamics (MD) trajectories of the wild-type natural substrate complex structures, to compare the observed motions and support the structure-based explanation of the results by ANM.

4.1. Principal Motions and Residue Fluctuations

The mechanisms of the cooperative molecular motions relevant to function are implied by the low frequency modes of motion (Bahar, 1999; Liu et al., 2004). The fluctuations in the principal directions refers to the main functional motion of the structure and thus all HIV-1 protease complex structures that are functional should display similar modes of motion. The motion of complex structures in the most cooperative

modes, the slowest first and second modes, is similar for all substrate and inhibitor complex structures. The direction of the fluctuations of residues for the ca-p2 substrate complex is presented as an example in Figures 4.1 and 4.2. The X and Z axes lie along the in-plane directions where X is the longest axis along which the protease lies, and the Y axis lies along the out-of-plane direction where positive Y direction coincides with the direction from N- to C- terminus of the peptide (see the ribbon diagrams in (a) panel). In the first slowest mode (Figure 4.1), both monomers of the protease rotate around two axes parallel to Z direction and the peptide fluctuates in negative Y direction. In the second slowest mode (Figure 4.2), there are two axes around which the monomers rotate, one parallel to X direction and the other parallel to Z direction. The monomers rotate around the X axis in opposite directions and the motion in the substrate is significant in the edges in the second slowest mode.

Besides the crystal structures, the ANM analysis is carried out for the representative structures extracted from the MD simulations. For this, the conformations of 11 ns MD simulations of substrate complex structures are clustered and the best member of the largest cluster of conformations are chosen for the ANM analysis. The modes of motion are highly correlated with those obtained by the ANM of crystal structures, implying that the dynamic fluctuations in the principal modes of motion are not affected by possible crystal contacts.

The distribution of mobilities among residues in the low frequency modes are represented by the mode shapes in Figures 4.3-4.6, where the mean square fluctuations of protease residues in the substrate and inhibitor complex structures in the first two modes are displayed. The fluctuations of protease residues in the inhibitor complex structures are in good correlation with those in the substrate complexes, and the minimum fluctuating residues correspond to the same regions. These minimum fluctuating regions in the most cooperative two modes correspond to residues 5-10 (dimerization region), 25-27 (active site), 45-55 (flap) and 80-90 (substrate cleft). The mobility of the flap region is supressed by binding of peptides. On the other hand, highly mobile regions correspond to residues 12-22, 36-44 and 61-73. Previous studies (Bahar et al., 1997; Bahar et al., 1998a) indicate that the minima in the global mode shapes corre-

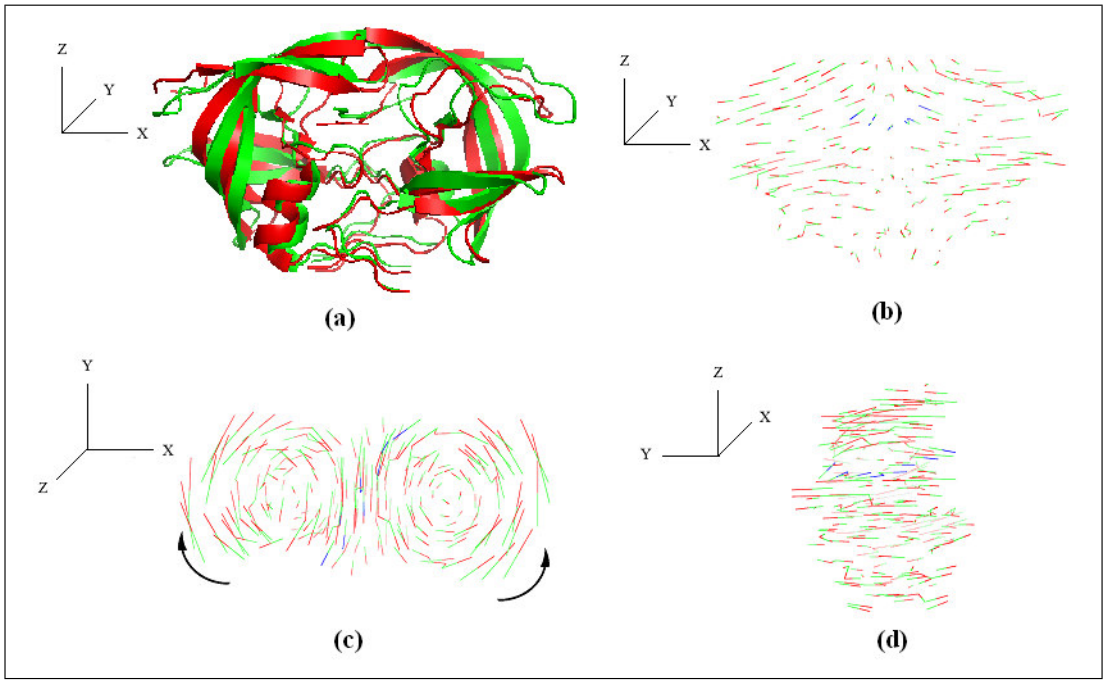


Figure 4.1. Motion of HIV-1 protease complex structures in the first slowest mode. (a) X and Z axes lie along the in-plane directions where X is the longest axis along which the protease lies, and Y axis lies along the out-of-plane direction where positive Y direction coincides with the direction from N- to C- terminus of the peptide. (b),(c),(d) The fluctuations of residues viewed from different directions. The monomers of the protease rotate around two axes parallel to Z direction and the peptide fluctuates in negative Y direction.

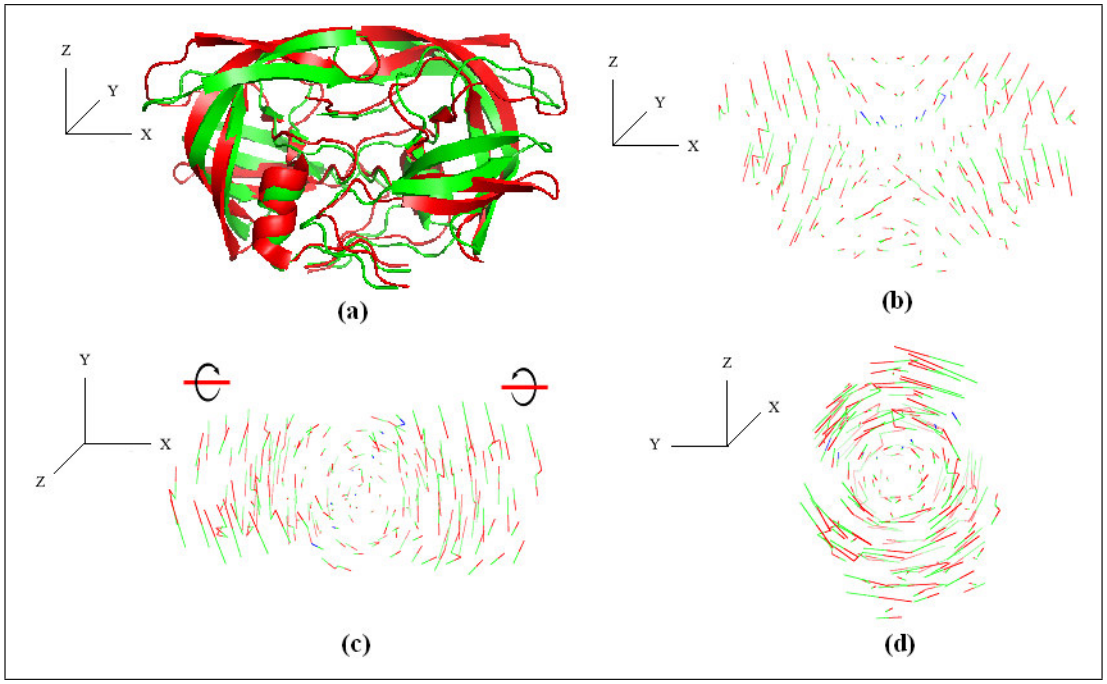


Figure 4.2. Motion of HIV-1 protease complex structures in the second slowest mode. (a) X and Z axes lie along the in-plane directions where X is the longest axis along which the protease lies, and Y axis lies along the out-of-plane direction where positive Y direction coincides with the direction from N- to C- terminus of the peptide. (b),(c),(d) The fluctuations of residues viewed from different directions. The monomers of the protease rotate around two axes, one parallel to X direction and the other parallel to Z direction. The monomers rotate around X axis in opposite directions and the motion in the substrate is significant in the edges.

spond to the regions with restricted motion that may act like hinges of the molecule, while the maxima correspond to substrate recognition regions of highest mobility that sample a large space. The mobile regions as well as residues that are important for protein stability or that take part in the key native contacts have been addressed for HIV-1 protease by previous GNM studies (Bahar et al., 1998a; Kurt et al., 2003; Liu et al., 2004; Micheletti et al., 2004). The flap region 45-55, although being part of the relatively mobile flap, has reduced mobility in the bound state. This is in consistency with its low tolerance to mutations (Bahar et al., 1998; Kurt et al., 2003). Nevertheless, the region 36-44 is located at the solvent-exposed parts of the flap and has the highest mobility. These regions surround and anchor the peptide in the cleft between the two monomers.

The fluctuations in the most cooperative modes in substrate and in inhibitor complexes are averaged separately and the deviation of each residue from the average is calculated. The patterns below each panel in Figures 4.3-4.6 distinguishes the residues of each structure that fluctuates above or below average; residues that fluctuate above average and belove average are colored red and blue, respectively. The highest deviation from average collective motion is in rt-rh and p1-p6 substrate complexes in the first mode, and in rt-rh and p2-nc complexes in the second mode. nc-p1 has the closest motion to average of substrate complexes in the first two modes. On the other hand, clustering of MD snapshots of substrate complexes separately shows that rt-rh complex has lowest number of clusters and p1-p6 has highest number of clusters by the same rmsd value. Thus, deviation from average collective motion for rt-rh might mean that its structural motion is relatively restricted and its sampled conformational space is rather limited. Contrary to rt-rh, p1-p6 has the ability to sample the conformational space rather freely compared to the other complex structures. The restricted motion of rt-rh may be associated with the tight binding of the substrate. In several previous studies (Altman et al. (2007); Hou et al., 2008), rt-rh is also shown as the tightest binding natural substrate. Altman et al. (2007) designed tighter binding substrate-like peptides to the inactivated protease using rt-rh/inactivated protease complex as an initial model. To this end, paradoxically tighter binding should reduce the conformational entropy of the protease, which in return should decrease its binding energy.

The patterns below the first two slow mode profiles of substrate complex structures (Figures 4.3-4.4) reveal that they can be grouped according to the magnitude of residue fluctuations. The difference in fluctuations, particularly in the regions 12-22, 36-44, 61-73 and 85-96 in both monomers, sorts the substrates as one group consisting of ca-p2, ma-ca and rt-rh, and the other group consisting of nc-p1, p1-p6, p2-nc and rh-in. The regions 12-22 and 36-44 are both highly fluctuating regions in the first two modes, whereas regions 61-73 and 85-96 are comparably higher fluctuating regions in the second mode while they have reduced mobility in the first mode. The two groups of substrate complex structures behave conversely in these regions of the two monomers: The region 12-22 in the first monomer of group 1 (ca-p2, ma-ca, rt-rh) is more mobile than that of group 2 (nc-p1, p1-p6, p2-nc, rh-in) in the first mode (Figure 4.3), while this region in the first monomer of group 2 is more mobile than that of group 1 in the second mode (Figure 4.4). The second monomers behave conversely; i.e. region 12-22 in the second monomer of group 2 is more mobile than that of group 1 in the first mode (Figure 4.3), and it is more mobile in the second monomer of group 1 in the second mode (Figure 4.4). The region 36-44 behaves as follows: it is more mobile in the second monomer of group 1 and in the first monomer of group 2 in the first mode (Figure 4.3), and in the second mode, it is more mobile in the first monomer of group 1 and in the second monomer of group 2 (Figure 4.4). The pattern below the first slowest mode profile (Figure 4.3) suggests that the region 61-73 is more mobile in the first monomer of group 1 and in the second monomer of group 2, and the region 85-96 is more mobile in the first monomer of group 2 and in the second monomer of group 1. Similarly, the pattern below the second slowest mode profile (Figure 4.4) suggests that these two regions behave conversely in the second mode. The grouping of the substrates is suggested in the work of Pettit et al. (2002) where they classify processing sites of substrates based on cleavage rate associated with a specific subset of P1 amino acids. The two groups are defined by the size of the amino acid in the P1' position; p2-nc and nc-p1 cleavage sites being in one group and ca-p2 and ma-ca cleavage sites being in the other group (Pettit et al., 2002). These substrates also fall into the same groups in our classification based on the residue fluctuations.

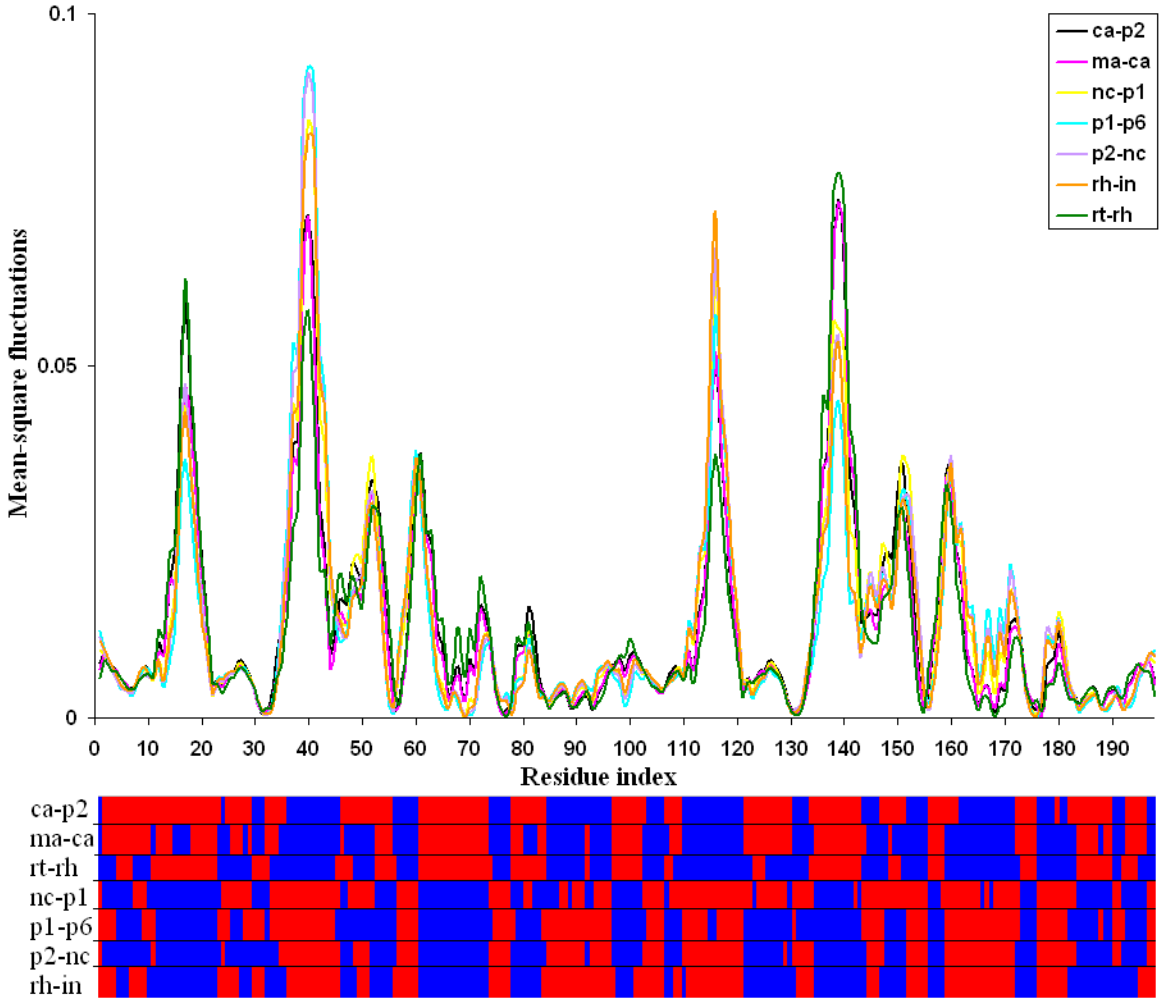


Figure 4.3. Mean square fluctuations of protease residues in the substrate complex structures in the first mode

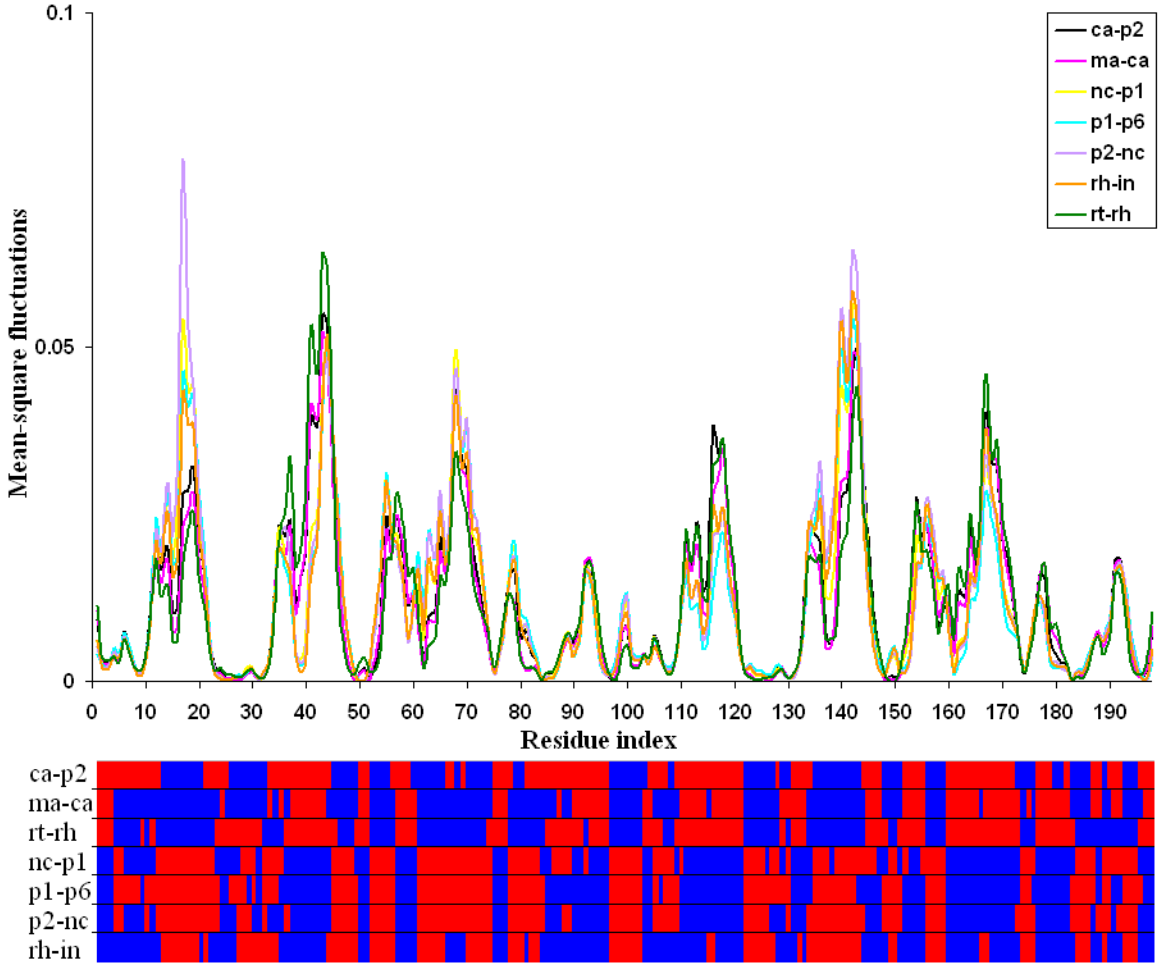


Figure 4.4. Mean square fluctuations of protease residues in the substrate complex structures in the second mode

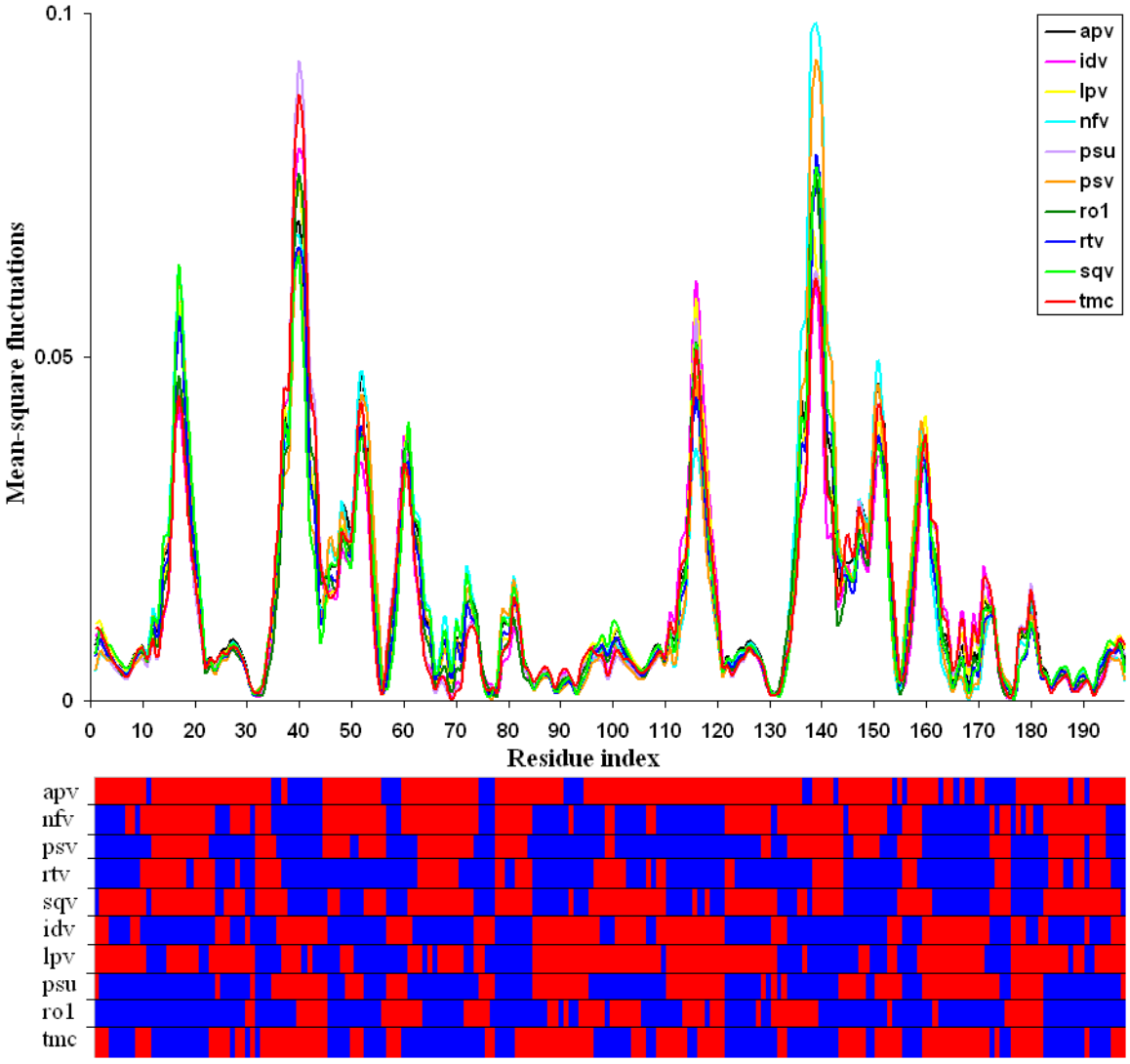


Figure 4.5. Mean square fluctuations of protease residues in the inhibitor complex structures in the first mode

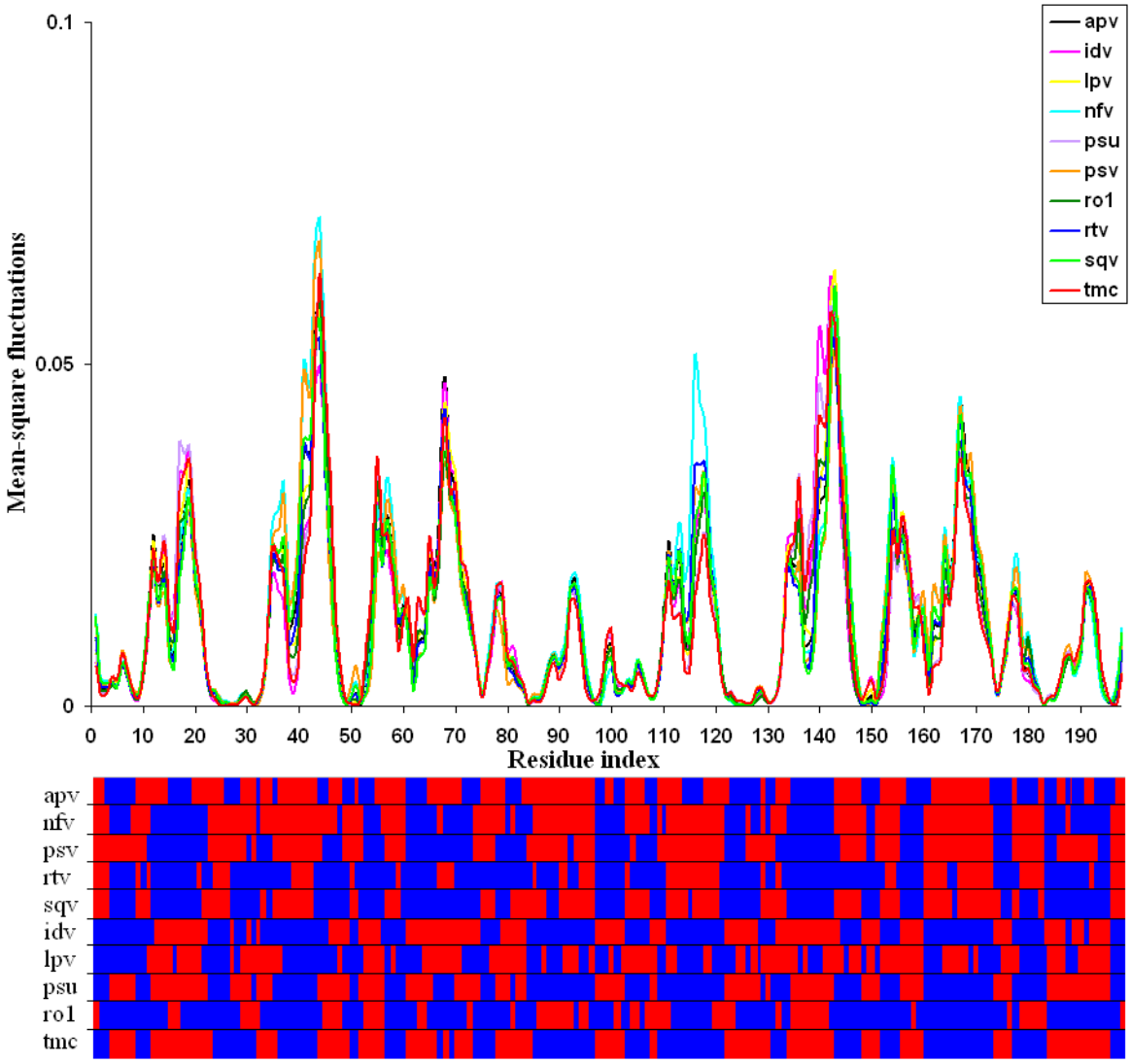


Figure 4.6. Mean square fluctuations of protease residues in the inhibitor complex structures in the second mode

In the case of inhibitors, the highest deviation from the average collective motion in the most cooperative two modes is in nfv, followed by idv, psu, psv, tmc and sqv. On the other hand, apv, rtv, ro1 and lpv have the closest motion to average behavior of inhibitor complex structures in the first two modes. When the patterns below the first two slow mode profiles of inhibitor complex structures that distinguish the residue fluctuations above or below average (Figures 4.5-4.6) are analyzed, the grouping of inhibitor complex structures is also noticed. Nevertheless, these two groups of inhibitor complex structures are not as clear as in the case of substrates. The analysis for particularly the regions 12-22, 36-44, 61-73 and 85-96 in inhibitor complexes, which are analyzed in detail for the substrates, can sort the inhibitors as one group consisting of apv, nfv, psv, rtv and sqv, and the other group consisting of idv, lpv, psu, ro1 and tmc. The two groups of inhibitor complexes also behave conversely in the most mobile regions of the two monomers as in substrates. Both minimum and maximum fluctuating regions of group 1 inhibitor complex structures in the first monomer coincide with those of group 1 substrate complex structures. This reveals that the fluctuations of the group of substrates consisting of ca-p2, ma-ca and rt-rh correlates with that of the group of inhibitors consisting of apv, nfv, psv, rtv and sqv. Similarly, the fluctuations of the group of the substrates consisting of nc-p1, p1-p6, p2-nc and rh-in correlates with that of the group of inhibitors consisting of idv, lpv, psu, ro1 and tmc. Observing similar profiles in the fluctuations of both subtrate and inhibitor complex structures may point to an intrinsic behavior for the protease structure; this will be further elaborated below.

PCA is performed on the conformations from 11 ns MD simulations of the natural substrate complex structures. The first few PCs collectively capture the majority of the total variance in the fluctuations. The average contribution of the first ten PCs of seven substrate complex structures is 87% of this MD trajectory. The first ten ANM modes of seven substrate complex structures correlate with 76% of the MD trajectory on the average. Root mean-square inner product (RMSIP) values between the first several PCs and the first several ANM modes are calculated to measure the coverage of the motion subspaces spanned by each approach. Table 4.1 summarizes the overlaps between these PCs and ANM modes for each substrate complex structure. Large RMSIP values can be seen even with three modes, and improvements are achieved as

more modes are included. These results suggest that the majority of the dynamics of protease complex structures can be explained by a small set of low-frequency ANM modes. This is also claimed in the recent work of Yang et al. (2008) where they identified essential motions of inhibitor bound HIV-1 protease for several data sets of X-ray structures, NMR ensembles and MD simulation and compared them with their coarse-grained elastic network model normal modes. Nevertheless, it should be noted that this is as much reflected by PCA of an MD trajectory of a given length. Further, the ANM modes in principle could represent large scale motions that could not spanned by 11 ns MD simulations. The length of the MD simulations may not be long enough to define the motion in the most cooperative modes, hence the grouping of substrate and inhibitor complex structures are not observed in MD simulated structures. Yet, the present MD simulations still provide assurance and at the same time could be complimentary for several other dynamic properties that are of interest here.

Table 4.1: Overlaps between PC and ANM mode spaces of the subsrate complex structures

ca-p2	**3 PCs**	**6 PCs**	**10 PCs**	**20 PCs**
3 ANM modes	0.61	0.55	0.57	0.46
6 ANM modes	0.66	0.60	0.64	0.58
10 ANM modes	0.74	0.68	0.70	0.67
20 ANM modes	0.83	0.80	0.80	0.79
ma-ca	**3 PCs**	**6 PCs**	**10 PCs**	**20 PCs**
3 ANM modes	0.66	0.53	0.55	0.46
6 ANM modes	0.69	0.59	0.63	0.56
10 ANM modes	0.73	0.66	0.68	0.62
20 ANM modes	0.77	0.75	0.78	0.74
nc-p1	**3 PCs**	**6 PCs**	**10 PCs**	**20 PCs**
3 ANM modes	0.66	0.66	0.63	0.49
6 ANM modes	0.70	0.70	0.69	0.59
10 ANM modes	0.73	0.74	0.74	0.66

Table 4.1: continued

20 ANM modes	0.80	0.81	0.81	0.78
p1-p6	**3 PCs**	**6 PCs**	**10 PCs**	**20 PCs**
3 ANM modes	0.59	0.58	0.60	0.47
6 ANM modes	0.65	0.66	0.66	0.56
10 ANM modes	0.68	0.71	0.70	0.63
20 ANM modes	0.80	0.81	0.81	0.78
p2-nc	**3 PCs**	**6 PCs**	**10 PCs**	**20 PCs**
3 ANM modes	0.59	0.67	0.57	0.48
6 ANM modes	0.62	0.69	0.64	0.56
10 ANM modes	0.68	0.74	0.70	0.66
20 ANM modes	0.77	0.81	0.80	0.78
rh-in	**3 PCs**	**6 PCs**	**10 PCs**	**20 PCs**
3 ANM modes	0.62	0.57	0.57	0.43
6 ANM modes	0.66	0.66	0.64	0.54
10 ANM modes	0.74	0.73	0.71	0.63
20 ANM modes	0.85	0.85	0.83	0.80
rt-rh	**3 PCs**	**6 PCs**	**10 PCs**	**20 PCs**
3 ANM modes	0.51	0.57	0.61	0.47
6 ANM modes	0.59	0.64	0.68	0.58
10 ANM modes	0.67	0.68	0.73	0.68
20 ANM modes	0.82	0.82	0.84	0.80

The eigenvalues that represent the frequencies of the individual 50 ANM modes for all substrate and inhibitor complex structures are plotted in ascending order in Figure 4.7. It is also clearly observed here that the frequency of motion is very similar for all the structures in the ten slowest modes, yet the eigenvectors in the slowest modes display some differences. In the later modes, lpv and tmc, which are known to be the strong binding inhibitors (Wang and Kollman, 2001; Prabu-Jeyabalan et al., 2006; Chellappan et al., 2007; Hou et al., 2007), have lower eigenvalues than other complex structures. Lower eigenvalue, i.e. lower frequency, suggests relatively more contribution

of the corresponding eigenvector to the overall motion and thus, this observation here may imply a more cooperative motion for the strong inhibitors in the same window of eigenvectors.

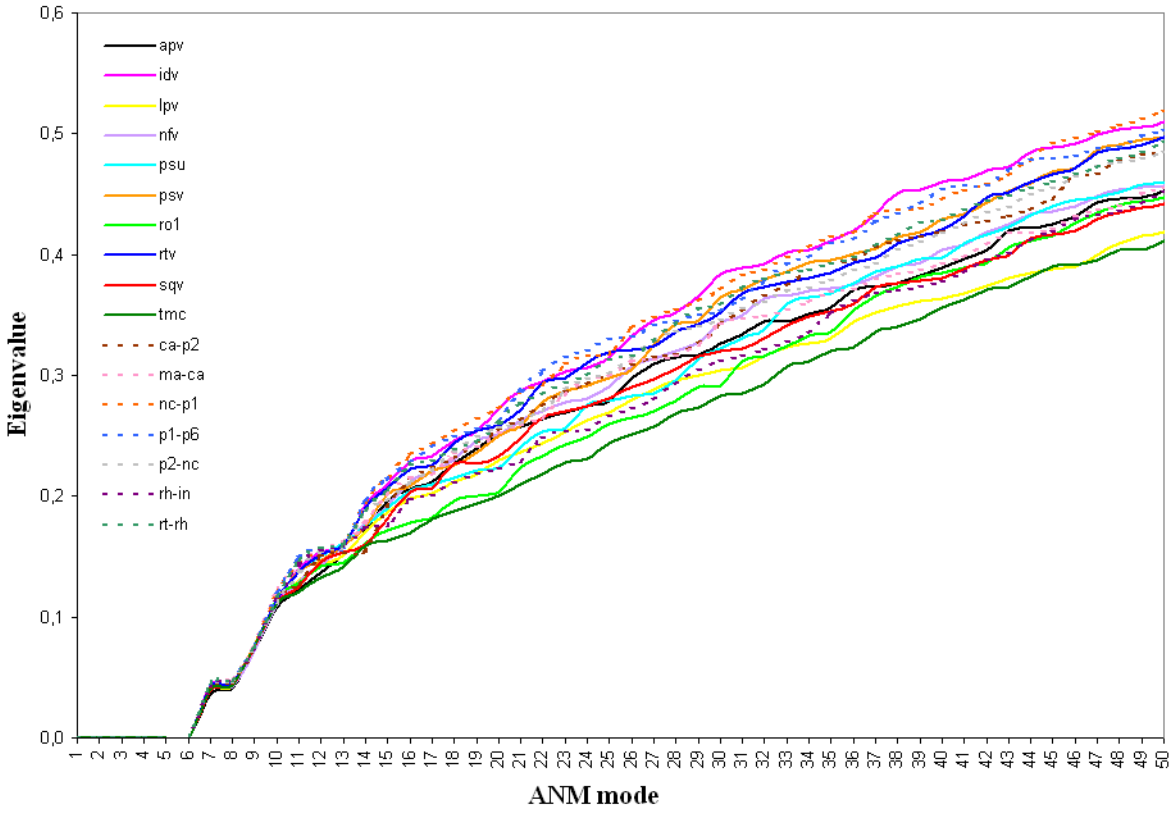


Figure 4.7. Eigenvalues from ANM

The correlation of the fluctuation of the two monomers of the protease in all the substrate and inhibitor complex structures are analyzed to observe the symmetry in the fluctuations between the two monomers of HIV-1 protease, which is a symmetric structure in unbound state. Table 4.2 shows these correlation coefficients for the first two modes and for the average of first ten modes. In the most cooperative modes, the average correlation coefficient between the monomers of the inhibitor complex structures are higher than the average of substrate complex structures, implying more symmetry in the inhibitor complexes. When substrate and inhibitor complex structures are analyzed separately, it is observed that the monomers correlate at the least in p1-p6 and nfv complex structures, respectively, in the first mode where the motion is observed along the Z axis (Figure 4.1). In the second slowest mode where the motion is observed along the X axis (Figure 4.2), the monomers of p1-p6, p2-nc, psu, nfv, psv

and idv complexes are less correlated. In the average of the first ten ANM modes, p2-nc and nfv have the least correlation between their monomers among the substrate and inhibitor complex structures, respectively. However, averaging the first ten PCs of MD simulated substrate complex structures showed minimum correlation of monomers in p1-p6. Lower correlation between the fluctuations of the two monomers indicates higher asymmetry in the fluctuations of the dimer structure, which is observed mainly in p1-p6 among substrate complexes and nfv among inhibitor complexes. These complex structures exhibits the highest deviation from average collective motion as well.

Table 4.2. The correlation coefficients between the magnitude of mean square fluctuations of the two monomers of HIV-1 protease complex structures

	mode 1	**mode 2**	**first 10 modes**
Substrates			
ca-p2	0.99	0.97	0.98
ma-ca	0.98	0.96	0.99
nc-p1	0.90	0.82	0.98
p1-p6	0.70	0.65	0.98
p2-nc	0.81	0.65	0.90
rh-in	0.81	0.77	0.99
rt-rh	0.85	0.79	0.98
Inhibitors			
apv	0.99	0.99	0.99
idv	0.90	0.86	0.98
lpv	0.99	0.99	0.96
nfv	0.89	0.86	0.84
psu	0.93	0.85	0.97
psv	0.90	0.86	0.92
ro1	0.99	0.97	0.97
rtv	0.97	0.97	0.91
sqv	0.93	0.93	0.97
tmc	0.93	0.87	0.98

4.2. Orientational Correlations

The inner products of eigenvectors that define the mode shapes are calculated to observe the orientational correlations between the fluctuations of the same residues in different complex structures. The orientational correlations of protease residues between the substrate and between the inhibitor complex structures in the most cooperative two modes are displayed in Figures 4.8-4.20. The value of each residue in the charts represent the dot product value of the fluctuation vectors that define the direction of motion for that residue of the two complex structures compared. Thus, the peaks with negative correlation values in the charts indicate the residues that fluctuate more diversely thereby causing the orientational difference between the two structures compared.

The groups of substrates and inhibitors identified with respect to the magnitude of residue fluctuations are more clearly observed with this analysis, especially in the first mode. Figures 4.8-4.10 show the orientational correlations of the substrate complex structures in the first mode; between the residues of the structures among group 1 (Figure 4.8), among group 2 (Figure 4.9), and between group 1 and 2 (Figure 4.10). The same applies for the inhibitor complex structures in Figures 4.11-4.13. The charts here clearly demonstrate the grouping according to the similarity in orientation of motion; the orientational correlations among the structures that fall into the same group of the structures are much higher compared to those between the structures that fall into the different groups. The least correlating residues between different structures even in the same groups correspond to 56, 69, 78 and 93 in both monomers of substrate and inhibitor complexes. However, the orientational correlation values of these residues differ in the two monomers of different groups. In group 1 of substrates and inhibitors (Figures 4.8 and 4.11), which are claimed to be correlated to each other based on magnitude of residue fluctuations, the orientational difference is mainly in the second monomer. Residue 93 has a lower correlation value in the first monomer, yet lower correlation of residues 69 and 78 in the second monomer is more dominant to cause the asymmetry within this group (Figures 4.8 and 4.11). On the other hand, in group 2 of substrates and inhibitors (Figures 4.9 and 4.12), the orientational differ-

ence is caused predominantly by residue 69 in the first monomer, although residues 78 and 93 have lower correlation values in the second monomer. When the orientational correlations between the two groups of substrates (Figure 4.10) and inhibitors (Figure 4.13) are compared, residue 69 which has a remarkably lower dot product value in the first monomer causes the orientational difference and thus the asymmetry between the monomers. In general, the asymmetry between the monomers of the substrate complex structures is higher than the inhibitor complex structures. Moreover, the correlation values between the residues among the substrate complex structures is much lower compared to the correlation values among the inhibitor complex structures, which is particularly noticeable in the orientational correlation of residues between the two different groups of substrates and inhibitors (Figures 4.10 and 4.13). This demonstrates the more similar motion in the inhibitor complexes, i.e. the dynamics of the protease does not change significantly upon binding of different inhibitors, whereas binding of different substrates allows the protease sample a larger conformational space, apparently due to the flexible nature of the substrate structures.

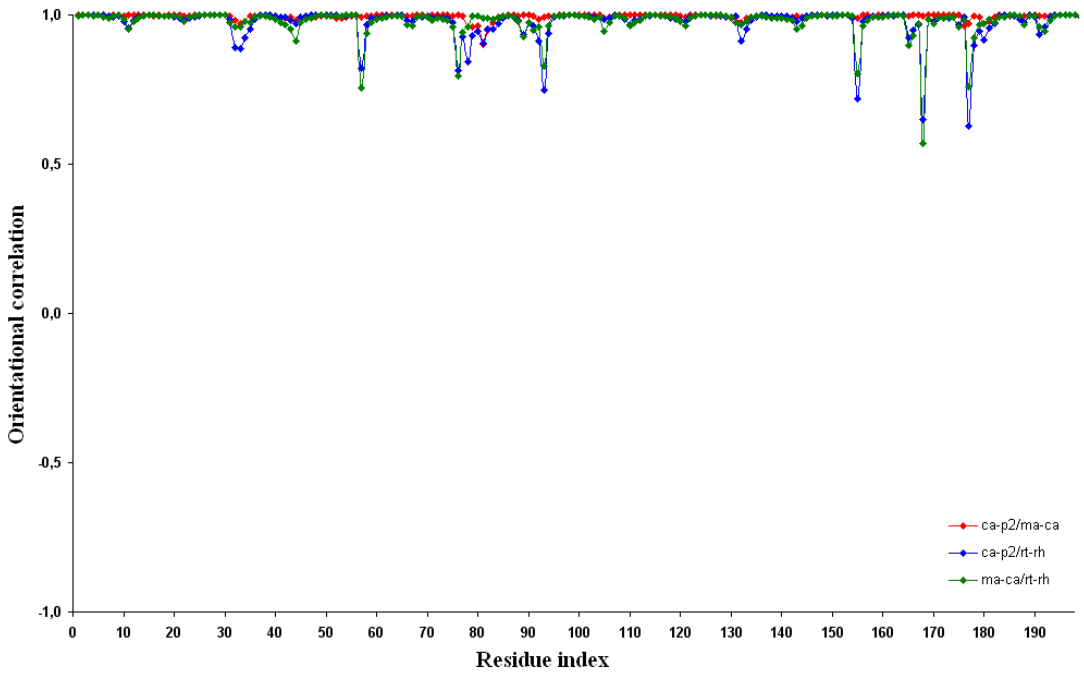


Figure 4.8. Orientational correlation of protease residues of substrate complexes in group 1 with those of the other substrate complexes in the same group in the first mode

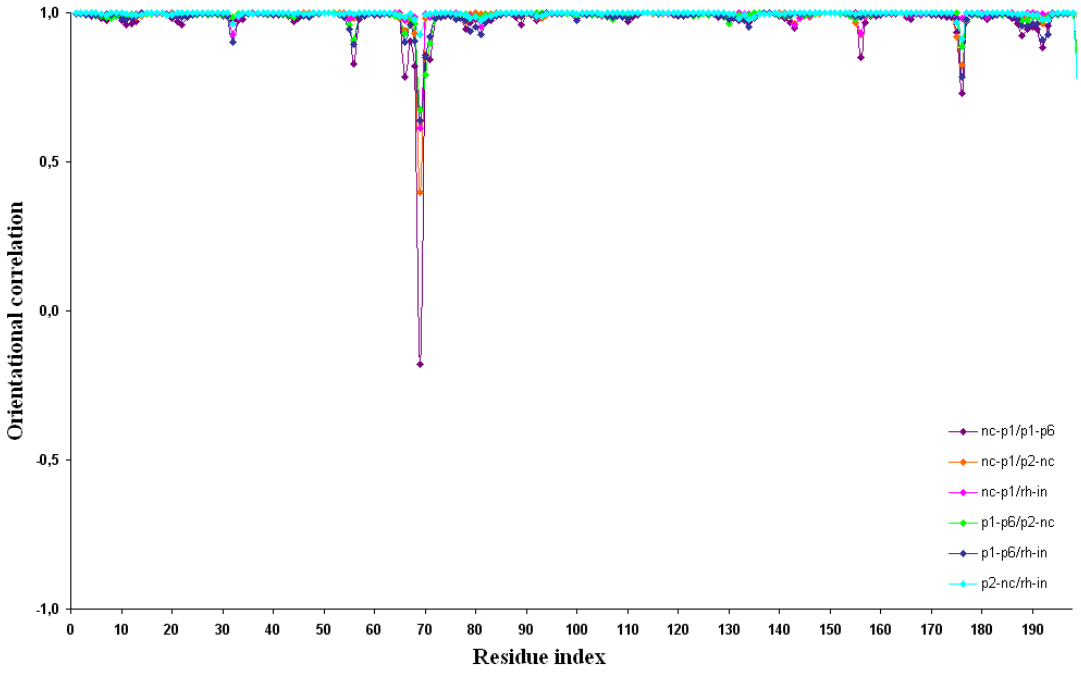


Figure 4.9. Orientational correlation of protease residues of substrate complexes in group 2 with those of the other substrate complexes in the same group in the first mode

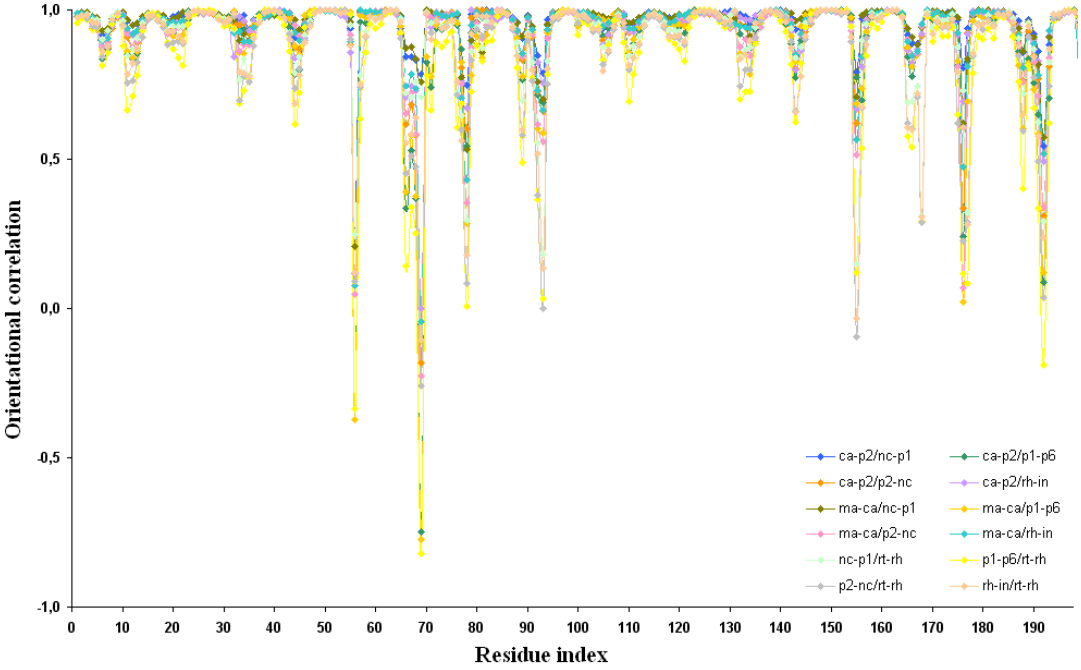


Figure 4.10. Orientational correlation of protease residues of substrate complexes in group 1 with those of the substrate complexes in group 2 in the first mode

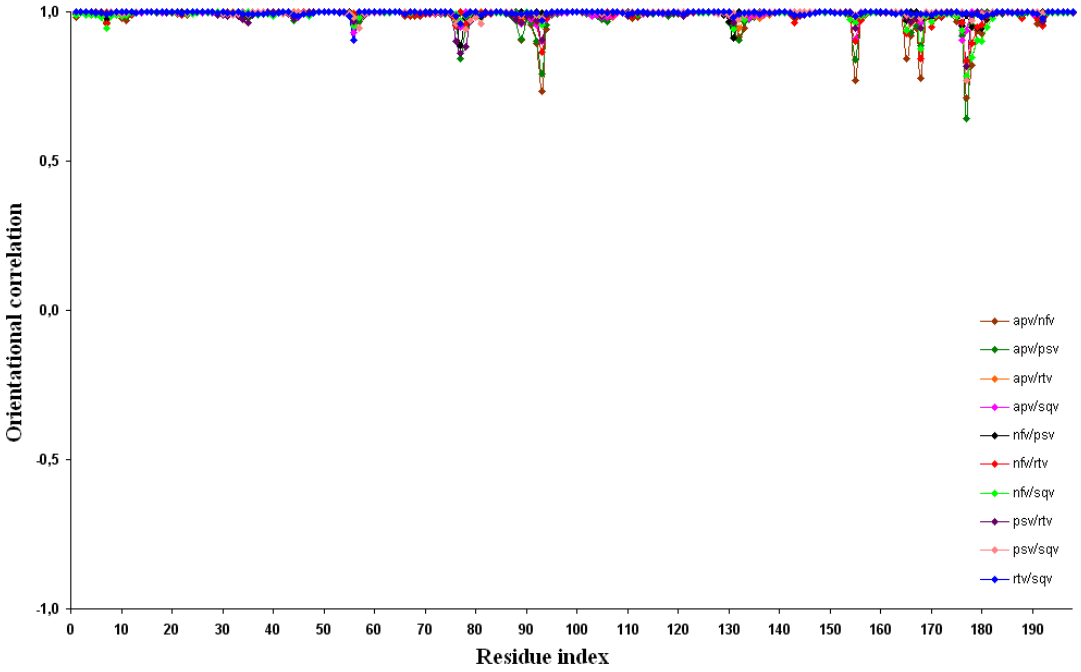


Figure 4.11. Orientational correlation of protease residues of inhibitor complexes in group 1 with those of the other inhibitor complexes in the same group in the first mode

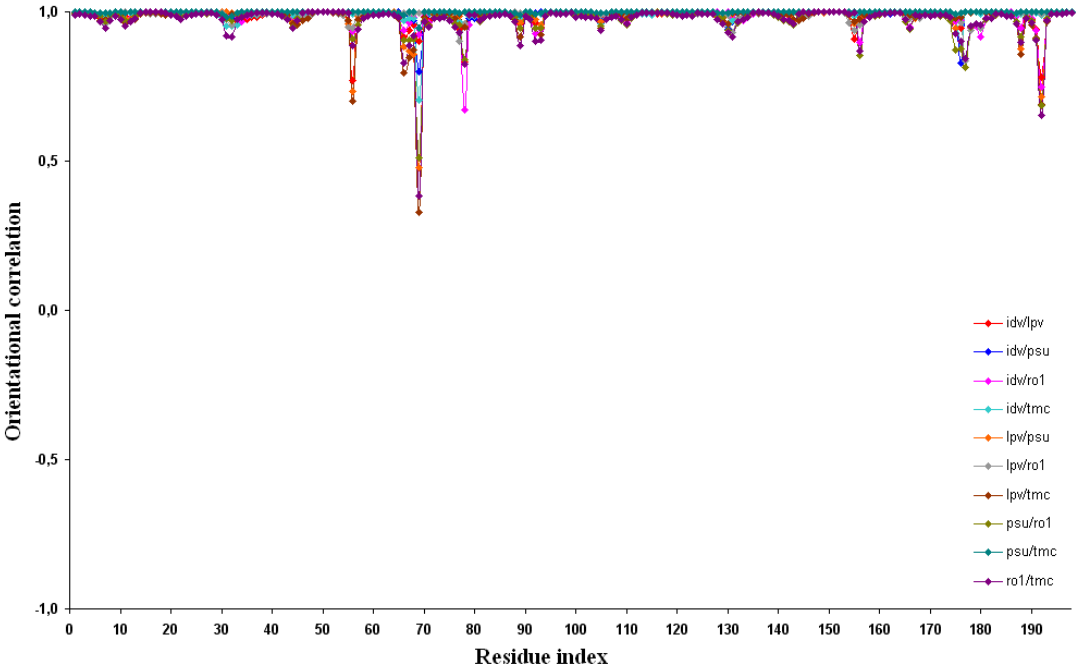


Figure 4.12. Orientational correlation of protease residues of inhibitor complexes in group 2 with those of the other inhibitor complexes in the same group in the first mode

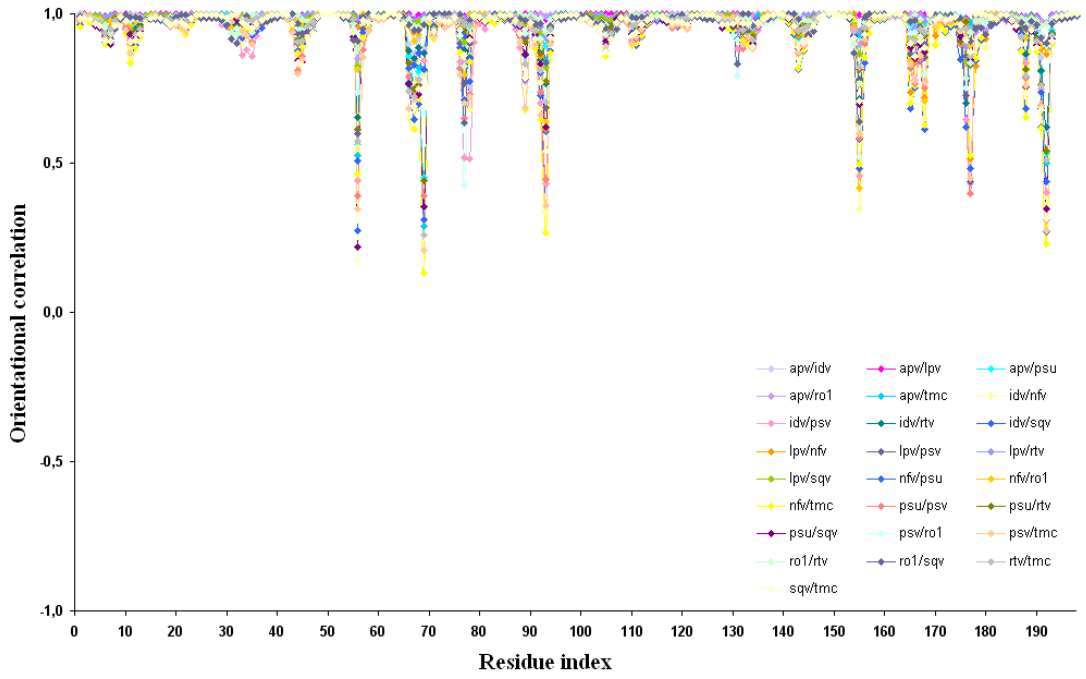


Figure 4.13. Orientational correlation of protease residues of inhibitor complexes in group 1 with those of the inhibitor complexes in group 2 in the first mode

The least correlating residues between different structures in the first mode, i.e. 56, 69, 78 and 93 in both monomers of substrate and inhibitor complexes, are observed to be in the minimum fluctuating regions in this mode (Figures 4.3 and 4.5). Residues 56 and 78 are very close in distance in space; 56 is the hinge point connecting the flexible flap loop (45-55) to the solvent-exposed upper arm of the flap (36-44), and 78 is the hinge point connecting the same flap loop (45-55) to the lower arm of the flap (57-77). These two arms in connection with the flap, namely the highly fluctuating region 36-44 and the minimum fluctuating region 61-73, appear to be the regions that imply the grouping of the substrates and inhibitors according to the magnitude of residue fluctuations. Besides, residues 69 and 93, which are the other least correlating residues in orientation, are also very close in space. 69 is the tip of the minimum fluctuating lower arm of the flap (61-73) and 93 is the tip of another minimum fluctuating loop (85-96) that also indicate the grouping of substrates and inhibitors in first mode. These specific loops and residues that bring on the grouping of structures are identified in Figure 4.14, where it is also noticed that these residues 56, 69, 78 and 93 lie along the two axes parallel to Z direction; i.e. the rotational axes around which the protease monomers rotate in the first slowest mode (see Figure 4.1).

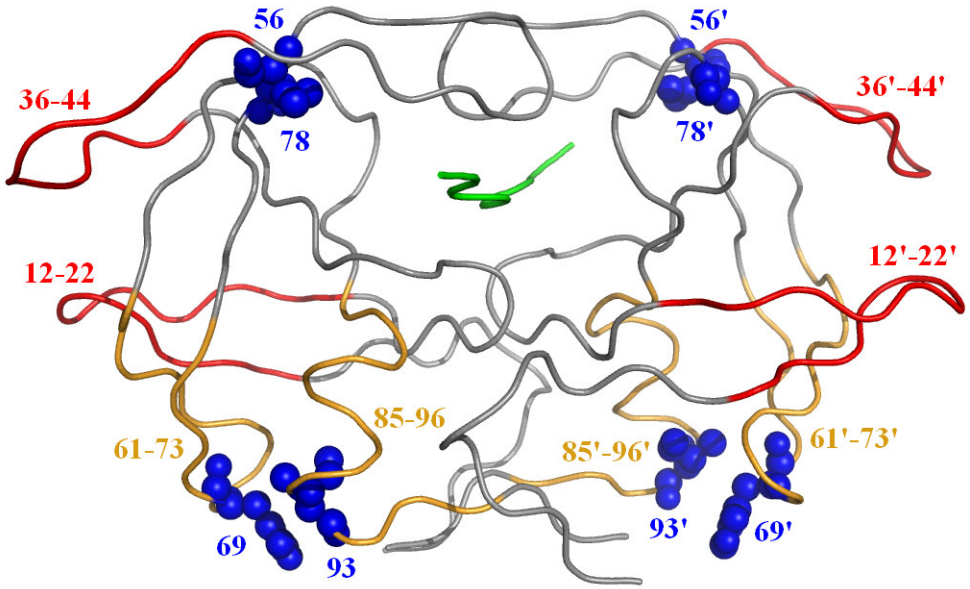


Figure 4.14. The regions causing the orientational difference in the first mode. The high fluctuating loops are displayed in red, the minimum fluctuating loops are displayed in orange, and the least correlating residues between different complex structures are displayed in blue.

In the second mode on the other hand, the orientational correlations do not pronounce the grouping of structures as clearly as in the first mode. Figures 4.15-4.17 show the orientational correlations of the substrate complex structures in the second mode; between the residues of the structures among group 1 (Figure 4.15), among group 2 (Figure 4.16), and between group 1 and 2 (Figure 4.17). The same applies for the inhibitor complex structures in Figures 4.18-4.20. The orientational correlations among the structures that fall into the same group of the structures are higher compared to those between the structures that fall into the different groups, similarly as in the first mode. Yet, the least correlating residues in the second mode correspond to 25-27, 49-51, 84 and 97 in both monomers of substrate and inhibitor complexes. The orientational correlation values of these residues differ in the two monomers of different groups of substrates particularly. The orientational difference and the asymmetry in the monomers is caused mainly by residues 25 and 49-51 of the second monomer in the first group of substrates (Figure 4.15) and by the same residues of the first monomer in group 2 of substrates (Figure 4.16). Residues 84 and 97 have the influence in the orientational difference between the two groups of substrates (Figure 4.17) and also

between the two groups of inhibitors (Figure 4.20). Further, the asymmetry between the monomers of the substrate complex structures is higher than the inhibitor complex structures in the second mode as well, pointing out to the more similar motion in the inhibitor complexes and sampling of a larger space by the substrate bound protease structures.

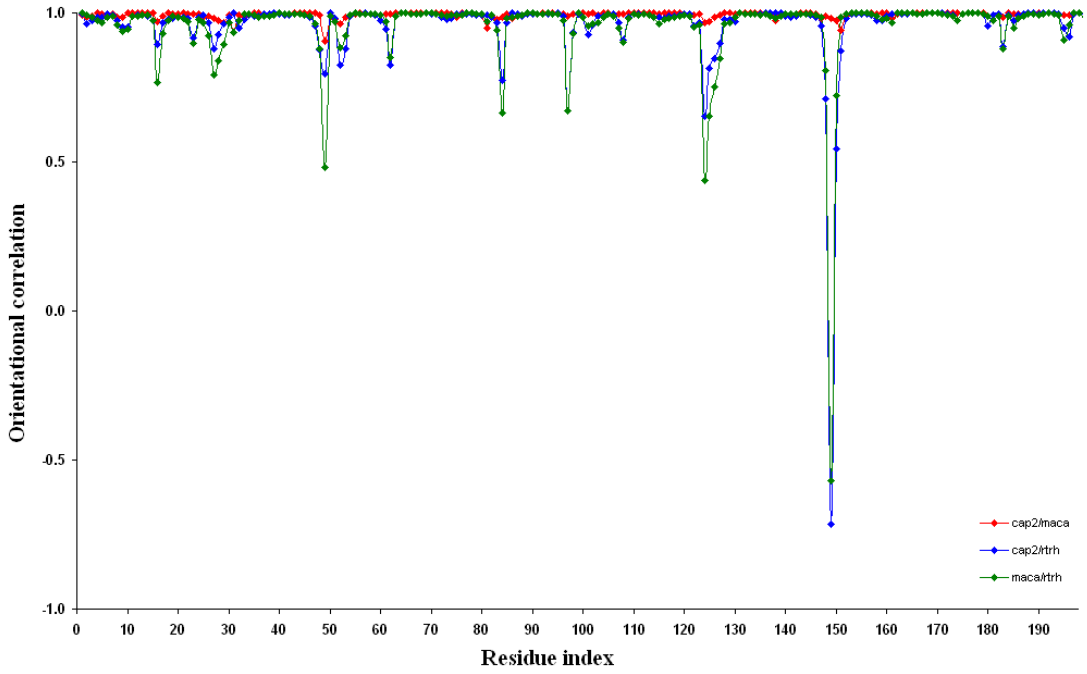


Figure 4.15. Orientational correlation of protease residues of substrate complexes in group 1 with those of the other substrate complexes in the same group in the second mode

The least correlating residues between different structures in the second mode, i.e. 25-27, 49-51, 84 and 97, are also observed to be in the minimum fluctuating regions in this mode (Figures 4.4 and 4.6). 49-51 are the residues at the flap tips, which are minimum fluctuating regions in the bound structures. Besides, 25-27 are the active site residues at the the tip of the loop right in the middle of the substrate cleft that is connected to the mobile 12-22 loop. This 12-22 region and the other regions 36-44, 61-73 and 85-96, are all highly fluctuating regions in the second mode which also imply the grouping of structures according to the magnitude of residue fluctuations. Residues 84 and 97, on the other hand, are the residues at the edges of the other mobile loop 85-96. These mobile loops and residues that cause the orientational difference between

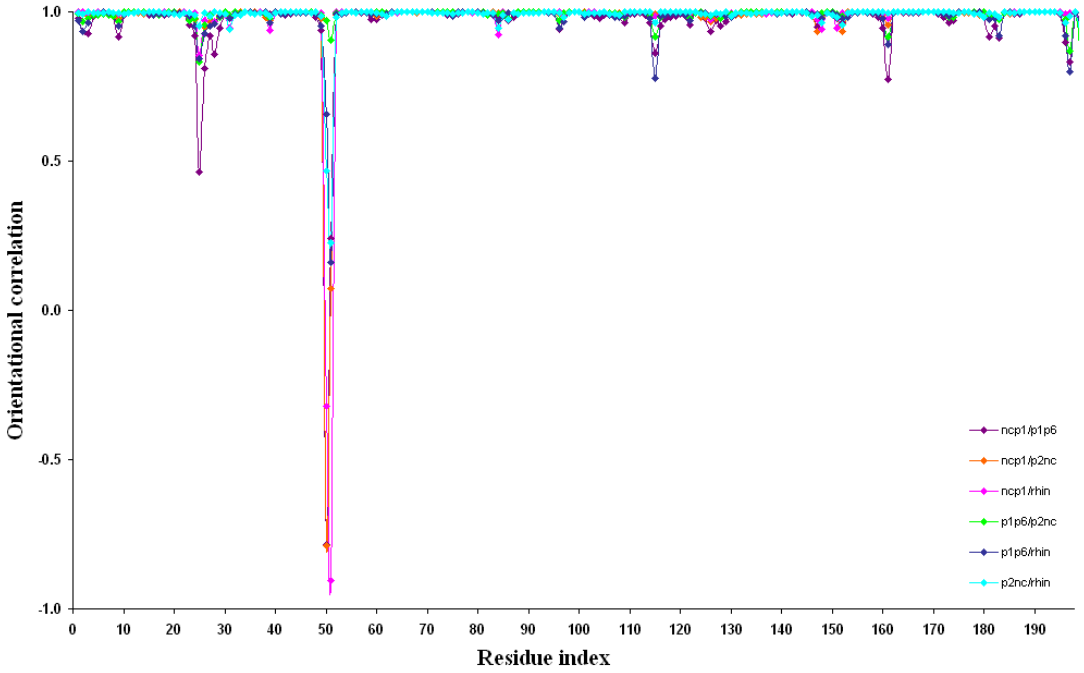


Figure 4.16. Orientational correlation of protease residues of substrate complexes in group 2 with those of the other substrate complexes in the same group in the second mode

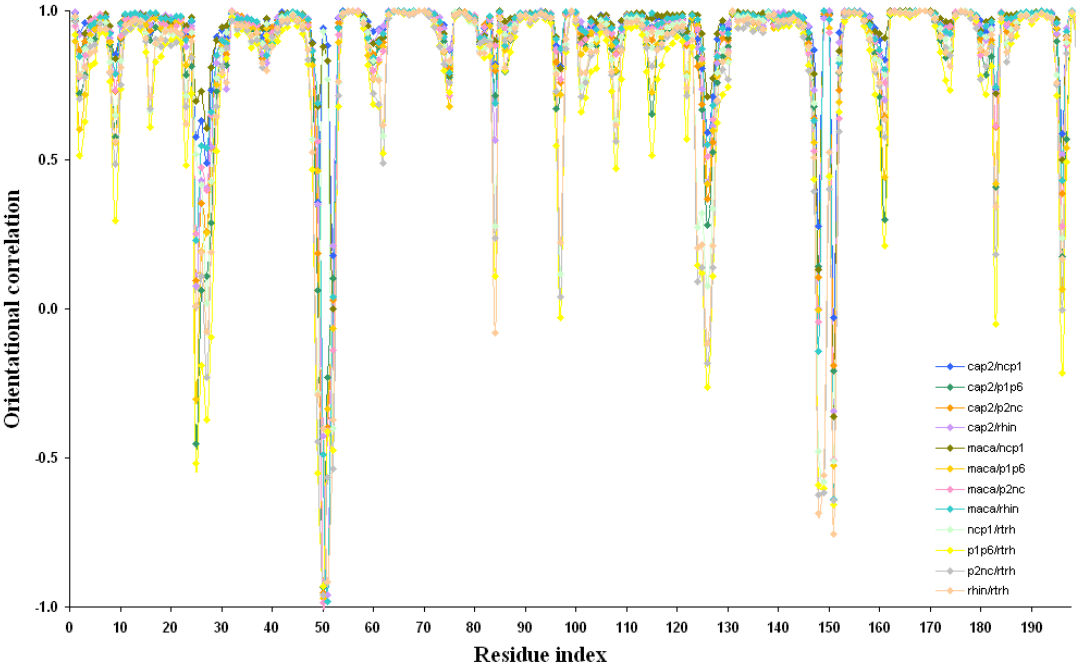


Figure 4.17. Orientational correlation of protease residues of substrate complexes in group 1 with those of the substrate complexes in group 2 in the second mode

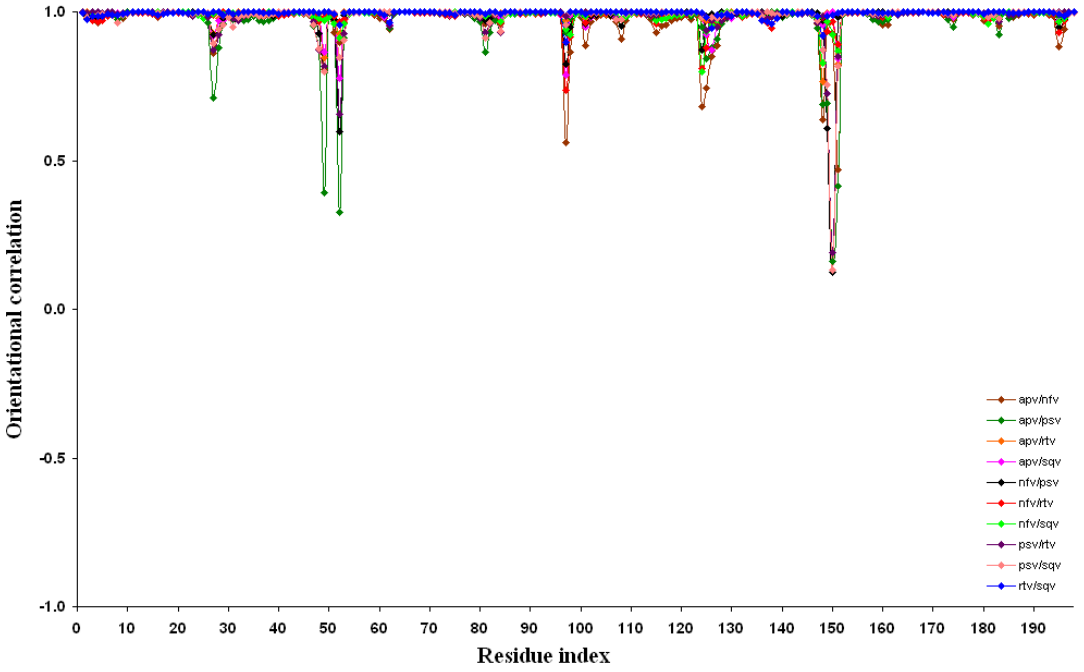


Figure 4.18. Orientational correlation of protease residues of inhibitor complexes in group 1 with those of the other inhibitor complexes in the same group in the second mode

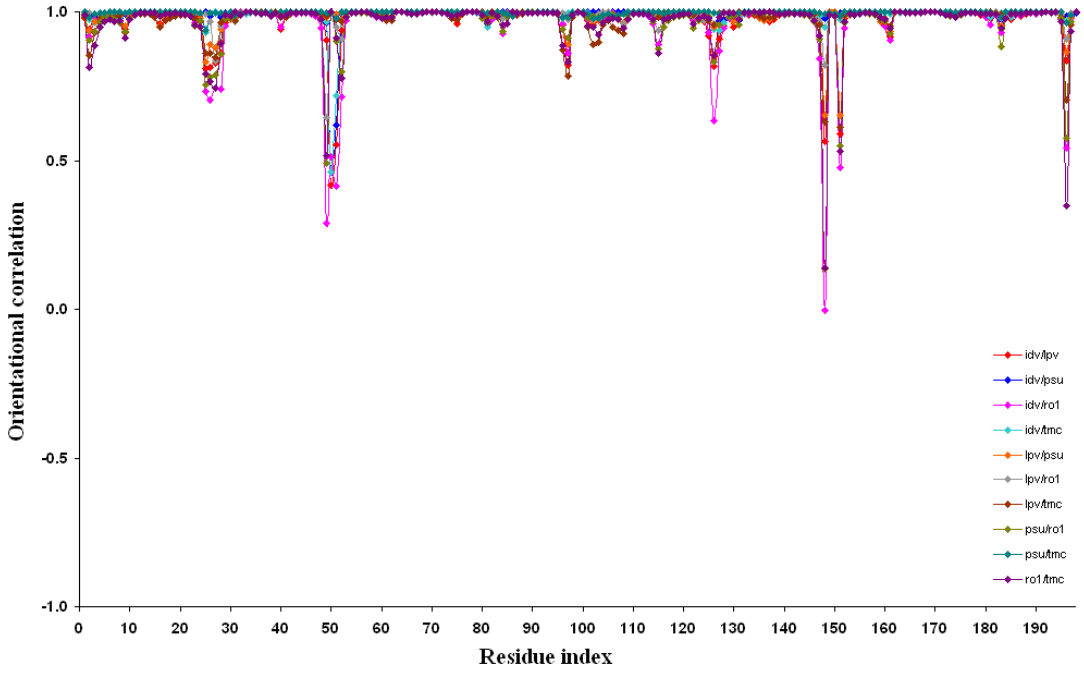


Figure 4.19. Orientational correlation of protease residues of inhibitor complexes in group 2 with those of the other inhibitor complexes in the same group in the second mode

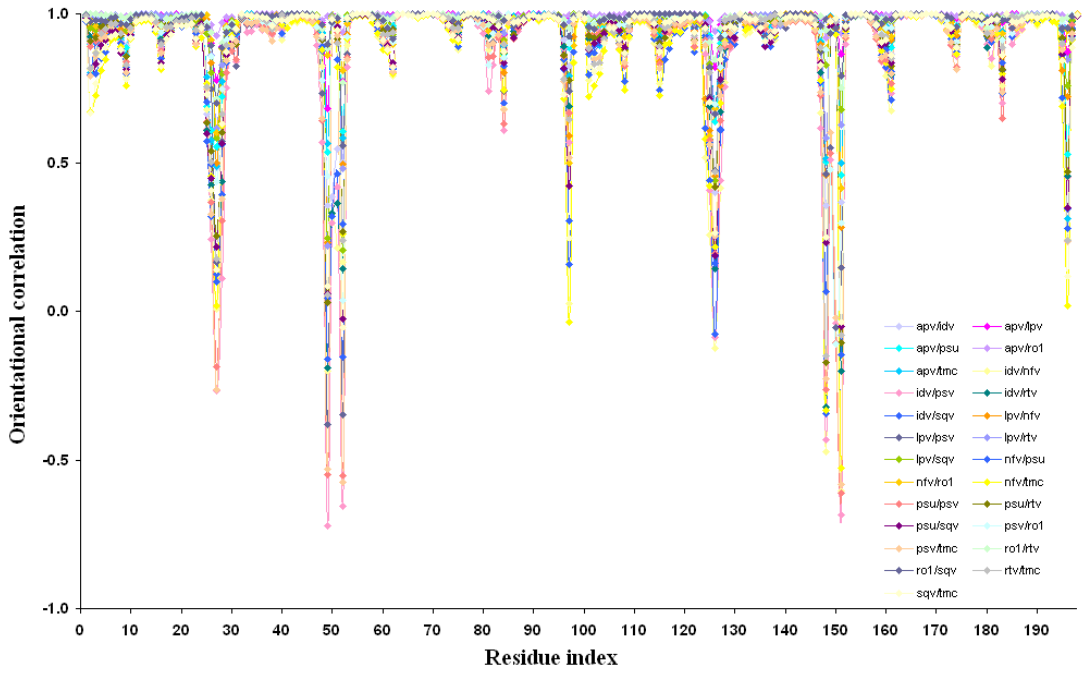


Figure 4.20. Orientational correlation of protease residues of inhibitor complexes in group 1 with those of the inhibitor complexes in group 2 in the second mode

the structures in the second slowest mode are identified in Figure 4.21, where it is also noticed that these hinge residues 25-27, 49-51, 84 and 97 lie along the axis parallel to Z direction; i.e. the rotational axis around which the monomers rotate in the second mode (see Figure 4.2).

Drug resistant mutations that occur in the protease after changing the inhibitor binding can also affect substrate recognition by changing the enzyme's substrate specificity. Then, the protease retains activity by the co-evolution of the substrate sequence. Mutations in the p1-p6 substrate cleavage site covary with the D30N/N88D protease mutations (Kolli et al., 2006). MD simulations carried out with the mutant and co-evolved variants of p1-p6 substrate complex structures are utilized here. The ANM analysis is performed for the representative members of the largest cluster of each structure. Figure 4.22 displays the orientational correlation of protease residues of the D30N mutant, the D30N-N88D mutant and the co-evolved p1-p6 complex structure with D30N, N88D and LP1'F mutations to those of the wild-type p1-p6 complex in the first mode.

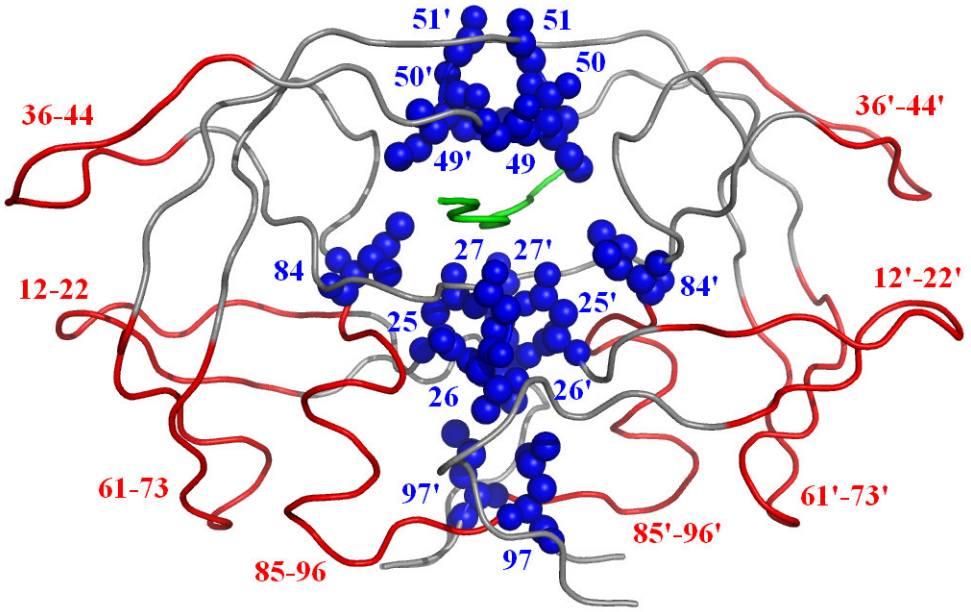


Figure 4.21. The regions causing the orientational difference in the second mode. The high fluctuating loops are displayed in red and the least correlating residues between different complex structures are displayed in blue.

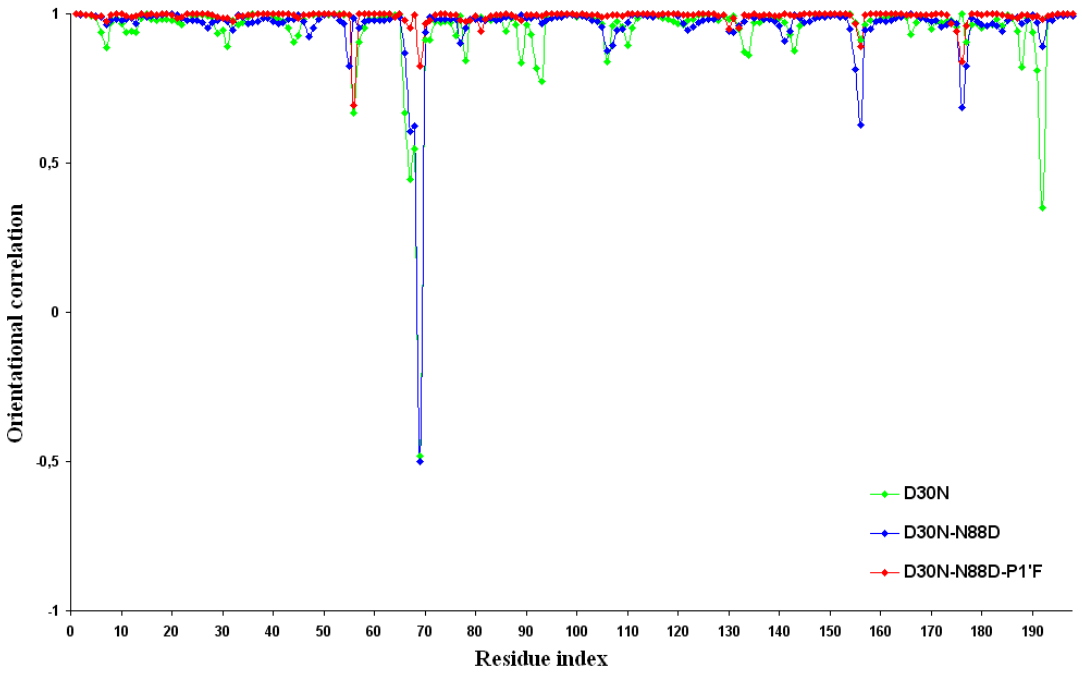


Figure 4.22. Orientational correlation of protease residues of D30N, N88D and D30N-N88D mutants and co-evolved p1-p6 substrate complexes to those of wild-type p1-p6 complex

Here the orientational difference between the structures is also caused by the minimum fluctuating residues. Correlation of the wild-type complex with the co-evolved structure is higher than the correlations with the mutant structures. Particularly, residue 69 of the first monomer is the least correlating residue that causes the difference in the orientation of translational motion (Figure 4.22). The mutation in the subtrate allows the protease residues fluctuate as in wild type and justify the existence of this mutation for the conservation of, at least, the fluctuations.

4.3. Correlations between the Direction of Fluctuations

The correlation between the fluctuations of residues are analysed for the substrate and inhibitor complex structures with respect to binding of peptides with the protease and dimerization of the monomers of the protease. In the binding and dimerization, the correlated fluctuations across the peptide and protease and across the interface of the two monomers of the protease are elaborated. The positively correlated atoms that fluctuate in the same direction are being focused in the present analysis. This in a way describes the association points between two interacting structures.

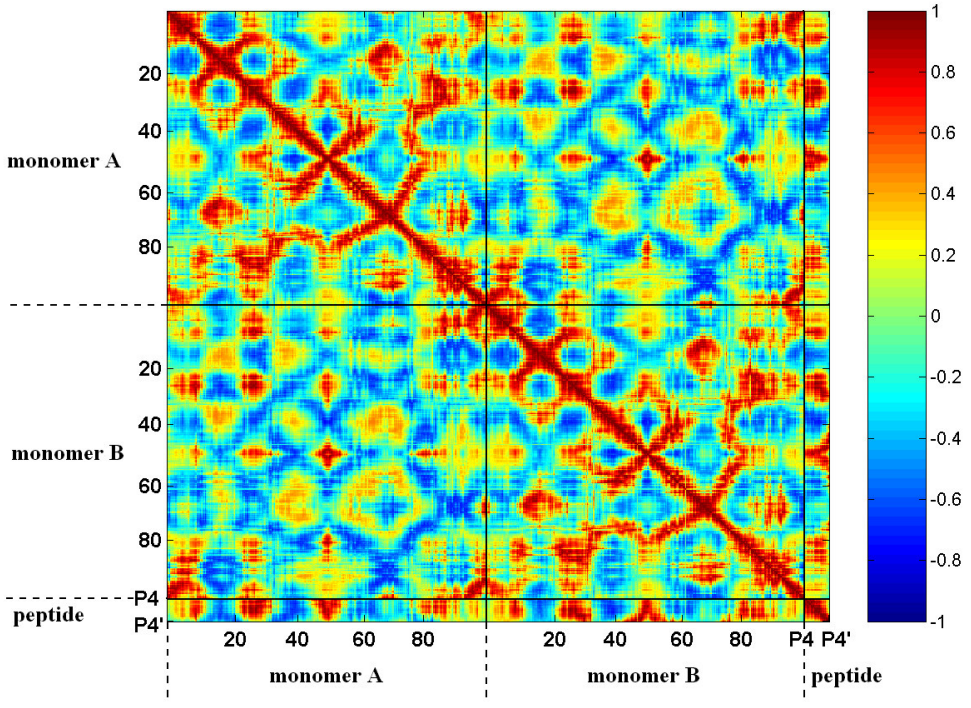


Figure 4.23. Cross correlations of residues in HIV-1 protease complex structure in the first ten ANM modes

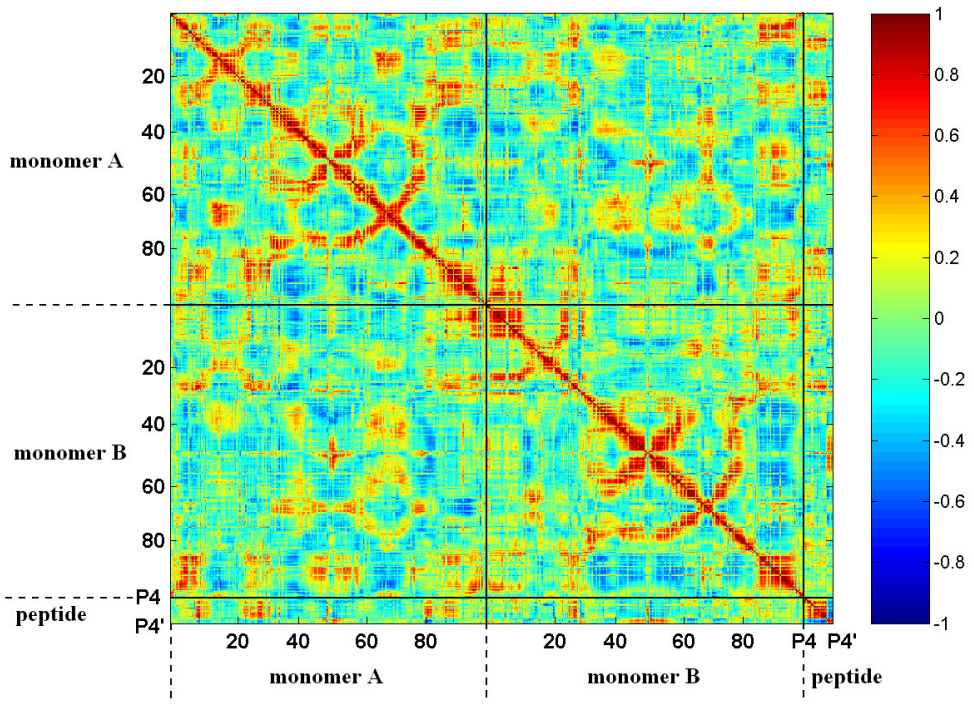


Figure 4.24. Cross correlations of residues in HIV-1 protease complex structure in the first ten PCs

The correlated dynamic fluctuations in the first ten ANM modes are displayed in Figure 4.23 for the ca-p2 substrate complex structure as an example, as very similar maps are obtained for the other substrate complex structures. Figure 4.24 shows the correlations in the first ten PCs of the same structure, which agree well with those in ANM except that they are less pronounced. The analysis of the fluctuations of the residues shows that the minima of the slowest modes shape (Figures 4.3-4.6) that correspond to 5-10 (dimerization region), 25-27 (active site), 45-55 (flap) and 80-90 (substrate cleft), i.e. the hinge regions of the two monomers in the most cooperative modes, and the N- and C- termini regions of the monomers are highly correlated with each other. As for the interaction between the protease and the peptide, the residues of the protease that display positively correlated fluctuations with the fluctuations of the peptide's residues also correspond to these hinge regions (Figures 4.23 and 4.24). The highly fluctuating regions on the other hand, such as solvent-exposed arms of the protease, display negative correlation between the monomers or with the peptide. Strong positive correlation between the substrate motion and the regions 24-30 and 45-55 is also found in previous works (Micheletti et al., 2004; Trylska et al., 2007). An

essential analysis of an MD trajectory of a substrate bound HIV-1 protease structure by Micheletti et al. (2004) reveals that regions 24-30 and 45-55 display a rotational "nutcracker-like" motion by embracing the substrate and the regions near the flaps elbows, 37-41 and 61-73, undergo a countermovement that results in a negative correlation with the substrate motion. Micheletti et al. (2004) conclude that a force applied around residues 40 and 63 should affect the protease-substrate coupling as mobile regions should be involved in any functionally relevant mechanical coupling.

4.3.1. Correlations between the Peptide and the Protease

To identify the critical residues for peptide binding, the peptide atoms that are positively correlated to each protease atom are specified. In the most cooperative modes, the number of correlated atom pairs is higher for inhibitor complex structure than substrate complex structures, despite the higher number of atoms that substrates comprise. The tighter binding of inhibitors than substrates is also suggested in previous works (Luque et al., 1998; Wang and Kollman, 2001; Hornak and Simmerling, 2007; Hou et al., 2008). The stronger binding of inhibitors obviously results in restricting the motion of the inhibitor complex structures compared to the substrates. Luque et al. (1998) also stated that the substrates have higher flexibility than the synthetic inhibitors in solution and thus binding of substrates cause higher conformational entropy loss. By contrast, because of their higher flexibility, the substrates are more adaptable to backbone rearrangements or conformational changes induced by the protease mutations. The capacity of the inhibitors to adapt to changes in the geometry of the binding pocket is more restricted because they are less flexible (Hornak and Simmerling, 2007). Further, with their tighter binding, the orientational space of the protease residues' fluctuations between the inhibitor complex structures is more restricted than between the substrate complex structures. This similarity in orientation of the fluctuations for inhibitor complexes, together with their similarity in three-dimensional shape and electrostatic character, may also have implications for multi-drug resistance.

Figures 4.25 and 4.26 display the positive correlations above 0.9 in the first ten slowest modes for the substrate and inhibitor complex structures respectively. When

less than ten modes are taken into account, the positively correlated interactions are only higher in number as expected and more scattered in the same regions. Moreover, taking different threshold values for positive correlations only changes the number of interactions; yet the distribution remains similar. The significant residues to binding are mainly located in four regions: residues 8-10, 25-27, 45-55 and 80-90; yet regions 25-27 (active site) and 45-55 (flaps) are more emphasized compared to the other two regions. The peptide atoms that are positively correlated to each protease atom are also analysed within the cross correlations of the first few PCs of MD simulated substrate complex structures; the peaks of significant residues to binding are found at the same regions of the protease.

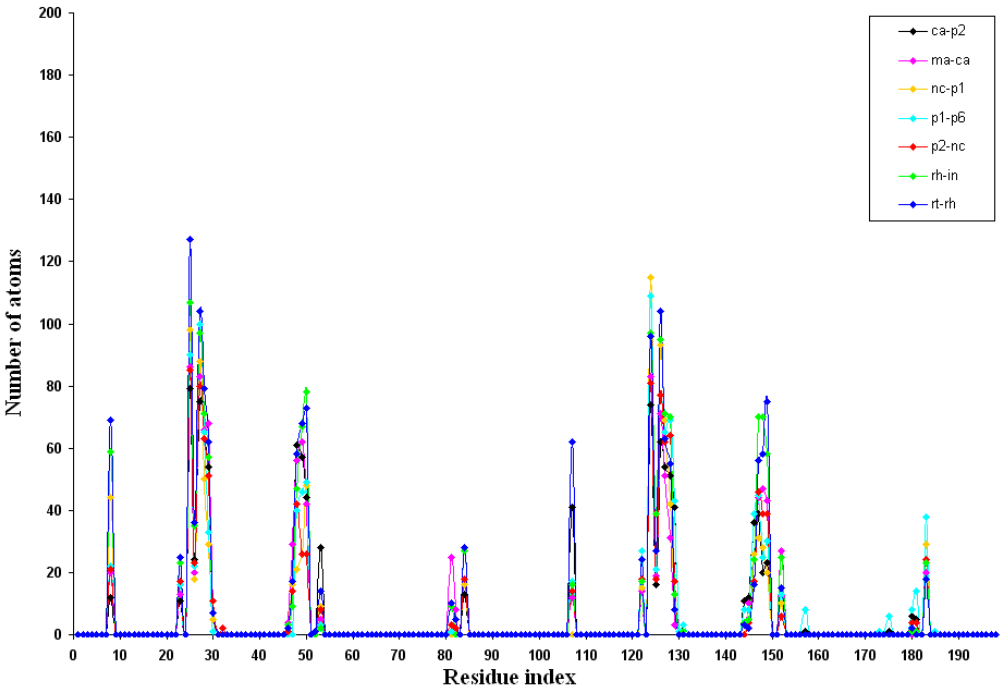


Figure 4.25. Number of peptide atoms positively correlated to each protease residue in substrate complexes in the first ten modes

The residues that are outside the strongly correlated regions but still interact with the peptide correspond to the residues in psu and psv complex structures which are the weakest inhibitors (Figure 4.26). These regions have further more interactions with the weak inhibitors when analyzed by the Gaussian Network Model (GNM) which is known as more robust in mean-square fluctuations (Cui and Bahar, 2005). Binding of these weak inhibitors to the positions other than the minimum fluctuating hinge regions in

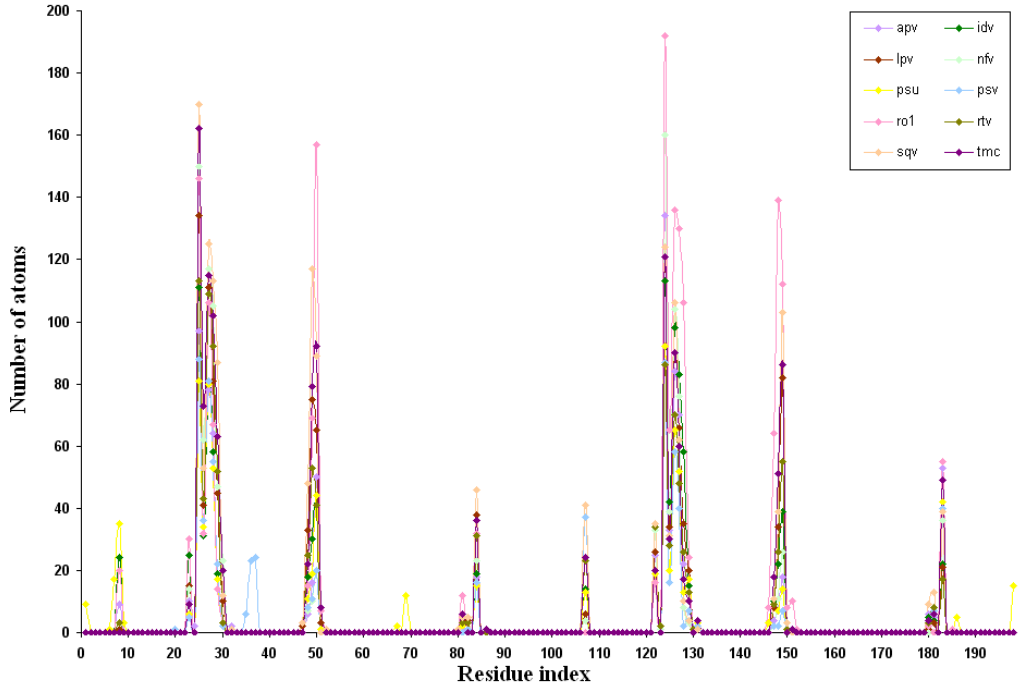


Figure 4.26. Number of peptide atoms positively correlated to each protease residue in inhibitor complexes in the first ten modes

the protease might play a role in decreasing the drug affinity, as the well conserved residues such as 25, 27, 28, 29 and 49 in the hinge regions are demonstrated to be critical for substrate binding in previous works (Wang and Kollman, 2001; Hou et al., 2008). Resistance-evading potent drugs should interact strongly with these residues.

4.3.2. Correlations across the Dimer Interface

To identify the critical residues in dimerization, the atoms of one monomer that are positively correlated to the other monomer are specified. Figures 4.27 and 4.28 display the interactions in the first ten ANM modes for substrate and inhibitor complex structures respectively. The interactions between the two monomers are observed at the same residues in the substrate and inhibitor complex structures. The cross correlations of the first few PCs of MD simulated substrate complex structures are investigated for the protease positions taking role in dimerization and they are also found on the same regions as in ANM. The only detail in PC correlations that should be noted is that the number of positively correlated residues between two monomers are higher than that between substrate and protease, when the same threshold value

as in ANM is taken. The significant residues to dimerization are mainly located in the same four specific regions as in binding to the peptide; 8-10, 25-27, 45-55 and 80-90. However, the regions 8-10 (dimerization), 25-27 (active site) and 45-55 (flaps) are more emphasized compared to the region 80-90. In addition, the N- and C- termini of the monomers are highly correlated with each other. The coupling between the binding site and the dimer interface is also suggested by the positive correlations between the active site and the C- and N- terminal residues of the monomers (Figures 7a and 7b). The importance of dimer interface for drug targeting is also stated by previous works (Hornak and Simmerling, 2007; Bowman et al., 2005) where they demonstrate that inhibitors that act as allosteric inhibitors binding at the dimer interface and alter the conformation of the protease can indirectly reduce the binding affinity of the substrate.

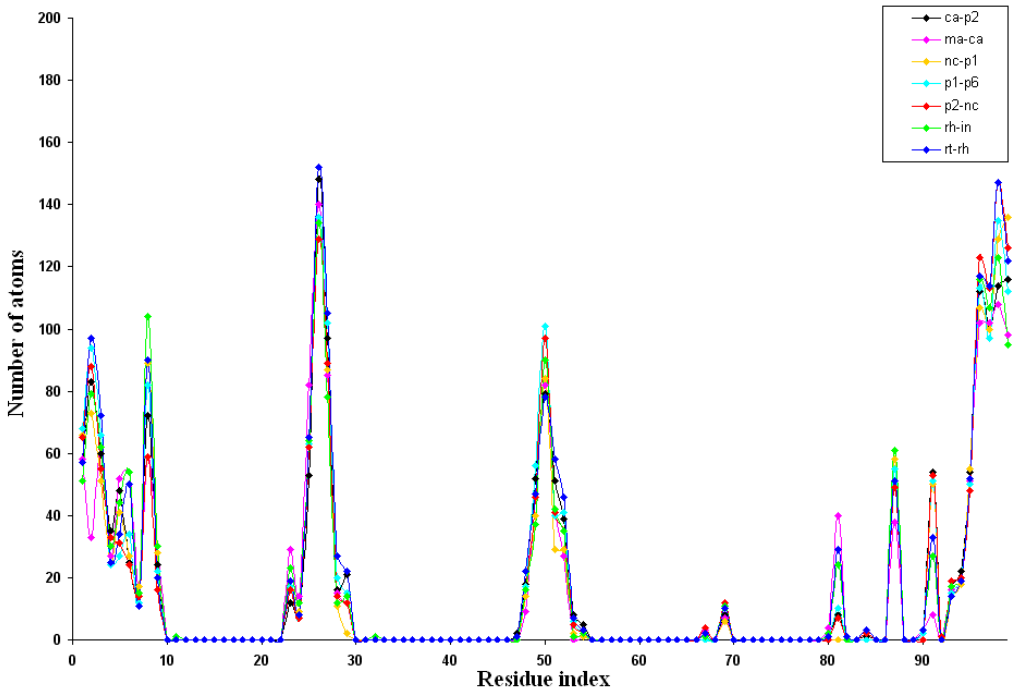


Figure 4.27. Number of atoms of one monomer positively correlated to the other monomer in substrate complexes in the first ten modes

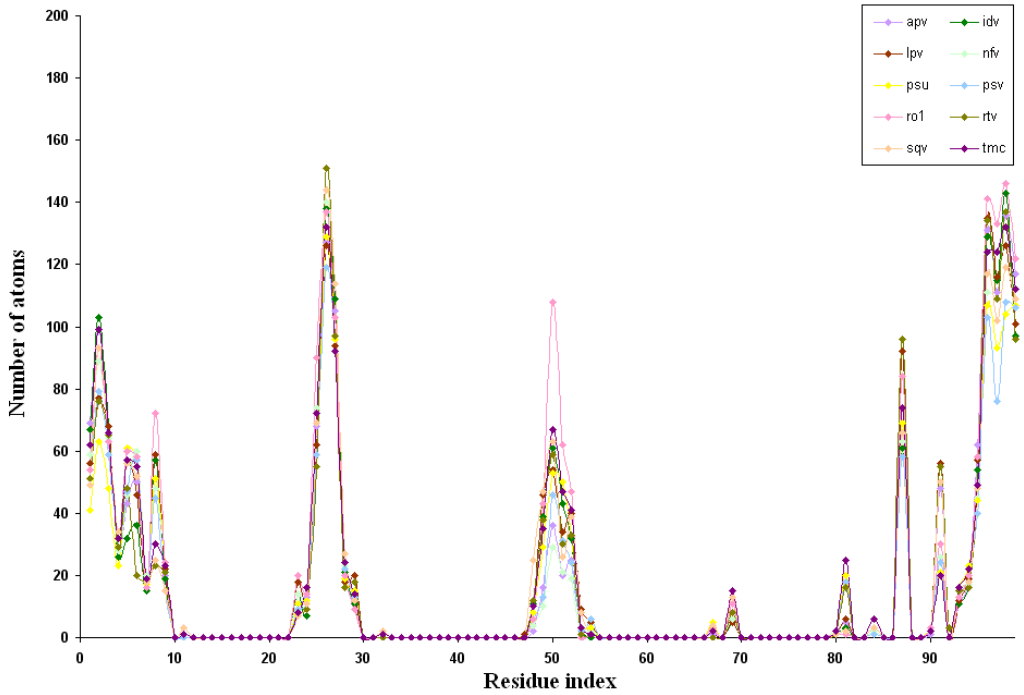


Figure 4.28. Number of atoms of one monomer positively correlated to the other monomer in inhibitor complexes in the first ten modes

5. PATHWAYS OF COMMUNICATION IN HIV-1 PROTEASE

The allosteric information transfer is fundamental to function and biological role of proteins (Changeux and Edelstein, 2005). This communication describes events where a signal at one region of a protein affects other distant regions in the protein via conformational changes (Tang et al., 2007). In a recent review, it is emphasized that allostery may not necessarily involve a change in backbone shape that leads to creation of new conformational species, but rather it leads to a change in their relative concentrations (Tsai et al., 2008). Yet, it must be noted that side-chain conformational changes, which could be important even in the absence of changes in backbone, are not taken into account in Tsai et al.'s review. Long-range interactions of residues are important in protein's binding processes and distant residues participating in substrate recognition control the structure or activity of the substrate binding site (Sel et al., 2003; Tsai et al., 2008). Proteins sample an ensemble of conformations as a result of their intrinsic dynamics and the ligand binds to a conformation that is optimal for interaction (Chennubhotla et al., 2008). The interactions between the protein and its ligand often induce local energetic and conformational changes at the binding site that subsequently propagate in a cooperative manner through the protein to produce collective conformational responses at distal regions (Ota and Agard, 2005; Chennubhotla et al., 2008). The changes in structure as a result of these cooperative changes lead to new functional states stabilized by rearrangement of intra- or inter-molecular interactions (Chennubhotla et al., 2008).

To be able to study the communication pathways in HIV-1 protease complex structure, a computation method is designed: The structure is considered a network of residues, where the extent of the interaction of each residue with others are determined as based on a scoring function that assigns a weight for each interaction. For the scoring function, two approaches are implemented. The first approach uses the correlations between the fluctuations of residues predicted by the coarse-grained elastic

network model, namely GNM (Bahar et al., 1997a; Haliloglu et al., 1997). The second approach uses the connectivity reflected by a modeled residue-specific Van der Waals potential function. Using these dynamic and energetic correlations, pathway analysis is performed to generate a network of interactions within the HIV-1 protease structure that could be plausible for its function. The networks are then visualized by the Pajek software (Batagelj and Mrvar, 1998).

5.1. Generation of Pathways

This analysis searches through the vast network for the most probable pathways of communication. The pathways are generated by a Monte Carlo approach, in which a probabilistic generation method based on random numbers is used. The generation of any given pathway starts with the identification of possible pairs; a possible pair represents a possible step that could be taken at a certain point on the pathway.

The generation of the pathways in detail is as follows: An NxN matrix, where N is the number of residues, is generated with the interaction values of each residue (Figure 5.1).

E	Glu	Arg	Val	Val	·	·	·
Glu	0	2.17	−0.07	−0.63			
Arg	2.17	0					
Val	−0.07		0				
Val	−0.63			0			
	·				·		
	·					·	
	·						

NxN

Figure 5.1. An example of an interaction matrix for a system of N residues

Then, a probability (weight) matrix (Figure 5.2) is constructed with all the elements in the interaction matrix. The diagonal elements of the probability matrix are set to zero, to prevent revisit of the starting residue along the pathways.

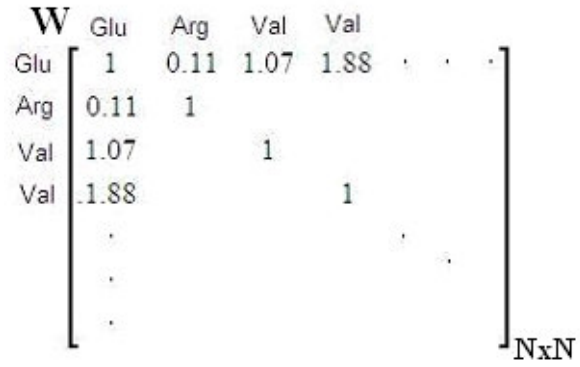


Figure 5.2. An example of a probability matrix for a system of N residues

The conditional probability matrix (Figure 5.3) is then generated by normalizing the rows in the probability matrix by

$$P(ij) = \frac{W(ij)}{\sum_{j=1}^{N} W(j)} \tag{5.1}$$

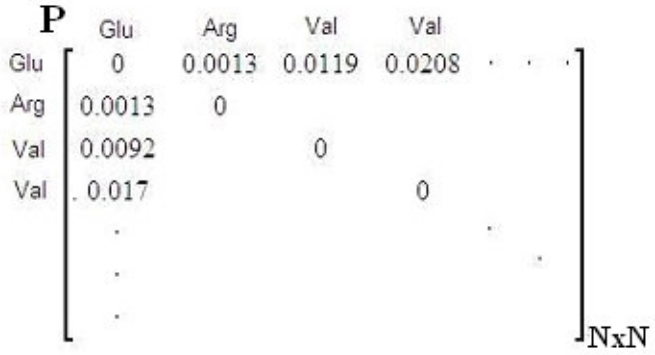


Figure 5.3. An example of a conditional probability matrix for a system of N residues

The row of the starting residue in the conditional probability matrix is selected and the probability values in that row are added from left to right in order to obtain the cumulative values that produce the probability distribution. Then, a random number between 0 and 1 is generated, and the residue with the range where the random number fits is selected as the residue in the next step.

Since Monte Carlo path generation is a probabilistic method, paths consisting of different lengths and different residues are generated. Several runs are carried out

to generate ensembles of pathways that represent the population, and the dominant pathways as well as the frequency of residues visited along the pathways are elaborated.

5.1.1. Prediction of Communication Pathways by GNM

The cross-correlations of residues predicted by the coarse-grained Gaussian Network Model (GNM) (Bahar et al., 1997a; Haliloglu et al., 1997) includes essential information about the coupled motions of molecular regions. It's possible to analyze the relations between distant and close regions using the cross correlation map; but the vast amount of information present about the motions makes it difficult to disclose a network of allosteric signals between remote residues. The pathways of communication here are generated using the cross-correlations between the fluctuations of atoms predicted by GNM, i.e. the probability matrix in the initial step of pathway prediction algorithm directly involves the cross-correlation values.

5.1.2. Prediction of Communication Pathways by Residue-Specific Potentials

A "simplified van der Waals" calculation is carried out for estimating the interaction energies between the atoms of the system. 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, as

$$E(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{5.2}$$

where ϵ is the well depth, σ is the collision diameter and r is calculated with the coordinates of each atom by

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{5.3}$$

While this potential works well in computing the interactions in a molecular

dynamics simulation, the repulsive term is too restrictive in assessing particular interactions of an experimentally determined crystal structure, where slight changes in position can cause a favorable interaction to be considered unfavorable. Thus, the attractive potential is retained and the repulsive potential is removed in the interaction energy calculations (Figure 5.4).

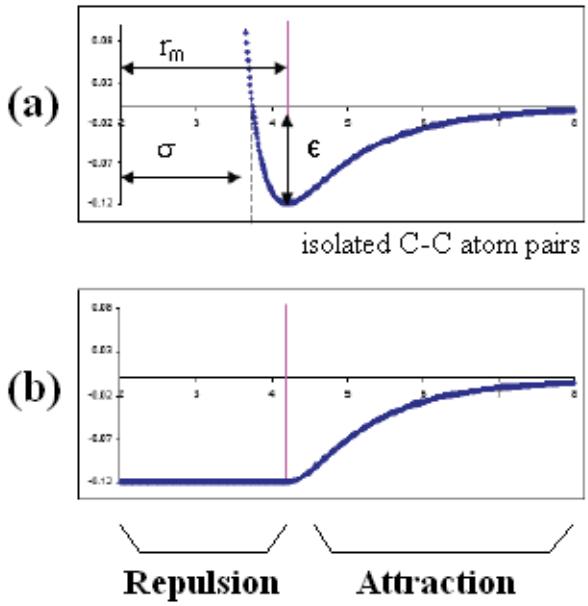


Figure 5.4. Plot of (a) Lennard-Jones potential function, (b) the simplified function.

The interaction matrix constructed in the initial step of pathway prediction algorithm here involves the interaction energies calculated for each atom of each residue. The probability matrix is constructed through $e^{-E_{ij}}$ with all the elements in the energy matrix. Due to the minus sign in the exponential, negative energy values which are more favorable get higher weights whereas the positive energy values get lower weights.

5.2. Pathway Analysis by GNM

Proteins are engaged in functional motions, and interactions, both within and between molecules. These motions can range from local motions, such as single amino acid side chain reorientations, to large scale global motions, such as domain-domain movements. Elastic network models, based on polymer mechanics, are successful in explaining the global motions (Cui and Bahar, 2005). GNM, which assumes isotropic fluctuations in the neighborhood of a single energy minimum (Bahar et al., 1997a;

Haliloglu et al., 1997), have been widely used to explain the collective dynamics of proteins. These collective motions can also determine communication patterns that are characteristic to the native framework of the protein structures (Chennubhotla and Bahar, 2007b; Chennubhotla et al., 2008). The lowest frequency global modes, i.e. the most cooperative collective motions that recruit a large number of residues and potentially play a role in accessing the functional substates, have an active part in facilitating allosteric communication (Bahar et al., 2007; Chennubhotla and Bahar, 2007b; Chennubhotla et al., 2008). Moreover, the role of residue fluctuations and cross-correlations are important in transmitting information; they are shown to be highly correlated with the communication pathway lengths in previous works (Atilgan et al., 2007; Chennubhotla and Bahar, 2007b).

Here, the average of the slowest five cooperative modes of the substrate bound HIV-protease structure (ca-p2) is analyzed for the key interactions of information flow across the network. The approach here uses the positive correlations within the structure in generating the communication pathways by a Monte Carlo path generation method (see Methods section). The mean-square fluctuations in the slowest five modes for the HIV-1 protease complex structure displayed in Figure 5.5 shows that the minima in both monomers of HIV-1 protease refer to residues 8-10 (dimerization region), 25-27 (active site), 45-55 (flaps) and 80-90 (substrate cleft). These minimum fluctuating sites refer to the hinge regions (flexible joints), which are revealed to act as messengers in the transmission of allosteric signals in recent works (Chennubhotla and Bahar, 2006; Bahar et al., 2007). The cross-correlation map for the slowest five modes is displayed in Figure 5.6. The positively correlated regions between the two protease monomers and those between the substrate peptide and protease here also correspond to the hinge regions noted in the slow mode profile (Figure 5.5).

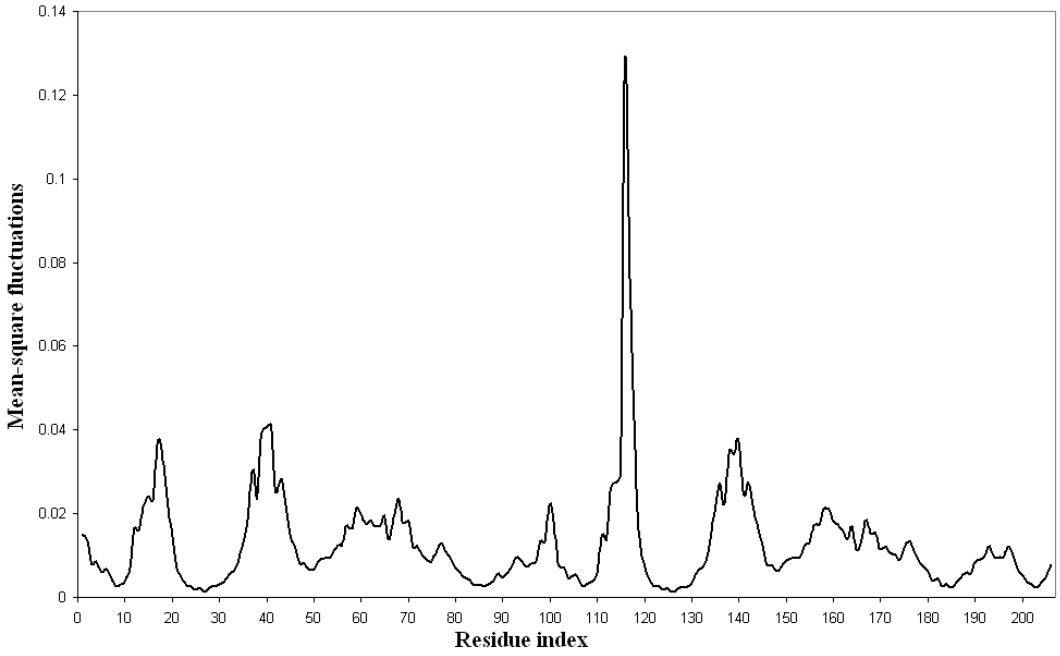


Figure 5.5. GNM slow mode profile for the most cooperative five modes. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

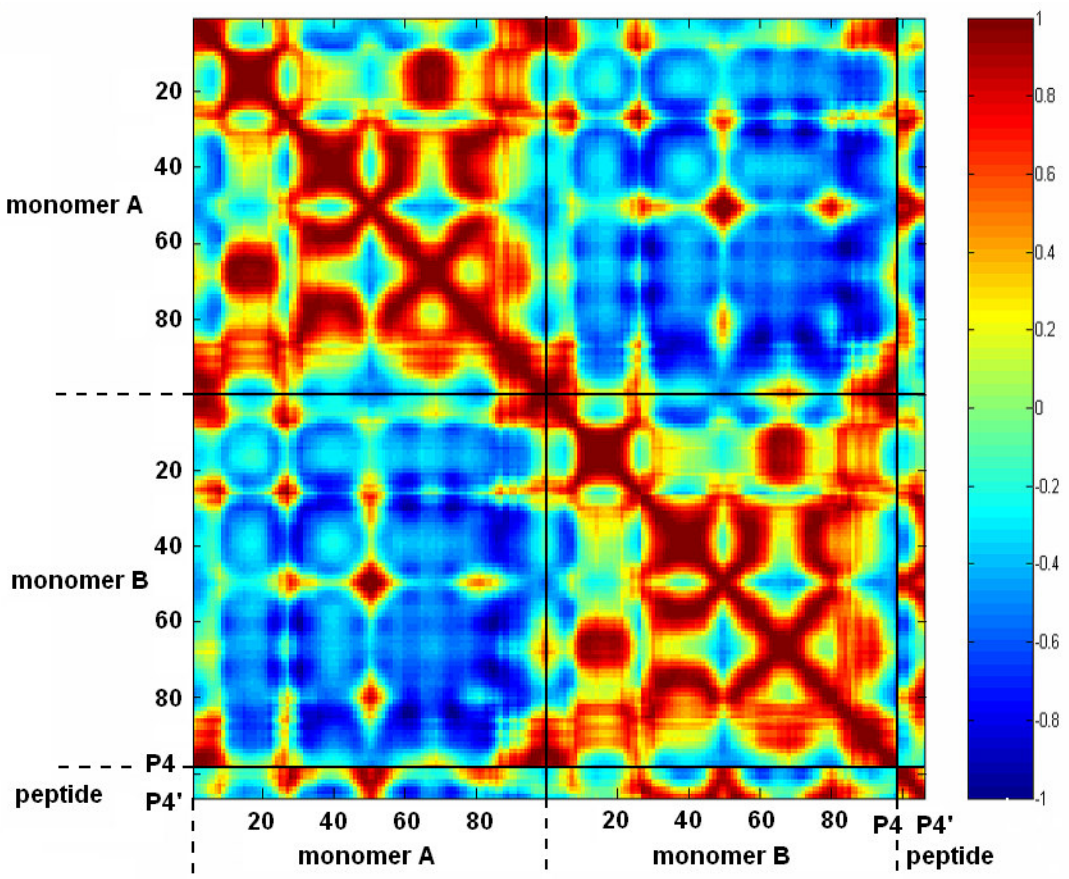


Figure 5.6. GNM cross-correlation map for the slowest five cooperative modes

5.2.1. Short Pathways starting at the Substrate

The pathways of communication between the substrate and the hinge regions of both monomers of ca-p2 complex structure of HIV-1 protease are analyzed. The residues visited on the second, third and fourth steps when the starting region of the pathways is the substrate (P4-P4' sites) are displayed on Figures 5.7-5.9 respectively. The hinge regions, 8-10, 25-27, 45-55 and 80-90, are visited most frequently on these short paths. However, the regions visited become more widespread on the fourth step and revisit of starting points cannot be observed in two steps, therefore the optimum number of steps for the analysis of shorter paths is taken as three.

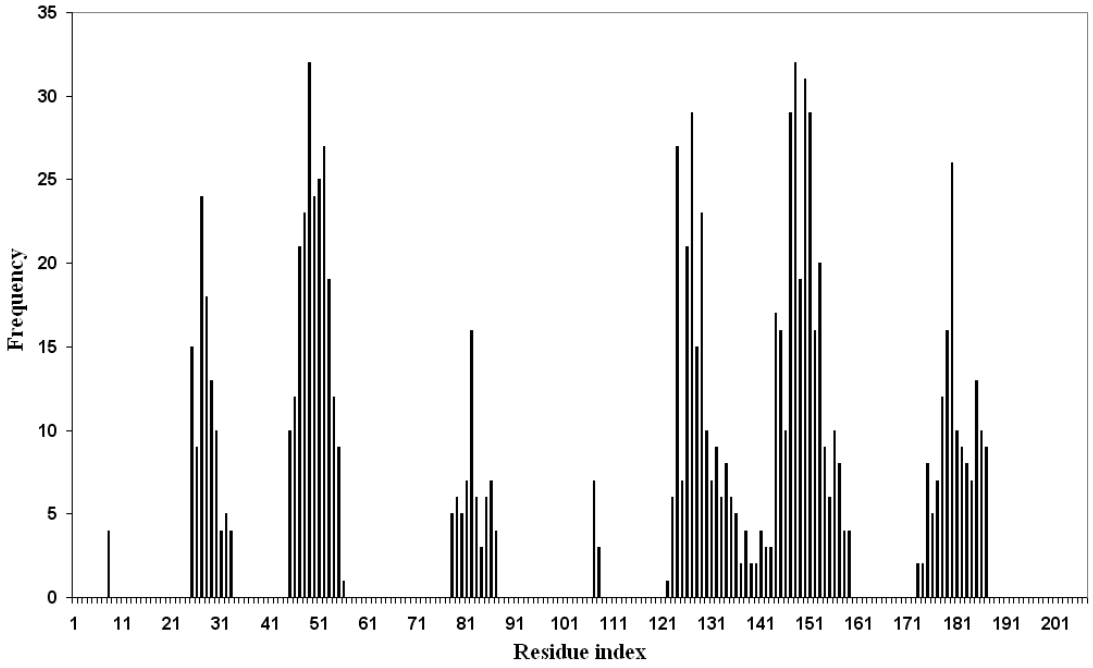


Figure 5.7. Frequency of residues visited at the second step on 1,000 paths starting at the substrate. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

5.2.2. Short Pathways starting at the Protease

The destinations at the third step when the pathways are started from different regions on the protease dimer, from both the hinge regions and the high fluctuating regions of both monomers, are analyzed. Figures 5.10-5.15 display the residues on the third step when the starting region involves the hinges, namely the active sites

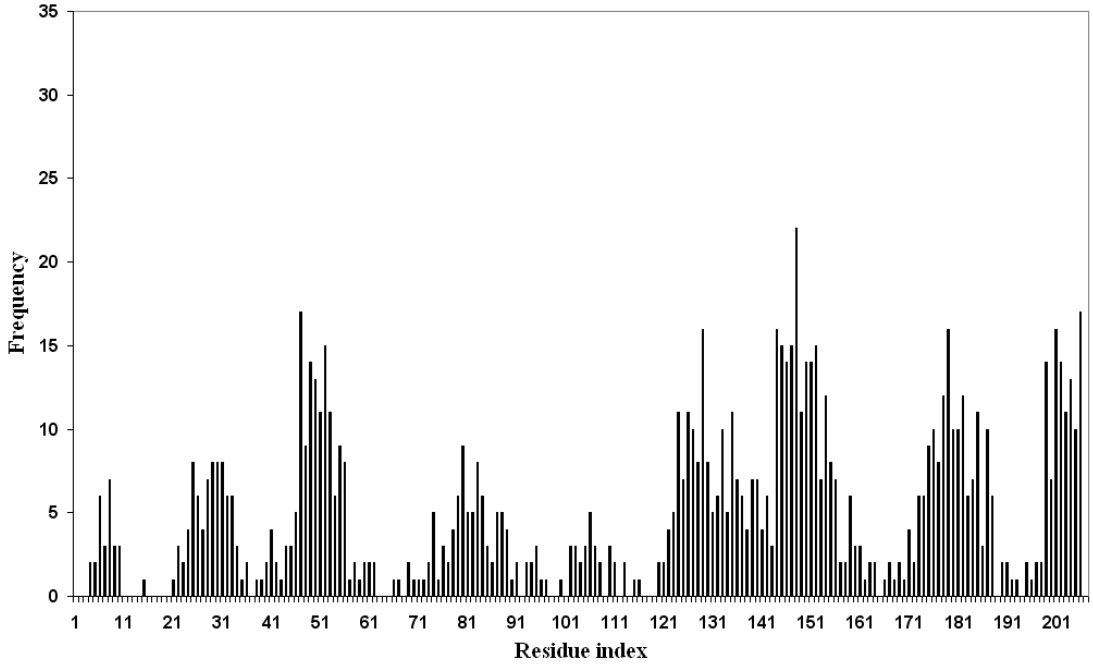


Figure 5.8. Frequency of residues visited at the third step on 1,000 paths starting at the substrate. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

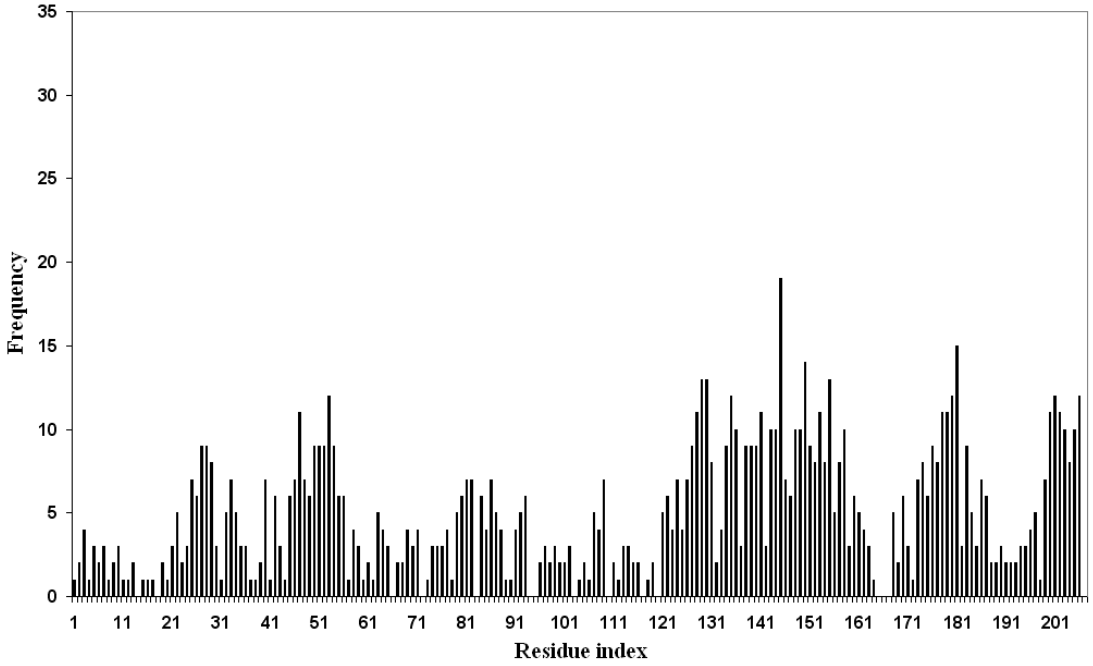


Figure 5.9. Frequency of residues visited at the fourth step on 1,000 paths starting at the substrate. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

(25-27), flaps (46-55), substrate clefts (80-90), dimerization sites (8-10), and two high fluctuating regions 15-18 and 35-38, successively. The paths starting from the active sites of monomer A and monomer B are displayed on the two panels in Figure 5.10; they can reach nearly all the residues on the same monomer of the starting region, but the destinations on the other monomer are only either the hinge regions or the substrate sites in three steps.

The pathways that start from the flap regions reach either the hinge regions of both monomers or the substrate sites in three steps (Figure 5.11). Yet, the regions reached on monomer B when the paths are started from the flaps of monomer A (Figure 5.11a) and the regions reached on monomer A when the paths are started from the flaps of monomer B (Figure 5.11b) are almost equal; i.e. communication can involve both monomers' residues equally. The dimerization region (8-10) is not involved as frequently in the third step of these pathways compared to the ones starting from the active sites.

The behavior observed in the pathways starting from the substrate cleft regions (Figure 5.12) is similar to those starting from the active sites; they can reach nearly all residues in the same monomer as the starting region in three steps and the destinations on the other monomer are only either the hinge regions or the substrate. Yet, the number of paths that involve the other monomer's hinge regions as destination points is fairly lower than the ones reaching the hinge regions of the starting monomer.

The pathways starting from the dimerization regions can reach the hinge regions of both monomers or the substrate sites in three steps (Figure 5.13). Yet, the flaps are not involved as frequently in the destination regions of these pathways; as similarly as the dimerization region is not involved in the destination of the paths starting from the flaps.

When the starting points of the pathways are the high fluctuating regions (Figures 5.14 and 5.15), the destination points are predominantly on the same monomer as the starting region. The pathways that start from residues 15-18 of one monomer never

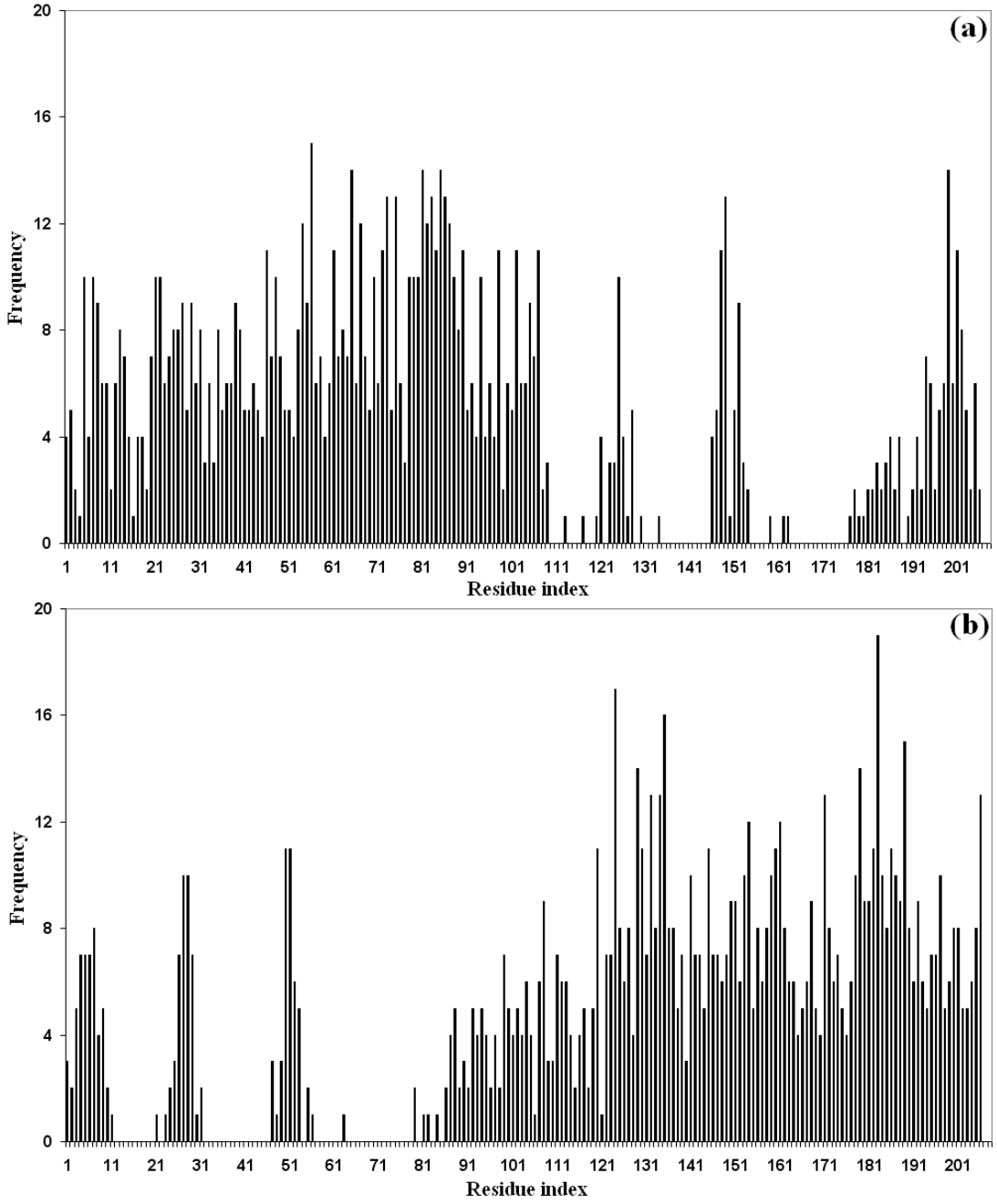


Figure 5.10. Frequency of residues visited at the third step on 1,000 paths starting at (a) the active site of monomer A (residues 25-27) (b) the active site of monomer B (residues 124-126). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

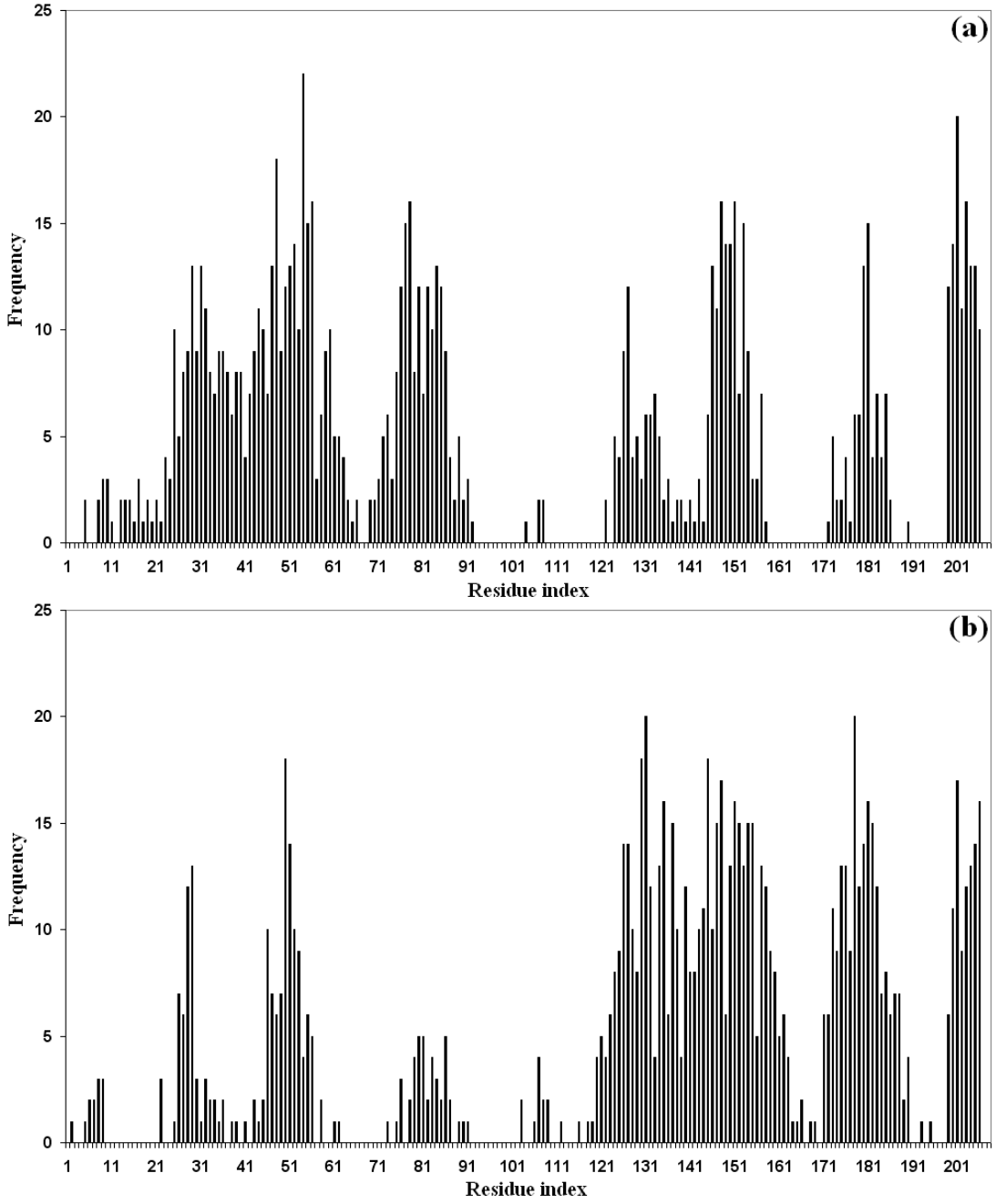


Figure 5.11. Frequency of residues visited at the third step on 1,000 paths starting at (a) the flap of monomer A (residues 46-55) (b) the flap of monomer B (residues 145-154). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

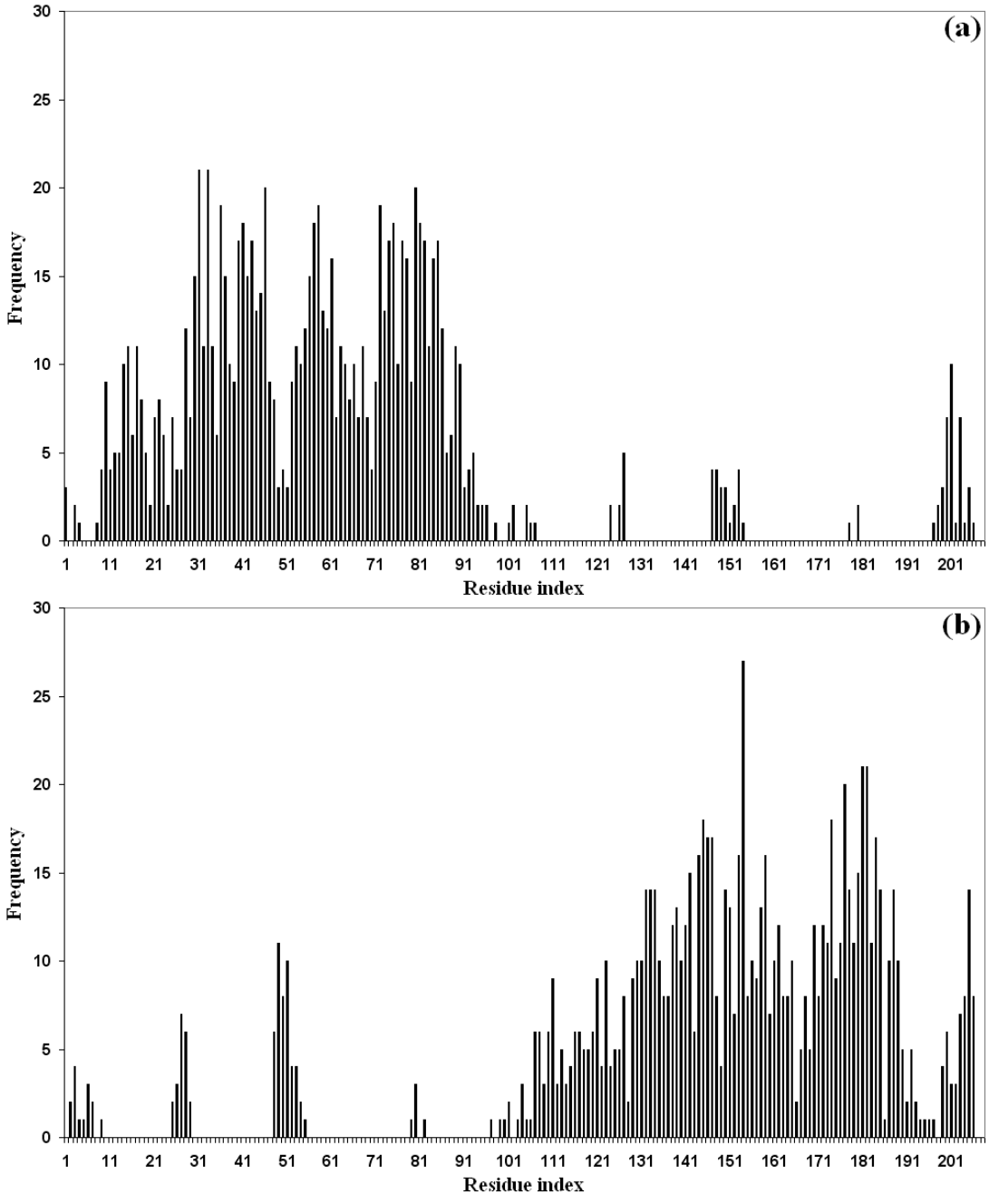


Figure 5.12. Frequency of residues visited at the third step on 1,000 paths starting at (a) the substrate cleft of monomer A (residues 78-85) (b) the substrate cleft of monomer B (residues 177-184). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

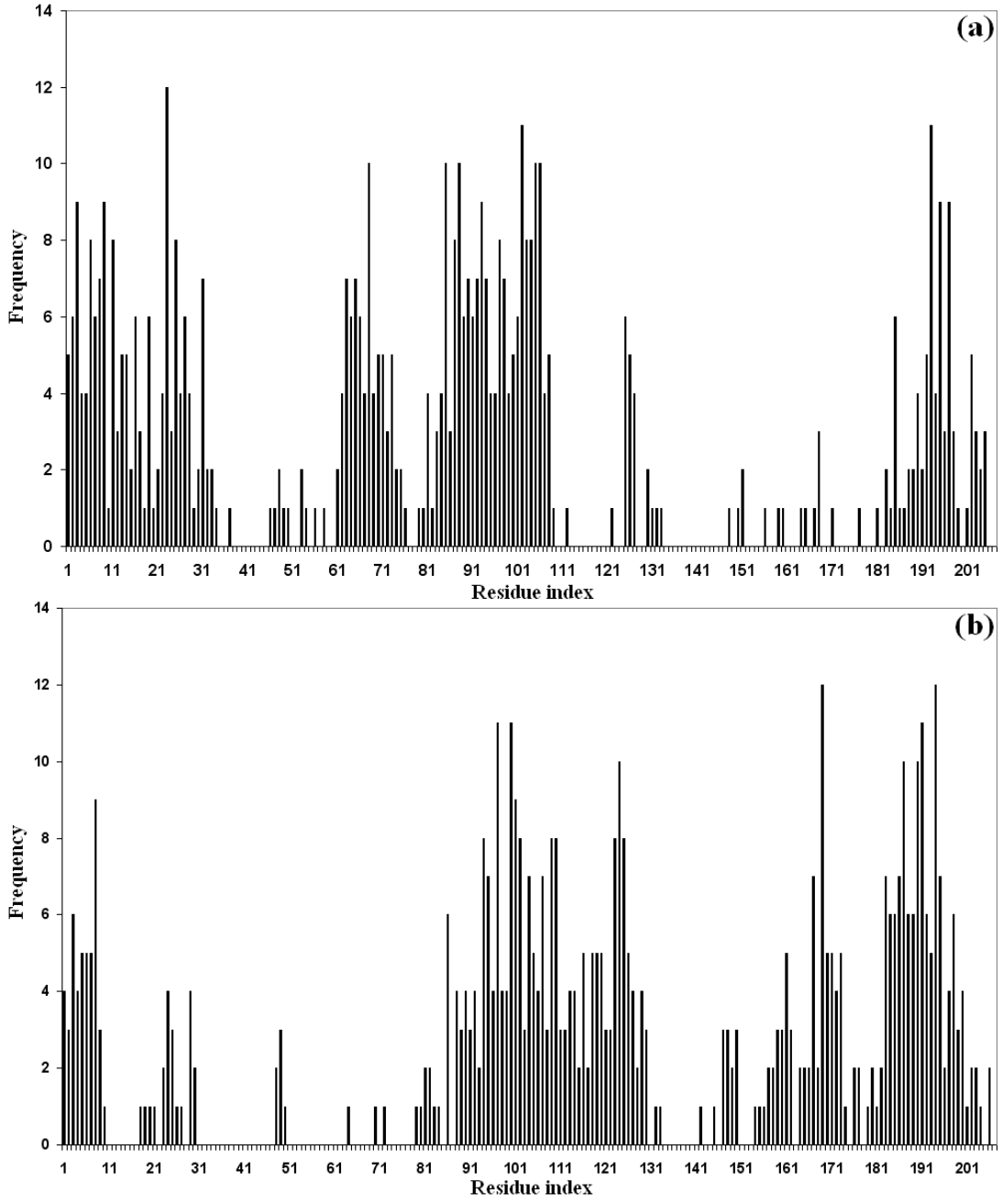


Figure 5.13. Frequency of residues visited at the third step on 1,000 paths starting at (a) the dimerization region of monomer A (residues 8-10) (b) the dimerization region of monomer B (residues 107-109). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

reaches the other monomer, yet the destination region can be one of the substrate sites in three steps (Figure 5.14). On the other hand, the three step pathways that start from residues 35-38 of one monomer can reach the substrate or the flap region of the other monomer as well (Figure 5.15).

5.2.3. Short Pathways starting at the Substrate and reaching Specified Regions of the Protease

Figures 5.16-5.21 display the cumulative frequency of residues visited in three steps when the starting point is any of the eight substrate sites. The residues visited along these three step pathways when the destination regions are specified as the active sites (25-27), flaps (45-55), dimerization regions (8-10), clefts (80-90) and a high fluctuating region of residues 15-18 of both monomers are given in Figures 5.17-5.21 successively. The frequency values on the y-axis vary according to the number of residues in different destination regions. All the paths starting from the substrate visit the flexible joints most frequently. However, the number of residues visited in monomer B is higher than the number of those visited in monomer A, especially in the region of residues 25-55. On the other hand, it is also noticed that the frequency of the residues visited in the region 80-90 is higher in monomer B than in monomer A (Figure 5.16).

When the destination points are specified and the frequencies of the residues visited along the three step pathways starting from the substrate are analyzed accordingly, the hinge regions are the regions mostly involved in communication even if the destination region involves high fluctuating residues. Yet, if the destination region is one of the flexible joints (Figures 5.17-5.20), the regions visited along the paths are denser in the hinge regions than they are in the high fluctuating region 15-18. Nearly all the residues are visited at least once along the paths that reach the high fluctuating region 15-18 (Figure 5.21). The higher number of residues visited in monomer B compared to the ones in monomer A applies here in the paths with specified destination points as well. Particularly in the paths that reach the flaps and the clefts, the number of residues that are visited in flaps and clefts of monomer B is quite higher than that of monomer A.

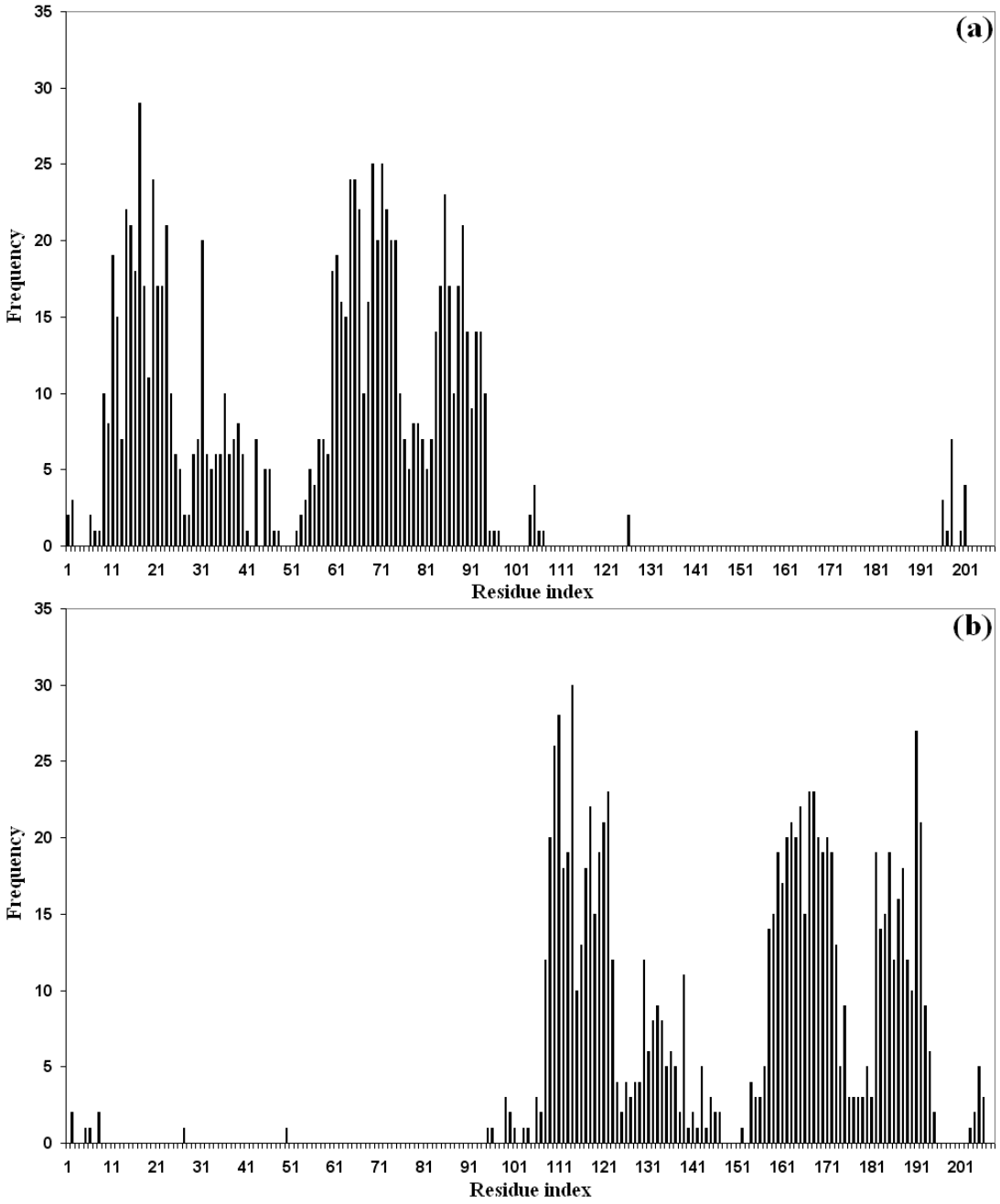


Figure 5.14. Frequency of residues visited at the third step on 1,000 paths starting at (a) the high fluctuating residues 15-18 of monomer A (b) the high fluctuating residues 114-117 of monomer B. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

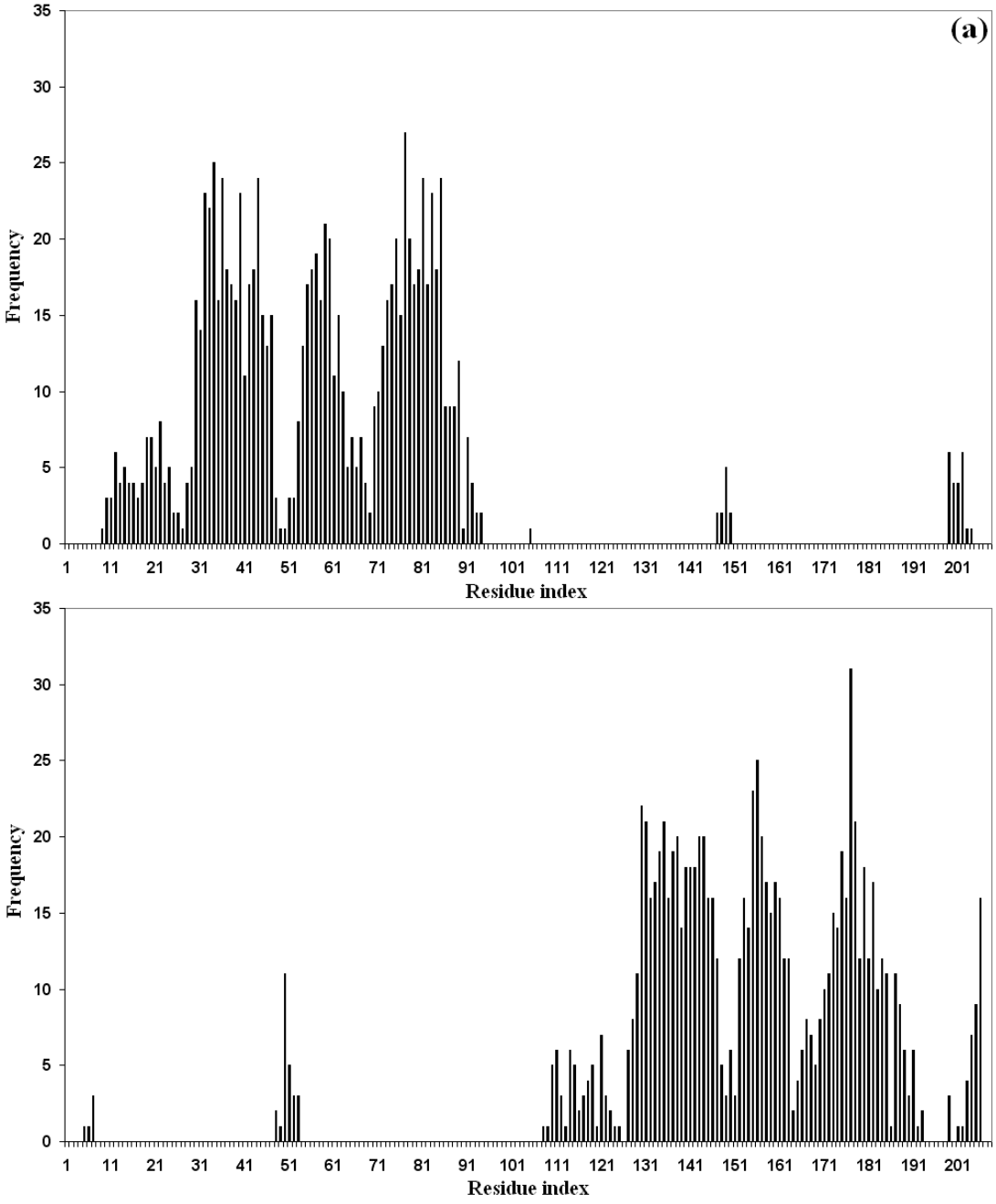


Figure 5.15. Frequency of residues visited at the third step on 1,000 paths starting at (a) the high fluctuating residues 35-38 of monomer A (b) the high fluctuating residues 134-137 of monomer B. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

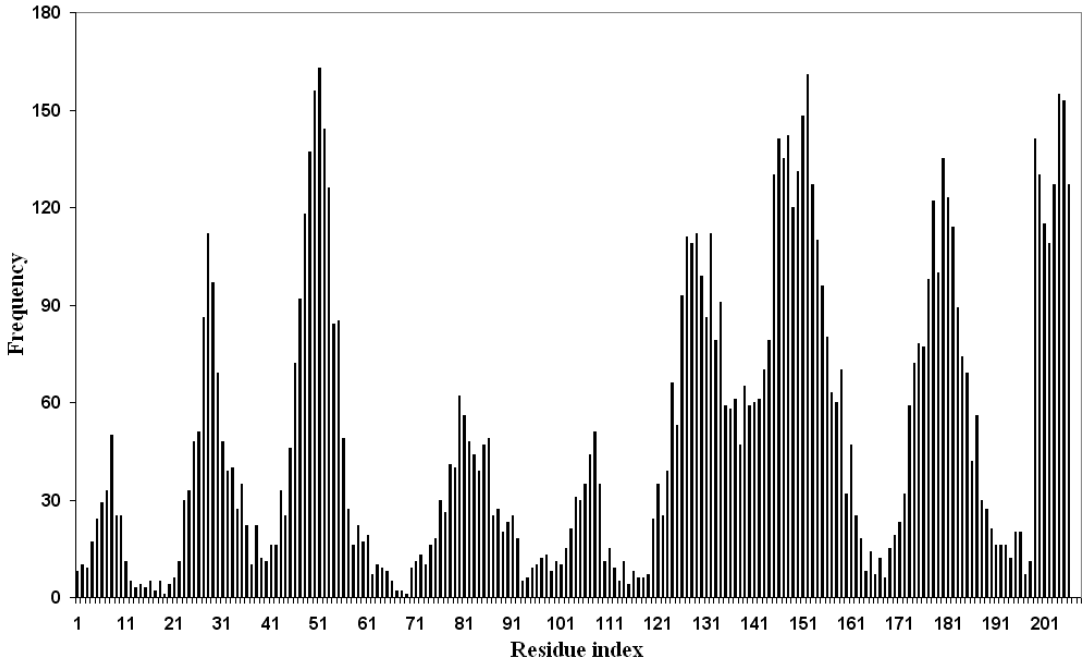


Figure 5.16. Frequency of residues visited in three steps on 10,000 paths starting at the substrate. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

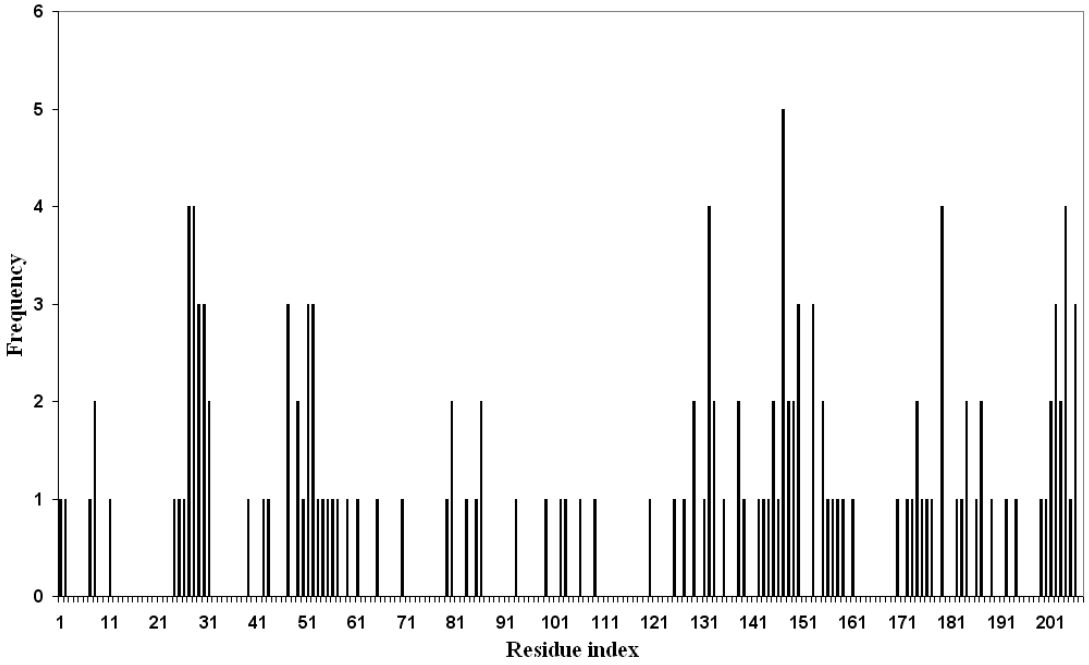


Figure 5.17. Frequency of residues visited in three steps on 10,000 paths starting at the substrate and reaching the active sites (residues 25-27). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

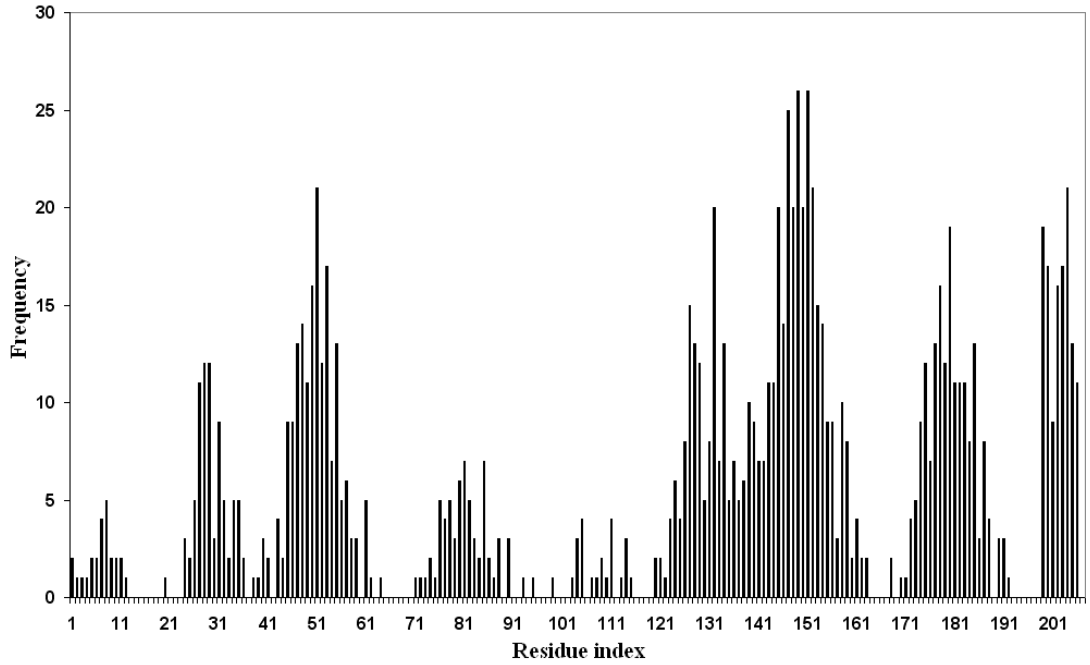


Figure 5.18. Frequency of residues visited in three steps on 10,000 paths starting at the substrate and reaching the flaps (residues 45-55). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

In general, the short pathways starting from the hinge regions or the substrate sites terminate either at the hinge regions of both monomers or at the substrate, whereas the communication starting from the highly mobile sites of one monomer remains within the same monomer. Hence, intermolecular communication is slower than intramolecular communication. Moreover, the time of communication within the core regions such as the active sites or the other substrate cleft residues is shorter than that when the solvent-exposed mobile regions are involved. The active residues are previously shown to be effective in communication (Chennubhotla and Bahar, 2007b); the pathways generated based on the correlations by GNM here also involve the active site residues most frequently. On the other hand, the negatively correlated regions between the two monomers (Figure 5.6), which are also the high fluctuating regions (Figure 5.5), are not observed frequently in the destination points of the short paths of three steps. This also agrees with the statements in Chennubhotla and Bahar's work (2007b), where the residues subject to large amplitude fluctuations and the anticorrelated residues are shown to increase the communication time.

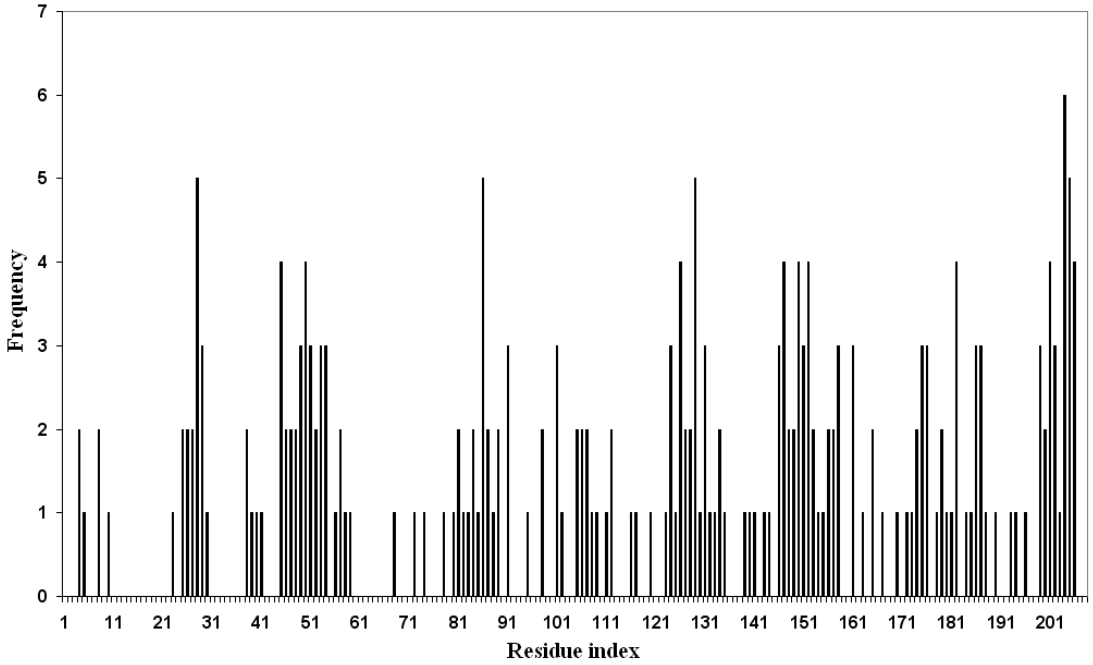


Figure 5.19. Frequency of residues visited in three steps on 10,000 paths starting at the substrate and reaching the dimerization regions (residues 8-10). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

5.2.4. Network Communication between Substrate and Active Sites

The network communication between the active sites of HIV-1 protease and the substrate sites, which is constructed using the pathways generated with the positive correlations of residues predicted by the GNM, is also visualized by the Pajek software (Batagelj and Mrvar, 1998). The residues visited along the three step pathways of communication between the eight substrate sites and the active site of protease monomer A and monomer B are displayed in the panels of Figures 5.22 and 5.23 successively. Thicker lines in the figures represent higher frequency of the corresponding interaction. The residues in the specific regions (active site (25-27), flap (45-55), cleft (80-90), dimerization regions (8-10) and N- and C- termini) of the protease are grouped together in the figures. These hinge regions are noticed as the most visited regions on the pathways and there are links between the two monomers. Yet, hinge regions of monomer B are mostly visited in three steps, no matter whether the pathways are started from the primed or unprimed site of the substrate.

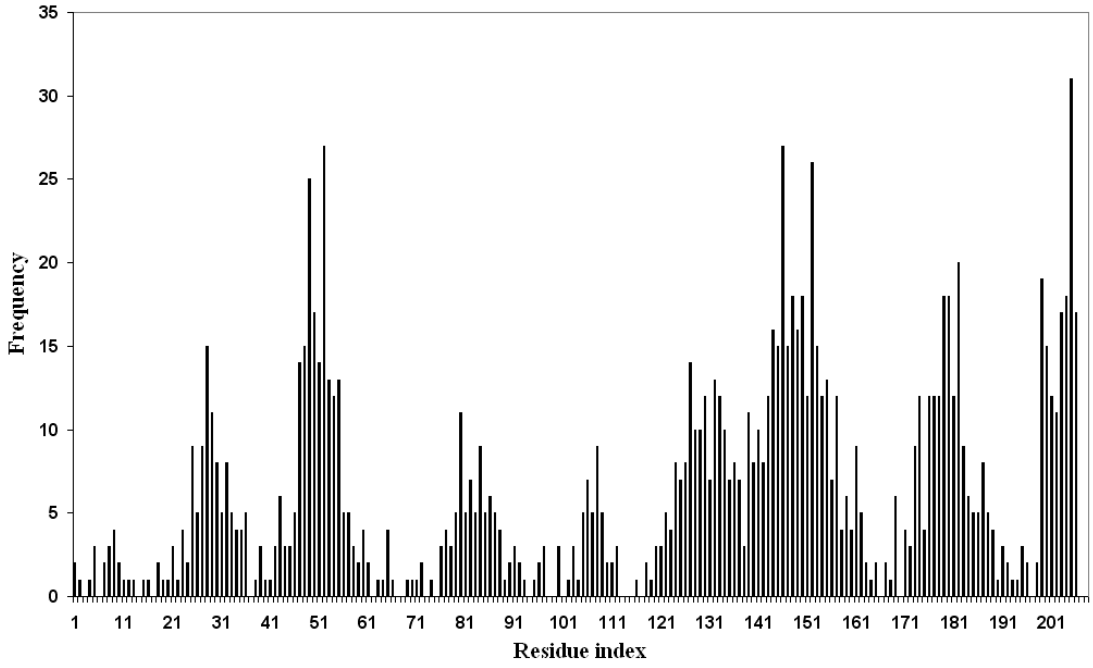


Figure 5.20. Frequency of residues visited in three steps on 10,000 paths starting at the substrate and reaching the substrate clefts (residues 80-90). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

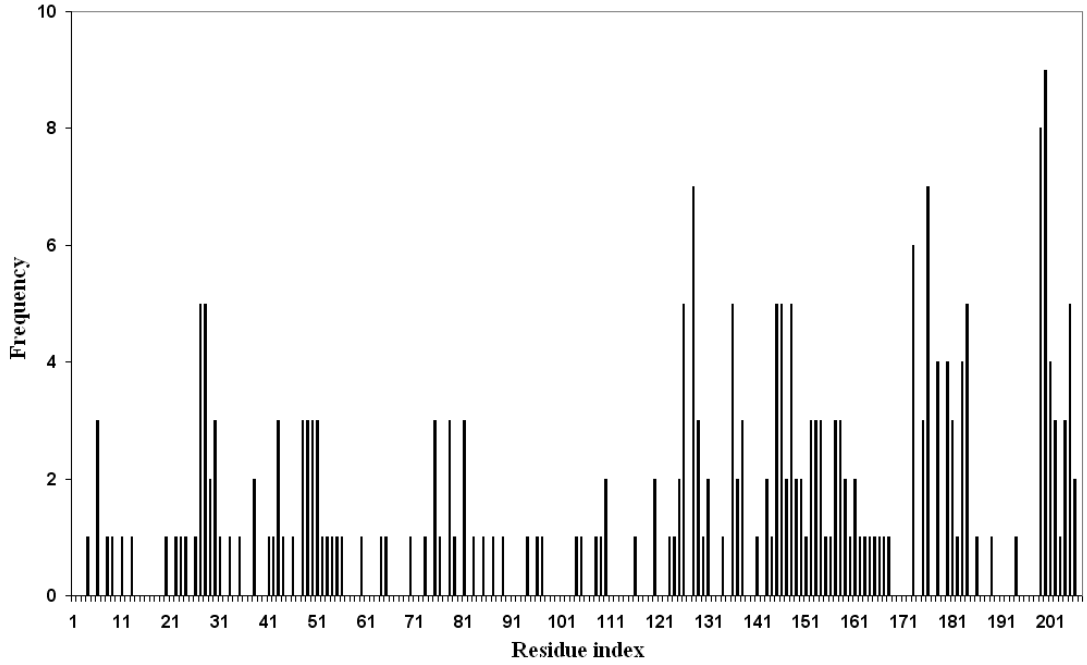


Figure 5.21. Frequency of residues visited in three steps on 10,000 paths starting at the substrate and reaching the high fluctuating residues 15-18. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

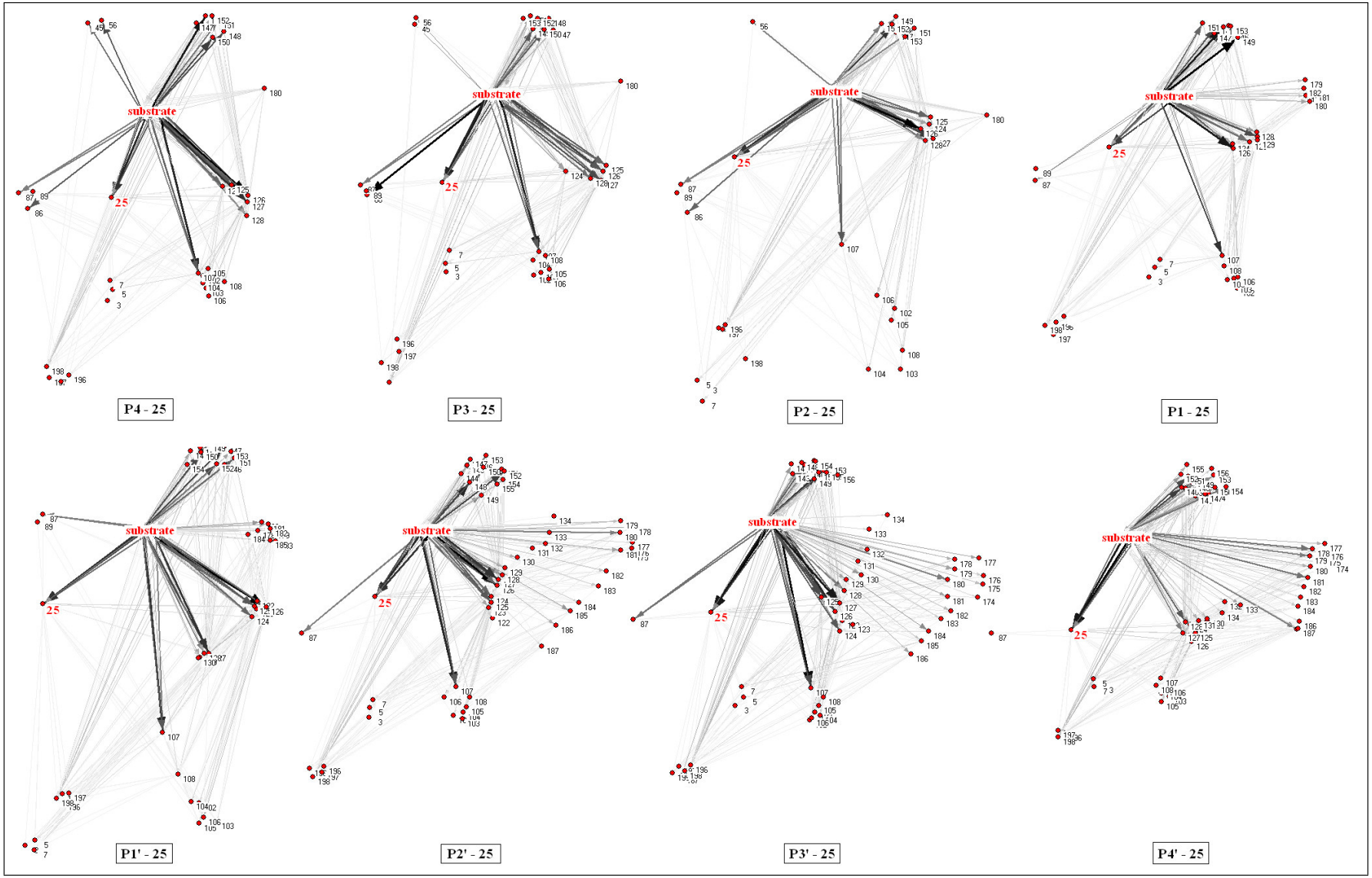


Figure 5.22. Network of interaction between the substrate sites and the active site of protease monomer A

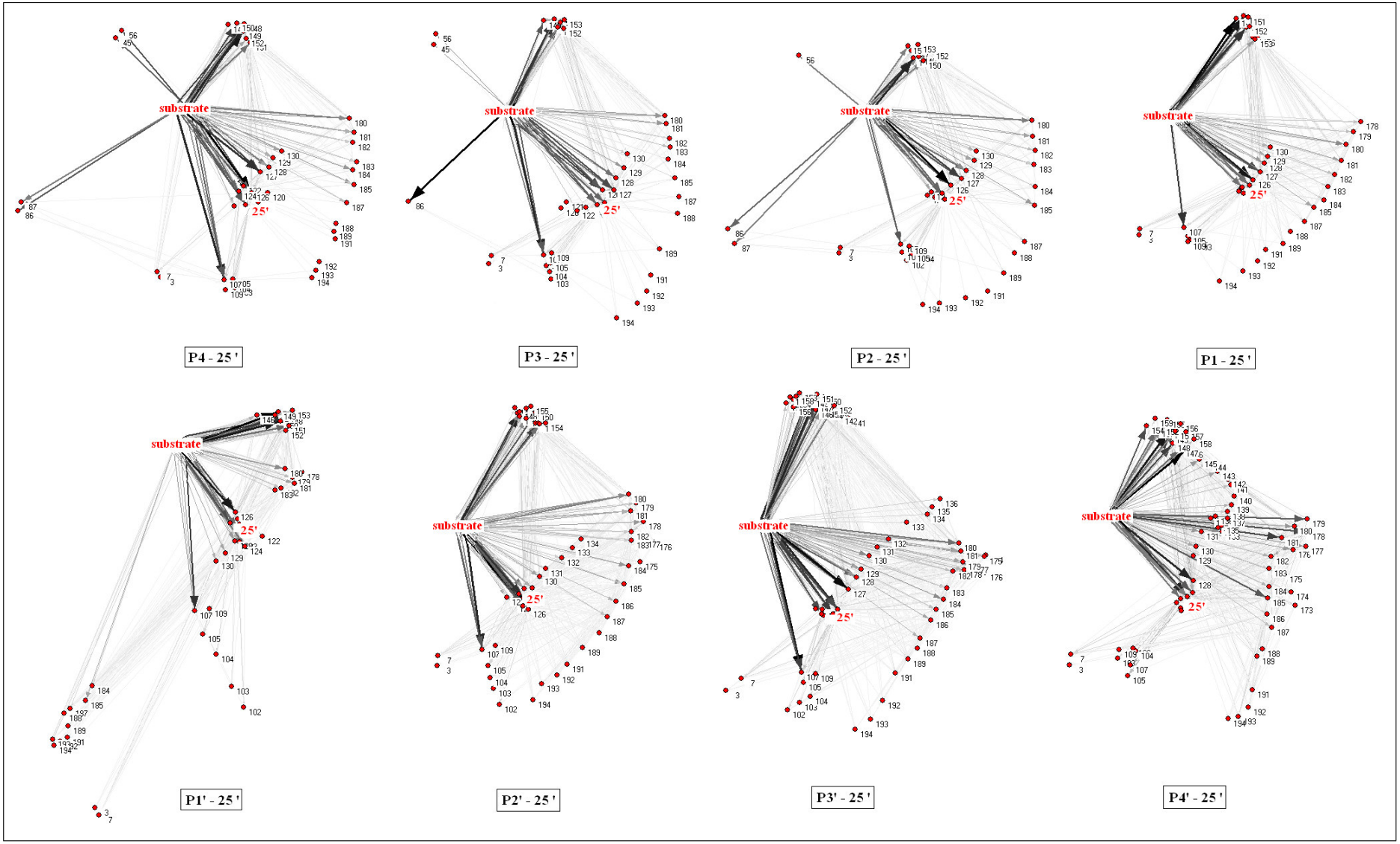


Figure 5.23. Network of interaction between the substrate sites and the active site of protease monomer B

5.3. Pathway Analysis by Residue-Specific Potentials

The second approach to generate the pathways of communication in complex structures of HIV-1 protease uses energetic correlations estimated by a simplified Van der Waals potential function. With this residue-specific approach, the identities of individual amino acids are considered, that is, sequence specificity is taken into account. Using these energetic correlations, the frequency of the residues visited along the information pathways between the substrate and the protease and the shortest paths of communication between the substrate and the active site residues are elaborated in both natural substrate structures and mutant p1-p6 complex structures.

5.3.1. Short Pathways starting at the Substrate

The destinations at the third step when the pathways are started at the substrate sites are analyzed; Figure 5.24 displays the residues on the third step for ca-p2 complex when the starting region involves all of the eight substrate sites. The only difference noticed when the cleavage site P1 is taken as the starting residue (Figure 5.25) is the frequency of the residues visited at the third step. With any substrate site as starting point, the destination residues at the third step are all residues of hinge regions (8-10, 25-27, 45-55, 80-90) and the substrate itself. Besides, the frequency of these regions in destination points differs in comparing the two monomers. When the pathways starting from all substrate sites are added up (Figure 5.24), asymmetry is observed in the active sites (25-27) and cleft (80-90) regions; the active site and the cleft of monomer B are more visited on the third step. On the other hand, when the destination points of the pathways starting at the P1 substrate site is observed (Figure 5.25), the number of times that the pathways reach the dimerization region (8-10) of monomer B is higher than the number of times they reach the dimerization region of monomer A. Besides, the flap region (45-55) of monomer A is more involved than that of monomer B in the destination point of the paths starting at P1. Meanwhile, it should also be noted that the destination regions at the third step of the pathways starting from the substrate sites in different natural substrate bound HIV-1 protease complex structures are quite similar, both in location and in frequency.

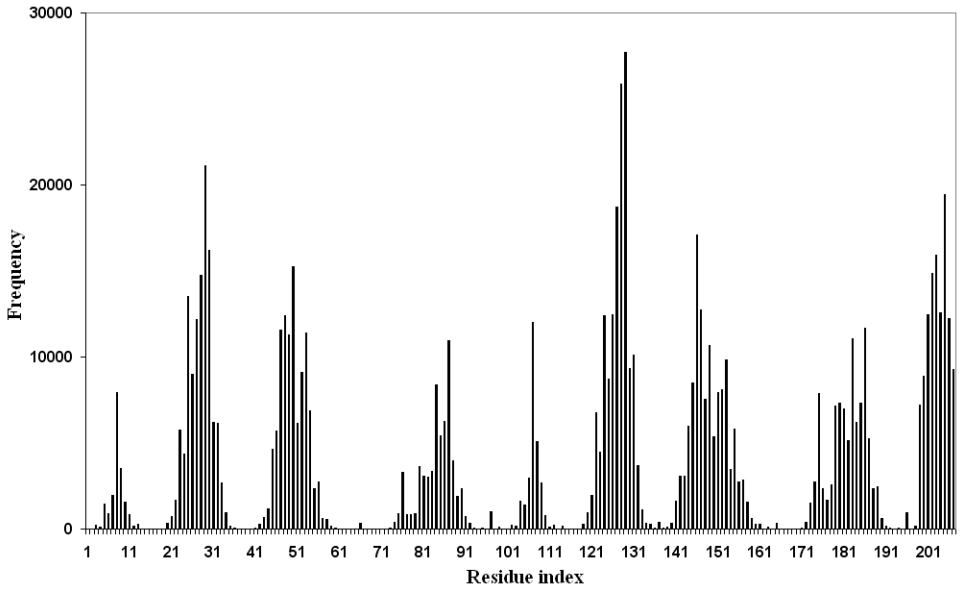


Figure 5.24. Frequency of residues visited at the third step on 100,000 paths starting at the substrate. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

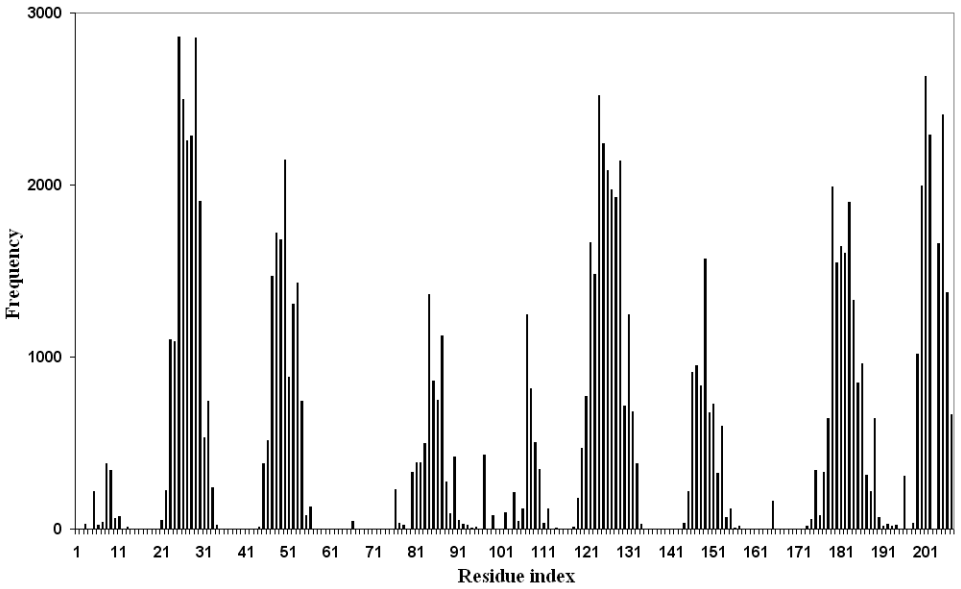


Figure 5.25. Frequency of residues visited at the third step on 100,000 paths starting at P1 site of substrate. Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide. (residue 203 on x-axis)

The observed destination regions in short pathways are similar to those generated by the previous approach that uses the GNM correlations. Nevertheless, taking residue specificity into account provides more pronounced results, i.e. the residues outside the specific hinge regions of the protease are hardly visited. The important role of these hinge regions, which correspond to the minimum fluctuating residues in global modes, is emphasized here; they act as messengers in information transfer.

5.3.2. Pathways starting at the Substrate and reaching Active Sites of the Protease

When the destination point is specified as the active site residue 25 on the three step pathways starting from the substrate, the hinge regions are the regions involved in communication. Figure 5.26 displays the cumulative frequency of residues visited along these pathways when the starting point is any of the eight substrate sites. The residues visited on the three step pathways between the substrate and the active sites of complex structures of different natural substrates are also identical, both in location and in frequency.

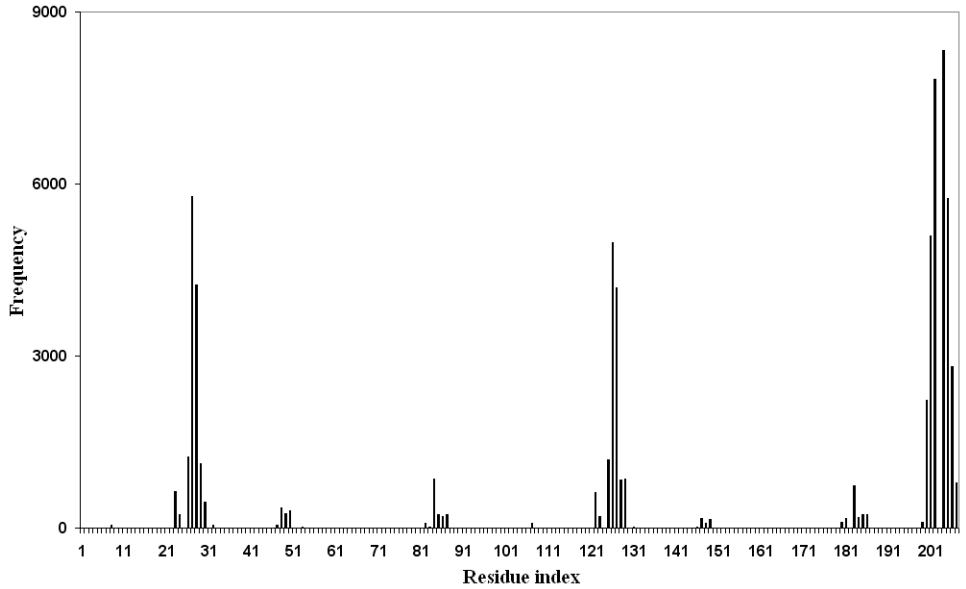


Figure 5.26. Frequency of visited residues in three steps, starting at the substrate and reaching the active sites of both monomers (residues 25 and 124 on x-axis). Residues 1-99 correspond to monomer A, residues 100-198 correspond to monomer B, and residues 199-206 correspond to substrate peptide.

Further, the dominant pathways between the substrate and active sites are analyzed. About 10% of the pathways that start at the substrate and reach the active sites occur more than once. Within these pathways that occur more than once, the maximum number of steps they take to reach residue 25 of either protease monomer is seven. Thus, analysis of three step pathways as the short pathways between substrate and active sites is reasonable. Moreover, about 3% of the pathways are directly reaching residue 25 at the second step. The key interacting residues, i.e. the residues visited along the pathways between the substrate sites and the active sites are 23-30, 48-50 and 82-87. The same residues are observed on the pathways between the substrate cleavage sites and protease active sites in all the natural substrate complex structures. The paths that occur most frequently in the ensemble of communication pathways starting at the substrate cleavage sites and reaching the protease active sites in different natural substrate complex structures are given in Appendix A.

5.3.3. Pathways starting at the Substrate Cleavage Site and reaching Active Sites of the Protease in Mutant Structures

Pathway analysis is performed on the best members of the largest clusters of MD simulated mutant p1-p6 complex structures, namely the D30N mutant, the N88D mutant, the D30N-N88D mutant and the co-evolved D30N-N88D-LP1'F structures. The pathways between the substrate cleavage site P1 and the active site of the protease are analyzed. The average number of steps the paths starting from the substrate take to reach residue 25 is identical in the wild-type and the co-evolved structure. The mutant structures have either longer or shorter paths; this implies adjustment of the wild-type structure through co-evolution. Moreover, the analysis of the most dominant pathways between substrate and active site reveal similarities between the wild-type and the co-evolved p1-p6 complex structures as well. Out of a total of 100,000 paths generated, the number of steps in the pathways that occur more than once is 1035, 1124, 992, 1333, and 1037 in the wild-type, D30N mutant, N88D mutant, D30N-N88D mutant and co-evolved D30N-N88D-LP1'F structures, respectively. Further, the sequence of residues involved in the first four most dominant pathways of the wild-type and the co-evolved structures are exactly identical. The co-evolving P1' site of substrate, is one of the

residues visited along these most dominant pathways. Yet, in the D30N-N88D double mutant structure, this site is visited along the pathways that occur less frequently. The number of pathways involving different residues, which implies variability, is 97, 97, 86, 110 and 87 in the wild-type, D30N mutant, N88D mutant, D30N-N88D mutant and co-evolved D30N-N88D-LP1'F structures, respectively. On the other hand, the number of paths that directly reach residue 25 in the first step is 272, 398, 316, 541, and 311 in the wild-type, D30N mutant, N88D mutant, D30N-N88D mutant and co-evolved D30N-N88D-LP1'F structures, respectively. The lower number of paths that reach the destination point in the first step in the wild-type and co-evolved structures indicate that less residues are visited along these pathways in these structures than the mutant structures. The key interacting residues, i.e. the residues visited along the pathways between the substrate cleavage sites and the active sites of the mutant structures are 8, 23-30, 48-50, 82-87 and 90. The paths that occur most frequently in the ensemble of communication pathways starting at the substrate cleavage sites and reaching the protease active sites in all the mutant complex structures are given in Appendix A.

5.3.4. Key Interactions

The key interacting residues that are found most frequently along the ensembles of the most dominant pathways, and thus that should be crucial in controlling the communication between HIV-1 protease and its substrates, are observed as 25, 26 and 87 (Figure 5.27). Residue 25 and 26 are active site residues that function in ligand recognition (Perryman et al., 2004) and their effectiveness in communication is also shown in the previous section where pathways of communication are generated using the contact topology information by the GNM. On the other hand, residue 87 interacts with residue 25 and also residue 90 whose mutations were reported to confer drug resistance (Wu et al., 2004). These residues are also stated as the conserved interconnectivity determinants that play important roles in information transfer within HIV-1 protease in previous works (del Sol et al., 2006a; del Sol et al., 2006b).

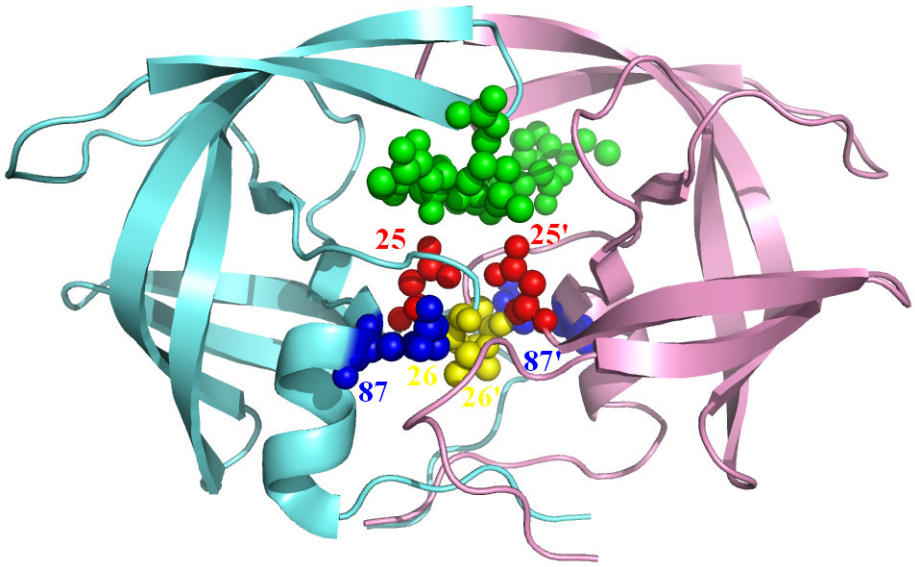


Figure 5.27. The key residues (25, 26, 87) in the shortest pathways of communication between the HIV-1 protease and its substrates

5.3.5. The Shortest Paths

Shortest pathways between the substrate sites and the active sites of both protease monomers are proposed, using the average step numbers that the residues are reached along the pathways and the average path lengths. The estimation of the shortest paths is as follows: Once the pathways are generated, the frequencies of the residues visited along these pathways are calculated. Then, the average step number of each residue is calculated by simply dividing the sum of their step numbers by their frequency. As expected, the average step number of the starting residue is 1. In addition, average path length of every residue is calculated by dividing the sum of the lenghts of the paths on which they appear by their frequency. The average step number and the average path length of the destination residue are equal to each other. Then, the mean and the standard deviation of the average step numbers are calculated. Starting from the mean value, the average step numbers are divided into ranges of one standard deviation length in both positive and negative directions. Later, the residues with the highest frequency value in every range constitute the steps of the so called shortest path between the starting and the destination residues.

The shortest pathways between the substrate sites and the active sites of both

protease monomers estimated for ca-p2 are given in Table 5.1. The shortest pathways estimated for different natural substrate complex structures involve the same residues (Table 5.2). The substrate sites are re-visited along these shortest pathways and there are links between the two protease monomers. The pathways starting from the primed site of the substrate involve residues of monomer B whereas those starting from the unprimed site involve residues of monomer A to reach the protease active sites.

Table 5.1. The shortest pathways between the substrate sites and the active sites of both protease monomers estimated for the ca-p2 complex structure

P4	P3	30.A	87.A	P2'	26.A	25.A
P3	29.A	P2	87.A	84.A	26.B	25.A
P2	P3	84.A	25.B	25.A		
P1	P2	25.B	87.A	25.A		
P1'	P2'	P1	25.B	23.A	89.A	25.A
P2'	30.B	29.B	87.B	26.B	26.A	25.A
P3'	P2'	30.B	87.B	25.B	87.A	25.A
P4'	P2'	76.B	25.B	87.A	25.A	
P4	P3	29.A	87.A	P2'	29.B	25.B
P3	29.A	P2	P1	P2'	87.B	25.B
P2	P1	25.A	30.B	26.B	25.B	
P1	P2	P2'	25.A	23.B	25.B	
P1'	P2'	25.A	87.B	24.B	25.B	
P2'	30.B	29.B	87.B	25.A	26.A	25.B
P3'	P2'	29.B	87.B	P1	24.B	25.B
P4'	P2'	32.B	85.B	26.B	25.B	

The shortest pathways between the substrate cleavage site P1 and residue 25 of monomer A are estimated for the best members of the largest clusters of MD simulated wild-type and mutant p1-p6 complex structures (Table 5.3). Observing similar residues along these shortest pathways as the shortest pathways of p1-p6 crystal structure implies that communication is not affected by conformational changes. This might be due to the involvement of hinge regions as messengers in information transfer.

Table 5.2. The shortest pathways between the substrate cleavage sites and the active sites of both protease monomers estimated for the natural substrate complex structures other than ca-p2

ma-ca	P1	P2	P1'	30.A	26.A	25.A	
nc-p1	P1	P1'	25.B	87.A	85.A	25.A	
p1-p6	P1	P2	P2'	25.B	87.A	23.A	25.A
p2-nc	P1	P1'	25.B	87.A	85.A	25.A	
rh-in	P1	P1'	25.B	29.A	25.A		
rt-rh	P1	P1'	25.B	87.A	23.A	25.A	
ma-ca	P1	P2	25.A	P2'	87.B	25.B	
nc-p1	P1	P1'	25.A	87.B	90.B	25.B	
p1-p6	P1	P2	P1'	25.A	30.B	24.B	25.B
p2-nc	P1	P2	25.A	87.B	25.B		
rh-in	P1	P2'	25.A	87.B	85.B	25.B	
rt-rh	P1	P2	P1'	25.A	26.A	90.B	25.B
ma-ca	P1'	P1	25.B	25.A			
nc-p1	P1'	P2'	P1	25.B	26.B	26.A	25.A
p1-p6	P1'	P2'	P1	25.B	87.A	26.A	25.A
p2-nc	P1'	P2'	P1	P2	87.A	26.A	25.A
rh-in	P1'	P2'	P1	25.B	84.A	90.A	25.A
rt-rh	P1'	P2'	P1	29.B	23.A	26.A	25.A
ma-ca	P1'	P1	25.A	26.B	25.B		
nc-p1	P1'	P2'	P1	25.A	87.B	23.B	25.B
p1-p6	P1'	P1	25.A	87.B	23.B	25.B	
p2-nc	P1'	P2'	25.A	87.B	23.B	25.B	
rh-in	P1'	P2'	P1	P2	29.B	26.B	25.B
rt-rh	P1'	P2'	P1	29.B	87.B	85.B	25.B

Table 5.3. The shortest pathways between the P1 cleavage site and the active site of protease monomer A estimated for the best members of the largest clusters of MD simulated mutant p1-p6 complex structures

wt p1-p6	P1	P2	P1'	25.B	87.A	25.A	
D30N	P1	P2	25.B	84.A	24.A	25.A	
N88D	P1	P2	25.B	26.A	85.A	25.A	
D30N-N88D	P1	P2	P1'	25.B	87.A	25.A	
D30N-N88D-LP1'F	P1	P2	P1'	25.B	87.A	90.A	25.A

The residues visited frequently along the shortest pathways generated are identical with those on the dominant pathways. Yet, the shortest paths estimated are longer than the most dominant pathways. The shorter dominant pathways are probably due to analysing the paths between closely interacting regions; i.e. the substrate sites and the protease active sites are close in space. Thus, the shortest pathway analysis should be improved further with the investigation of more distant interactions.

6. CONCLUSIONS AND FUTURE STUDIES

6.1. Conclusions

Understanding the molecular recognition events that confer drug resistance in HIV-1 protease is crucial to the development of drugs. Combined computational methodologies used in this thesis puts three different perspectives together to study the recognition and binding processes in HIV-1 protease complex structures within the paradigm of sequence, structure and dynamics. In the first part, substrate specificity investigated by a sequence search threading method explores the potential substrate sequence space. In the second part, the analysis of the fluctuations of the ligand bound HIV-1 protease structures identifies the functionally plausible dynamic motion comparatively between different substrate and inhibitor complexes. In the third part, the residue interactions that are possibly crucial in the binding interactions of HIV-1 protease are identified by a communication pathway analysis.

The biased sequence search threading methodology introduced in the first part uses low resolution knowledge-based potentials and efficiently explores the potential substrate sequence space. Different template structures are used to provide a structure space as a base for the differences between the behavior of various substrates. By determining the relationship between the substrate sequence and three-dimensional structure of the protease, it is possible to probe which substrate sequences are more likely to tolerate changes in HIV-1 protease due to drug resistant mutations and which are not. The low energy substrate sequences generated by the biased search are correlated with the natural substrates. Octameric sequences are predicted using the probabilities of residue positions in the sequences generated by BSST in three ways: considering each position in the substrate independently, considering pairwise interdependency and considering triplewise interdependency. The prediction of octameric sequences using the triplewise conditional probabilities produces the most accurate results, implying that there is a complex interdependence between the different substrate residue positions. This likely reflects that HIV-1 protease recognizes the overall shape of the substrate

more than its specific sequence. Overall, the BSST methodology based on low resolution knowledge-based potentials provided a powerful methodology for accurately predicting HIV-1 protease substrate specificity.

In the second part of this thesis, the elaborated analysis of the structural fluctuations of HIV-1 protease in interaction with its substrates and inhibitors enhance the understanding of the dynamics of HIV-1 protease in relation to its function. The analysis of the fluctuations of ligand bound complex structures by atomistic Anisotropic Network Model (ANM) displays that all HIV-1 protease complex structures display similar molecular motion in the low frequency modes that are related to the main function. The minimum fluctuating residues in these modes of motion, i.e. the hinge regions, correspond to the dimerization region, the active site, the flaps and the substrate cleft of the protease, which are also positively correlated with each other. As for the interaction between the protease and the peptide, the residues of the protease that display positively correlated fluctuations with the fluctuations of the peptide's residues also correspond to these hinge regions. That is; the same sites are associated with both dimerization and binding to ligands. Further, despite the similarity in the coopearative modes between the substrate and inhibitor complex structures, the detailed comparative analysis of the direction of the fluctuations of residues between the structures suggested some differences: The substrate and inhibitor structures are observed to gather into two groups each, according to the magnitude of residue fluctuations and orientational correlations of residues. The residues that lead to this grouping of complex structures with respect to their direction of fluctuations lie along the rotational axes around which the protease monomers rotate in the most cooperative modes. The latter analysis also implies that the protease bound to different subtrates and to different inhibitors displays an enhanced orientational space sampled for the former. The extent of the coupling of the protease with its substrates and inhibitors implies tighter binding for the inhibitors. Because of their higher flexibility, the substrates should be more adaptable to backbone rearrangements or conformational changes induced by the protease mutations. Moreover, the examination of the structural and dynamic properties of the mutant and co-evolved structures of p1-p6 substrate contributes to the understanding of the binding as well as the drug resistance mechanism of HIV-1

protease. The higher correlation of the wild-type complex with the co-evolved structure than the other mutant structures, justify the existence of this mutation for the conservation of dynamic fluctuations. The conservation of the direction of the fluctuations, in particular, with respect to the rotational axes in the most cooperative modes justifies the co-evolution. In the third part of this thesis, the network of key interactions within the native topology of HIV-1 is searched by an ensemble of pathways generated by a newly designed computation tool. Focusing on the binding process, an ensemble of pathways of communication is generated between the binding site and substrate sites. The scoring of the interaction between the two residues is carried out by two approaches; using the coupling between the fluctuations predicted by the Gaussian Network Model (GNM) which reflects topological features of the structure with no specificity in interactions, and using the intensity of the interactions based on the residue specific potential functions. The communication, as observed similarly by both approaches, is achieved by the hinge regions (flexible joints) that act as messengers in the information transfer between the residues. Moreover, the time of communication within the hinge residues found in core regions is shorter than that when the solvent-exposed mobile regions are involved. Further, the most dominant pathways estimated between the substrate and protease active sites define the key interacting residues. The active site and the substrate cleft residues in the core regions that either function in ligand recognition or interact with the residues that confer drug resistance are effective in communication. The additional analysis of conformations from molecular dynamics simulations suggest similar communication pathways as the crystal structures, implying that communication is not affected by conformational changes as probably the hinges are involved as messengers. Also, the adjustment of the p1-p6 variant structures through co-evolution is reflected by the similarity in the most dominant pathways of communication in the wild-type and co-evolved complex structures.

With all, the present thesis might contribute both to the overall understanding of the plasticity of the ensemble of ligand bound HIV-1 protease conformations and sequences and to the technology of drug design.

6.2. Future Studies

Very little sequence homology exists between the various substrate sequences in HIV-1 protease and natural variation exists in the substrate sequences between different viral subtypes of HIV (i.e. 1A, 1B, 1C). Altered specificity in different subtypes may indicate protease variants that warrant subtype specific inhibitor, i.e. subtypes which are unlikely to respond to currently available inhibitors, as "non-B" HIV proteases, are more important world-wide. The biased sequence search threading technique introduced in this thesis is applicable for predicting which substrate sites can tolerate the changes and which cannot, by an efficient exploration of the sequence space. Hence, the technique's ability to predict sequence variability can be utilized to examine the natural variation that exists within the subtypes. Yet, minimization of the modeled protease variants on which the sequences of subtypes will be threaded is of importance here, which can be handled by Monte Carlo/Metropolis simulations. Understanding the range of adaptability of the substrate sites within the subtypes should provide an important test and validation of a set of sequences that can then be applied to variations that occur with drug resistance. Further, this general method could also potentially be optimized to predict the substrate specificity of other proteases with complex substrate specificities.

Certain substrate sites are more likely to mutate in response to particular mutations in the drug resistant protease variants. This affects substrate specificity and different substrate sites may be more susceptible to change depending on the protease mutation. An example of this evolutionary communication has already been seen in the co-evolution of the nc-p1 cleavage site where Ala in P2 mutates to Val in response to V82A mutation in the protease associated with IDV or RTV therapy (Prabu-Jeyabalan et al., 2004). The p1-p6 is another cleavage site that undergoes co-evolution with HIV-1 protease; mutations in the P1' site of p1-p6 substrate covary with the D30N/N88D mutations in the protease (Kolli et al., 2006). Despite the knowledge of mutation types, the exact mechanism by which these changes as well as compensatory protease mutations cause resistance or facilitate protease activity is poorly understood. The examination of the structural and dynamic properties of the mutant structures involved

in the co-evolution of p1-p6 with D30N/N88D protease mutations here contributes to the understanding of the binding as well as the drug resistance mechanism of HIV-1 protease. Therefore, the structural analysis performed in this thesis can be utilized to elaborate the co-evolution of the nc-p1 substrate cleavage site with V82A protease mutation. The conformations of mutant and co-evolved variants of p1-p6 substrate complex structures utilized here are obtained from MD simulations of in silico created variant structures. Yet, the crystal structure of AP2V/V82A nc-p1/protease variant has already been solved (Prabu-Jeyabalan et al., 2004) and deposited in the PDB. Thus, MD simulations of this variant structure can be carried out and conformations generated from these simulations can be analyzed with respect to both dynamic fluctuations and ensemble of short pathways. This might help in investigation of the mechanism of drug resistance and co-evolution in the HIV-1 protease system.

The structural fluctuations and orientational correlations of wild-type and mutant HIV-1 protease complex structures are analyzed in the most cooperative modes of the ANM utilized in this thesis. The principal components of the conformations generated by the MD simulations of the same structures are also calculated and their correlation with the ANM modes is studied. The measure of the coverage of the motion subspaces spanned by each approach shows that the majority of the dynamics of protease complex structures can be explained by a small set of low-frequency ANM modes. Nevertheless, it should be noted that this is as much reflected by the PCA of an MD trajectory of a given length. In other words, these ANM modes in principle could represent large scale motions that could not spanned by the 11 ns MD simulations. The length of the MD simulations may not be long enough to define the motion in the most cooperative modes, hence the grouping of substrate and inhibitor complex structures are not observed in the MD simulated structures. Although the present MD simulations still provide assurance and could be complimentary for the dynamic properties that are of interest here, the simulations can be elongated in order to search the conformational space more thoroughly.

The communication analysis by a newly designed computation tool in this thesis is carried out between the binding site and substrate sites of the HIV-1 protease

complex structures. The confirmation of the role of the hinge regions in information transfer, the revelation of the location of the key interacting residues in the dominant paths of communication, and the adjustment of the variant structures with co-evolution reflected by the similarity in the communication pathways of wild-type and co-evolved structures suggest that this analysis can further be applied to investigate the communication between different sites within the structure. Both inter- and intramolecular communication can be inspected; i.e. the interactions between the substrate and different regions of the protease as well as those between the protease sites that display correlated mutations can be elaborated in the ligand bound HIV-1 protease structures. The shortest path prediction algorithm should also be improved by the analysis of communication between distant sites. This would help to predict the patterns of drug resistant mutations and to design potential binding sites for allosteric inhibitors to regulate the HIV-1 protease dynamics.

APPENDIX A: DOMINANT PATHWAYS OF COMMUNICATION IN SUBSTRATE COMPLEX STRUCTURES

Table A.1. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in ca-p2 complex structure

Starting point	Step 1	Step 2	Step 3	Frequency
P1	25.A			450
P1	27.A	25.A		119
P1	P1'	25.A		83
P1	P2	25.A		78
P1	28.A	25.A		66
P1	27.B	25.A		43
P1	25.B	25.A		39
P1	27.A	28.A	25.A	27
P1	27.A	26.A	25.A	21
P1	P2	28.A	25.A	19
P1	28.A	27.A	25.A	13
P1	P2	P1'	25.A	11
P1	P2	27.A	25.A	11
P1	25.B	27.A	25.A	10
P1	P1'	P2	25.A	9
P1	P1'	27.B	25.A	8
P1	25.B	27.B	25.A	8
P1	P3	28.A	25.A	7
P1	28.B	27.B	25.A	7
P1	28.A	26.A	25.A	6

Table A.2. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer B in ca-p2 complex structure

Starting point	Step 1	Step 2	Step 3	Frequency
P1	25.B			688
P1	27.A	25.B		83
P1	P1'	25.B		75
P1	28.B	25.B		71
P1	27.B	25.B		50
P1	23.B	25.B		40
P1	25.A	25.B		31
P1	84.B	25.B		24
P1	P2	25.B		16
P1	23.B	24.B	25.B	16
P1	27.B	28.B	25.B	11
P1	27.A	25.A	25.B	10
P1	27.B	26.B	25.B	10
P1	P2	27.A	25.B	9
P1	P1'	27.B	25.B	9
P1	84.B	85.B	25.B	9
P1	P1'	28.B	25.B	7
P1	27.A	26.A	25.B	7
P1	28.A	27.A	25.B	7
P1	P2'	28.B	25.B	6
P1	25.A	26.A	25.B	6
P1	P2	P1'	25.B	5
P1	27.A	24.B	25.B	5
P1	28.B	27.B	25.B	5

Table A.3. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer A in ca-p2 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.A				606
P1'	P1	25.A			78
P1'	27.B	25.A			78
P1'	23.A	25.A			36
P1'	P2	25.A			29
P1'	25.B	25.A			24
P1'	84.A	25.A			24
P1'	P1	27.A	25.A		21
P1'	P1	P2	25.A		19
P1'	P1	28.A	25.A		16
P1'	23.A	24.A	25.A		14
P1'	27.B	26.A	25.A		11
P1'	27.B	26.B	25.A		11
P1'	P1	25.B	25.A		9
P1'	P1	27.B	25.A		9
P1'	84.A	85.A	25.A		9
P1'	P2	P1	25.A		8
P1'	P2'	25.A			8
P1'	84.A	28.A	25.A		8
P1'	P2'	27.B	25.A		7
P1'	27.B	27.A	25.A		7
P1'	P1	27.A	28.A	25.A	6
P1'	27.B	25.B	25.A		6
P1'	28.B	27.B	25.A		6

Table A.4. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer B in ca-p2 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.B				486
P1'	P1	25.B			114
P1'	27.B	25.B			82
P1'	28.B	25.B			52
P1'	25.A	25.B			24
P1'	27.B	26.B	25.B		22
P1'	P1	27.A	25.B		19
P1'	P2'	28.B	25.B		16
P1'	P1	28.B	25.B		12
P1'	27.B	28.B	25.B		12
P1'	P2	P1	25.B		11
P1'	P1	27.B	25.B		11
P1'	28.B	27.B	25.B		11
P1'	25.A	27.A	25.B		9
P1'	P1	25.A	25.B		7
P1'	P1	84.B	25.B		6
P1'	P2'	28.B	27.B	25.B	6
P1'	P2'	30.B	28.B	25.B	6
P1'	28.B	26.B	25.B		6
P1'	P2'	25.B			5
P1'	P3'	P2'	28.B	25.B	5
P1'	25.A	26.B	25.B		5

Table A.5. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in ma-ca complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				349
P1	27.A	25.A			113
P1	P1'	25.A			104
P1	28.A	25.A			57
P1	27.B	25.A			35
P1	25.B	25.A			26
P1	P2	28.A	25.A		20
P1	27.A	26.A	25.A		18
P1	P2	27.A	25.A		15
P1	27.A	28.A	25.A		12
P1	28.A	27.A	25.A		12
P1	28.B	25.A			10
P1	P1'	84.A	25.A		9
P1	P2'	P1'	25.A		9
P1	P2	28.A	27.A	25.A	8
P1	25.B	27.B	25.A		8
P1	P3	29.A	28.A	25.A	7
P1	P2	25.A			7
P1	P1'	27.A	25.A		7
P1	P3	28.A	25.A		6
P1	P2	P1'	25.A		6
P1	P1'	P2'	25.A		6
P1	25.B	26.B	25.A		6
P1	P3	27.A	25.A		5
P1	P1'	27.B	25.A		5
P1	P1'	28.B	25.A		5
P1	25.B	27.A	25.A		5

Table A.6. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer B in ma-ca complex structure

Starting point	Step 1	Step 2	Step 3	Frequency
P1	25.B			477
P1	P1'	25.B		74
P1	27.B	25.B		54
P1	27.A	25.B		52
P1	28.B	25.B		52
P1	23.B	25.B		22
P1	P2	25.B		18
P1	84.B	25.B		18
P1	P2	27.A	25.B	12
P1	28.A	25.B		12
P1	P1'	27.B	25.B	10
P1	P2	P1'	25.B	9
P1	23.B	24.B	25.B	9
P1	25.A	27.A	25.B	9
P1	28.B	27.B	25.B	9
P1	P3	25.B		8
P1	P1'	28.B	25.B	8
P1	P2'	25.B		8
P1	27.A	28.A	25.B	8
P1	P1'	27.A	25.B	7
P1	25.A	25.B		7
P1	25.A	27.B	25.B	7
P1	27.B	26.B	25.B	7
P1	P1'	P2'	25.B	6
P1	P2	28.A	25.B	5
P1	27.B	28.B	25.B	5
P1	84.B	85.B	25.B	5

Table A.7. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer A in ma-ca complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.A				566
P1'	P1	25.A			84
P1'	27.B	25.A			71
P1'	27.A	25.A			52
P1'	84.A	25.A			40
P1'	23.A	25.A			25
P1'	P1	27.A	25.A		23
P1'	25.B	25.A			22
P1'	28.B	25.A			16
P1'	P1	28.A	25.A		12
P1'	27.A	28.A	25.A		10
P1'	23.A	24.A	25.A		9
P1'	27.A	26.A	25.A		9
P1'	27.B	26.B	25.A		9
P1'	P2'	25.A			8
P1'	28.B	27.B	25.A		8
P1'	84.A	85.A	25.A		7
P1'	P2	P1	25.A		6
P1'	P2'	P1	25.A		6
P1'	P2'	27.B	25.A		6
P1'	P1	25.B	25.A		5
P1'	P1	27.A	26.A	25.A	5
P1'	23.A	27.B	25.A		5
P1'	25.B	26.A	25.A		5
P1'	25.B	27.A	25.A		5

Table A.8. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer B in ma-ca complex structure

Starting point	Step 1	Step 2	Step 3	Frequency
P1'	25.B			361
P1'	P1	25.B		103
P1'	P2'	25.B		75
P1'	27.B	25.B		75
P1'	28.B	25.B		59
P1'	27.A	25.B		39
P1'	25.A	25.B		30
P1'	P2'	28.B	25.B	24
P1'	P2'	27.B	25.B	15
P1'	P1	27.A	25.B	13
P1'	P2'	P1	25.B	10
P1'	28.B	27.B	25.B	10
P1'	P1	28.B	25.B	9
P1'	P1	P2'	25.B	8
P1'	P2'	84.B	25.B	8
P1'	25.A	26.B	25.B	8
P1'	P2	P1	25.B	7
P1'	P1	27.B	25.B	7
P1'	25.A	26.A	25.B	7
P1'	27.B	26.B	25.B	7
P1'	25.A	27.A	25.B	6
P1'	25.A	27.B	25.B	6
P1'	27.B	28.B	25.B	6
P1'	P1	P2	25.B	5
P1'	23.A	27.B	25.B	5
P1'	25.A	28.A	25.B	5

Table A.9. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in nc-p1 complex structure

Starting point	Step 1	Step 2	Step 3	Frequency
P1	25.A			356
P1	27.A	25.A		98
P1	P2	25.A		60
P1	28.A	25.A		59
P1	P1’	25.A		47
P1	27.B	25.A		37
P1	27.A	28.A	25.A	25
P1	25.B	25.A		24
P1	P3	P2	25.A	20
P1	P3	28.A	25.A	20
P1	P2	28.A	25.A	16
P1	P3	27.A	25.A	13
P1	27.A	26.A	25.A	13
P1	28.A	27.A	25.A	11
P1	P1’	27.B	25.A	10
P1	P2	27.A	25.A	9
P1	P2’	P1’	25.A	9
P1	P1’	P2’	25.A	8
P1	28.A	26.A	25.A	7
P1	P3	25.A		6
P1	P1’	84.A	25.A	6
P1	23.B	26.A	25.A	6
P1	25.B	27.B	25.A	6
P1	27.A	27.B	25.A	6

Table A.10. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer B in nc-p1 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.B				511
P1	27.A	25.B			65
P1	28.B	25.B			61
P1	27.B	25.B			53
P1	P1’	25.B			39
P1	84.B	25.B			34
P1	23.B	25.B			29
P1	P1’	P2’	25.B		27
P1	P2	27.A	25.B		18
P1	25.A	25.B			16
P1	P2’	25.B			15
P1	27.B	28.B	25.B		14
P1	P3	25.B			12
P1	P1’	27.B	25.B		10
P1	23.B	24.B	25.B		10
P1	84.B	28.B	25.B		10
P1	P1’	P2’	28.B	25.B	9
P1	P2’	28.B	25.B		9
P1	P2’	P1’	25.B		8
P1	84.B	85.B	25.B		8
P1	27.B	26.B	25.B		7

Table A.11. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer A in nc-p1 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.A				455
P1'	27.B	25.A			59
P1'	P2'	25.A			40
P1'	P1	25.A			32
P1'	P2	25.A			30
P1'	84.A	25.A			30
P1'	23.A	25.A			25
P1'	P2'	P1	25.A		15
P1'	P2'	27.B	25.A		14
P1'	P1	27.A	25.A		12
P1'	25.B	25.A			10
P1'	P1	28.A	25.A		8
P1'	27.B	26.A	25.A		8
P1'	P2	27.A	25.A		7
P1'	25.B	27.A	25.A		7
P1'	P2	28.A	25.A		6
P1'	P1	P2	25.A		6
P1'	P2'	23.A	25.A		6
P1'	23.A	24.A	25.A		5
P1'	27.B	24.A	25.A		5
P1'	P2'	25.B	25.A		4
P1'	P2'	84.A	25.A		4
P1'	8.A	9.A	24.A	25.A	4

Table A.12. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer B in nc-p1 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.B				327
P1'	P2'	25.B			153
P1'	27.B	25.B			71
P1'	P1	25.B			52
P1'	P2'	28.B	25.B		36
P1'	P2'	27.B	25.B		28
P1'	27.B	26.B	25.B		25
P1'	25.A	25.B			20
P1'	27.B	28.B	25.B		16
P1'	P2'	P1	25.B		12
P1'	P1	27.A	25.B		11
P1'	25.A	27.A	25.B		10
P1'	P2'	27.B	26.B	25.B	9
P1'	P2'	27.B	28.B	25.B	9
P1'	P1	27.B	25.B		8
P1'	P2'	84.B	25.B		8
P1'	P1	28.B	25.B		7
P1'	P1	P2'	25.B		5
P1'	P1	84.B	25.B		5
P1'	27.B	25.A	25.B		5
P1'	P1	27.B	28.B	25.B	4
P1'	P2'	28.B	27.B	25.B	4
P1'	P2'	29.B	28.B	25.B	4
P1'	P3'	P2'	25.B		4
P1'	25.A	27.B	25.B		4
P1'	27.B	P2'	25.B		4
P1'	27.B	26.A	26.B	25.B	4

Table A.13. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				294
P1	P2	25.A			107
P1	27.A	25.A			90
P1	28.A	25.A			53
P1	P1'	25.A			38
P1	27.B	25.A			29
P1	P2	27.A	25.A		26
P1	P2	28.A	25.A		26
P1	27.A	28.A	25.A		24
P1	25.B	25.A			20
P1	27.A	26.A	25.A		18
P1	P3	27.A	25.A		11
P1	P2	P1'	25.A		10
P1	P2	84.A	25.A		9
P1	P3	28.A	25.A		8
P1	25.B	27.A	25.A		8
P1	P2	P3	28.A	25.A	7
P1	P1'	P2	25.A		7
P1	P2	P3	25.A		6
P1	P2	28.A	27.A	25.A	6
P1	25.B	27.B	25.A		6
P1	27.B	26.B	25.A		6
P1	P1'	P2'	27.B	25.A	5
P1	P1'	27.B	25.A		5
P1	23.B	27.A	25.A		5
P1	25.B	26.B	25.A		5
P1	28.A	27.A	25.A		5

Table A.14. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer B in p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.B				485
P1	27.A	25.B			63
P1	28.B	25.B			51
P1	27.B	25.B			39
P1	P1'	25.B			38
P1	P2	25.B			32
P1	23.B	25.B			28
P1	84.B	25.B			25
P1	P2	27.A	25.B		16
P1	P1'	27.B	25.B		10
P1	23.B	24.B	25.B		9
P1	25.A	25.B			9
P1	27.A	25.A	25.B		9
P1	25.A	27.B	25.B		8
P1	27.B	28.B	25.B		8
P1	27.B	26.B	25.B		7
P1	P3	27.A	25.B		6
P1	P2	P1'	25.B		6
P1	27.A	23.B	25.B		6
P1	27.A	26.A	25.B		6
P1	P2	28.A	27.A	25.B	5
P1	P1'	28.B	25.B		5
P1	27.A	27.B	25.B		5
P1	28.A	25.A	25.B		5
P1	84.B	85.B	25.B		5

Table A.15. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer A in p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.A				440
P1'	27.B	25.A			63
P1'	84.A	25.A			45
P1'	P1	25.A			41
P1'	P2	25.A			29
P1'	23.A	25.A			23
P1'	25.B	25.A			23
P1'	P1	P2	25.A		16
P1'	P2	P1	25.A		13
P1'	P2	28.A	25.A		13
P1'	23.A	24.A	25.A		12
P1'	P1	27.A	25.A		10
P1'	25.B	27.A	25.A		9
P1'	P2'	25.A			8
P1'	P2'	27.B	25.A		7
P1'	84.A	85.A	25.A		7
P1'	P1	28.A	25.A		6
P1'	25.B	26.B	25.A		6
P1'	P2	P1	27.A	25.A	5
P1'	25.B	27.B	25.A		5
P1'	27.B	23.A	25.A		5
P1'	27.B	24.A	25.A		5
P1'	28.B	27.B	25.A		5

Table A.16. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer B in p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.B				388
P1'	27.B	25.B			84
P1'	28.B	25.B			46
P1'	P1	25.B			45
P1'	P2'	28.B	25.B		23
P1'	25.A	25.B			21
P1'	27.B	26.B	25.B		18
P1'	P2	P1	25.B		13
P1'	P2'	27.B	25.B		10
P1'	27.B	28.B	25.B		10
P1'	P1	27.B	25.B		9
P1'	P1	28.B	25.B		9
P1'	P1	27.A	25.B		8
P1'	25.A	26.B	25.B		7
P1'	28.B	27.B	25.B		7
P1'	23.A	27.B	25.B		6
P1'	25.A	27.B	25.B		6
P1'	P1	84.B	25.B		5
P1'	P2'	P3'	28.B	25.B	5
P1'	P2'	25.B			5
P1'	P2'	28.B	26.B	25.B	5
P1'	P2'	P1	25.B		4
P1'	P2'	27.B	26.B	25.B	4
P1'	P2'	29.B	28.B	25.B	4
P1'	P3'	P2'	28.B	25.B	4
P1'	27.B	28.B	26.B	25.B	4

Table A.17. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in p2-nc complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				396
P1	27.A	25.A			112
P1	P1'	25.A			87
P1	28.A	25.A			71
P1	P2	25.A			62
P1	27.B	25.A			45
P1	25.B	25.A			32
P1	P2	28.A	25.A		20
P1	27.A	28.A	25.A		19
P1	27.A	26.A	25.A		15
P1	P1'	27.A	25.A		12
P1	P1'	27.B	25.A		10
P1	25.B	27.A	25.A		10
P1	P2	27.A	25.A		9
P1	28.A	26.A	25.A		9
P1	25.B	27.B	25.A		8
P1	28.A	27.A	25.A		8
P1	P3	28.A	25.A		7
P1	P2	P1'	25.A		7
P1	25.B	26.A	25.A		7
P1	P2	27.A	26.A	25.A	6
P1	P2	84.A	25.A		6
P1	28.B	27.B	25.A		6
P1	P1'	84.A	25.A		5
P1	25.B	26.B	25.A		5
P1	27.A	27.B	25.A		5

Table A.18. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer B in p2-nc complex structure

Starting point	Step 1	Step 2	Step 3	Frequency
P1	25.B			721
P1	27.A	25.B		75
P1	P1'	25.B		55
P1	27.B	25.B		51
P1	28.B	25.B		51
P1	23.B	25.B		34
P1	84.B	25.B		33
P1	P1'	27.B	25.B	17
P1	27.B	26.B	25.B	13
P1	23.B	24.B	25.B	9
P1	25.A	25.B		9
P1	28.A	27.A	25.B	9
P1	P2	25.B		8
P1	27.A	26.A	25.B	8
P1	28.B	27.B	25.B	8
P1	P2	27.A	25.B	7
P1	P1'	25.A	25.B	7
P1	P2'	P1'	25.B	7
P1	27.A	25.A	25.B	7
P1	27.A	26.B	25.B	7
P1	P2'	28.B	25.B	6
P1	23.B	85.B	25.B	6
P1	27.B	28.B	25.B	6
P1	84.B	23.B	25.B	6
P1	P3	27.A	25.B	5
P1	P2	P1'	25.B	5
P1	27.A	24.B	25.B	5

Table A.19. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer A in p2-nc complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.A				548
P1'	27.A	25.A			56
P1'	P1	25.A			55
P1'	27.B	25.A			49
P1'	84.A	25.A			32
P1'	P2	25.A			29
P1'	23.A	25.A			22
P1'	25.B	25.A			19
P1'	P1	27.A	25.A		16
P1'	23.A	24.A	25.A		14
P1'	P2'	25.A			13
P1'	P2	28.A	25.A		11
P1'	P2'	27.B	25.A		9
P1'	27.A	26.A	25.A		9
P1'	P1	P2	25.A		8
P1'	27.B	26.B	25.A		8
P1'	84.A	85.A	25.A		8
P1'	P1	27.B	25.A		7
P1'	P1	28.A	25.A		7
P1'	25.B	26.A	25.A		7
P1'	P1	27.A	28.A	25.A	6
P1'	25.B	27.A	25.A		6
P1'	27.B	25.B	25.A		6
P1'	P2	P1	25.A		5
P1'	P1	27.A	26.A	25.A	5
P1'	27.B	24.A	25.A		5
P1'	27.B	26.A	25.A		5

Table A.20. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer B in p2-nc complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.B				367
P1'	27.B	25.B			108
P1'	P1	25.B			100
P1'	27.A	25.B			27
P1'	P2'	28.B	25.B		26
P1'	25.A	25.B			22
P1'	27.B	26.B	25.B		17
P1'	P2'	27.B	25.B		16
P1'	P1	27.A	25.B		14
P1'	27.B	28.B	25.B		13
P1'	P2	P1	25.B		11
P1'	P2'	P1	25.B		7
P1'	P1	28.B	25.B		6
P1'	P2'	25.B			6
P1'	P2'	28.B	27.B	25.B	6
P1'	25.A	27.A	25.B		6
P1'	27.A	25.A	25.B		6
P1'	27.A	27.B	25.B		6
P1'	P2'	84.B	25.B		5
P1'	25.A	26.B	25.B		5
P1'	27.A	26.A	25.B		5
P1'	27.B	26.A	25.B		5
P1'	P1	25.A	25.B		4
P1'	P1	27.B	25.B		4
P1'	P2'	27.B	28.B	25.B	4
P1'	P3'	P2'	28.B	25.B	4
P1'	23.A	27.B	25.B		4

Table A.21. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in rh-in complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				621
P1	27.B	25.A			65
P1	28.A	25.A			60
P1	27.A	25.A			56
P1	P1'	25.A			37
P1	84.A	25.A			35
P1	23.A	25.A			28
P1	P2	25.A			14
P1	25.B	25.A			13
P1	27.A	28.A	25.A		11
P1	P4	25.A			8
P1	27.B	26.B	25.A		8
P1	P1'	27.B	25.A		7
P1	P2'	28.A	25.A		7
P1	28.A	27.A	25.A		6
P1	28.B	27.B	25.A		6
P1	P1'	P2'	28.A	25.A	5
P1	P2'	P1'	25.A		5
P1	27.A	26.A	25.A		5
P1	27.A	26.A	26.B	25.A	5
P1	27.B	26.A	25.A		5
P1	27.B	27.A	25.A		5
P1	84.A	28.A	25.A		5
P1	84.A	85.A	25.A		5

Table A.22. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer B in rh-in complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.B				364
P1	27.B	25.B			93
P1	P2	25.B			63
P1	28.B	25.B			50
P1	27.A	25.B			48
P1	P1'	25.B			47
P1	25.A	25.B			32
P1	27.B	28.B	25.B		20
P1	P2	28.B	25.B		17
P1	P1'	27.A	25.B		12
P1	P2	27.B	25.B		11
P1	27.B	26.B	25.B		10
P1	28.B	27.B	25.B		10
P1	25.A	27.B	25.B		8
P1	28.A	25.A	25.B		8
P1	P1'	27.B	25.B		7
P1	P3	28.B	25.B		6
P1	P2	P1'	25.B		6
P1	25.A	27.A	25.B		6
P1	28.A	27.A	25.B		6
P1	P4	27.B	25.B		5
P1	P2	28.B	27.B	25.B	5
P1	25.A	26.A	25.B		5
P1	25.A	26.A	26.B	25.B	5
P1	25.A	26.B	25.B		5
P1	27.B	P1'	25.B		5
P1	28.B	86.B	25.B		5

Table A.23. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer A in rh-in complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.A				310
P1'	27.A	25.A			73
P1'	P1	25.A			64
P1'	P2'	28.A	25.A		34
P1'	27.B	25.A			30
P1'	P2'	27.A	25.A		27
P1'	25.B	25.A			16
P1'	27.A	26.A	25.A		14
P1'	27.A	28.A	25.A		14
P1'	P2'	25.A			12
P1'	P2'	P1	25.A		9
P1'	P1	27.B	25.A		7
P1'	P1	28.A	25.A		7
P1'	P2'	84.A	25.A		7
P1'	P2	P1	25.A		6
P1'	P2'	27.A	28.A	25.A	6
P1'	27.B	P1	25.A		6
P1'	P2'	27.A	26.A	25.A	5
P1'	P2'	28.A	27.A	25.A	5
P1'	27.A	27.B	25.A		5
P1'	27.B	26.B	25.A		5

Table A.24. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer B in rh-in complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.B				360
P1'	P1	25.B			49
P1'	27.B	25.B			44
P1'	27.A	25.B			43
P1'	23.B	25.B			30
P1'	P2'	27.A	25.B		27
P1'	84.B	25.B			22
P1'	P1	27.B	25.B		19
P1'	P2'	25.B			18
P1'	P2	25.B			17
P1'	25.A	25.B			15
P1'	84.B	85.B	25.B		11
P1'	P2'	P1	25.B		8
P1'	27.B	26.B	25.B		8
P1'	27.B	28.B	25.B		8
P1'	P1	28.B	25.B		7
P1'	P2	P1	25.B		6
P1'	P2	28.B	25.B		6
P1'	27.A	26.A	25.B		6
P1'	P1	P2	25.B		5
P1'	23.B	24.B	25.B		5
P1'	23.B	27.A	25.B		5
P1'	23.B	85.B	25.B		5
P1'	25.A	27.A	25.B		5
P1'	27.A	23.B	25.B		5

Table A.25. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in rt-rh complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				338
P1	27.A	25.A			109
P1	P2	25.A			59
P1	28.A	25.A			45
P1	27.B	25.A			35
P1	P1’	25.A			27
P1	25.B	25.A			25
P1	27.A	26.A	25.A		24
P1	27.A	28.A	25.A		19
P1	P2	27.A	25.A		11
P1	P2	28.A	25.A		11
P1	P3	27.A	25.A		10
P1	25.B	27.B	25.A		10
P1	P3	P2	25.A		8
P1	P3	29.A	28.A	25.A	8
P1	P1’	27.B	25.A		8
P1	P2’	P1’	25.A		8
P1	P3	25.A			7
P1	P2’	27.B	25.A		7
P1	28.A	26.A	25.A		7
P1	25.B	26.A	25.A		6
P1	27.A	25.B	25.A		6
P1	P3	27.A	26.A	25.A	5
P1	P2	P3	25.A		5
P1	25.B	27.A	25.A		5

Table A.26. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer B in rt-rh complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.B				593
P1	27.A	25.B			64
P1	28.B	25.B			51
P1	27.B	25.B			37
P1	84.B	25.B			27
P1	23.B	25.B			25
P1	P1'	25.B			22
P1	P3	25.B			21
P1	23.B	24.B	25.B		16
P1	P2	27.A	25.B		12
P1	25.A	25.B			12
P1	27.B	26.B	25.B		12
P1	P1'	27.B	25.B		11
P1	27.A	25.A	25.B		11
P1	28.B	27.B	25.B		10
P1	P2	25.B			9
P1	P3	27.A	25.B		8
P1	27.A	26.B	25.B		7
P1	28.A	27.A	25.B		7
P1	84.B	85.B	25.B		7
P1	P3	P2	25.B		6
P1	P2'	27.B	25.B		6
P1	25.A	27.B	25.B		6
P1	27.A	27.B	25.B		5
P1	27.A	28.A	25.A	25.B	5
P1	27.B	28.B	25.B		5

Table A.27. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer A in rt-rh complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.A				368
P1'	27.B	25.A			44
P1'	P1	25.A			35
P1'	27.A	25.A			34
P1'	P2'	25.A			32
P1'	P2'	27.B	25.A		24
P1'	23.A	25.A			23
P1'	P2	25.A			16
P1'	25.B	25.A			15
P1'	84.A	25.A			15
P1'	P1	27.A	25.A		11
P1'	23.A	24.A	25.A		11
P1'	P2'	P1	25.A		9
P1'	P2'	P1	P2	25.A	6
P1'	27.A	26.A	25.A		6
P1'	27.A	28.A	25.A		6
P1'	27.B	25.B	25.A		6
P1'	27.B	26.A	25.A		6
P1'	P2	P1	25.A		5
P1'	P2'	P2	25.A		5
P1'	25.B	27.B	25.A		5

Table A.28. Dominant pathways between the P1' cleavage site and the active site residue 25 of protease monomer B in rt-rh complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1'	25.B				276
P1'	27.B	25.B			67
P1'	P1	25.B			42
P1'	P2'	28.B	25.B		38
P1'	P2'	25.B			25
P1'	27.A	25.B			25
P1'	P2'	27.B	25.B		21
P1'	27.B	28.B	25.B		20
P1'	27.B	26.B	25.B		15
P1'	P2'	P1	25.B		13
P1'	P1	27.B	25.B		12
P1'	P1	27.A	25.B		10
P1'	25.A	25.B			10
P1'	P2	P1	25.B		9
P1'	P2'	84.B	25.B		9
P1'	P2'	27.B	26.B	25.B	6
P1'	P2'	28.B	27.B	25.B	6
P1'	P1	28.B	25.B		5
P1'	27.B	P1	25.B		5
P1'	27.B	27.A	25.B		5
P1'	P2'	P3'	25.B		4
P1'	P2'	28.B	29.B	25.B	4
P1'	P2'	84.B	85.B	25.B	4
P1'	8.A	29.B	28.B	25.B	4
P1'	P3'	25.B			4
P1'	P3'	29.B	28.B	25.B	4
P1'	27.B	25.A	25.B		4

Table A.29. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in the best members of the largest cluster of MD simulated wild-type p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				272
P1	P2	25.A			131
P1	27.A	25.A			85
P1	P1'	25.A			54
P1	25.B	25.A			46
P1	P2	27.A	25.A		35
P1	27.B	25.A			29
P1	28.A	25.A			28
P1	27.A	26.A	25.A		25
P1	P2	28.A	25.A		21
P1	P2	P1'	25.A		19
P1	27.A	28.A	25.A		13
P1	27.A	25.B	25.A		11
P1	P3	P2	25.A		10
P1	P2	P3	25.A		9
P1	P2	84.A	25.A		8
P1	25.B	26.B	25.A		7
P1	25.B	27.B	25.A		7
P1	P3	28.A	25.A		7
P1	P3	25.A			6
P1	P3	27.A	25.A		6
P1	P2	P3	27.A	25.A	6

Table A.30. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in the best members of the largest cluster of MD simulated D30N mutant p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				398
P1	27.A	25.A			87
P1	P2	25.A			83
P1	P1'	25.A			73
P1	27.B	25.A			38
P1	27.A	26.A	25.A		32
P1	28.A	25.A			29
P1	25.B	25.A			27
P1	27.A	28.A	25.A		19
P1	P3	P2	25.A		15
P1	P2	P1'	25.A		14
P1	P2	28.A	25.A		13
P1	P3	27.A	25.A		11
P1	P3	28.A	25.A		11
P1	P2	27.A	25.A		11
P1	25.B	27.A	25.A		9
P1	P3	25.A			9
P1	25.B	26.A	25.A		8
P1	28.B	25.A			8
P1	28.B	27.B	25.A		7
P1	P3	27.A	26.A	25.A	7
P1	28.A	27.A	25.A		6
P1	23.B	26.A	25.A		6
P1	P1'	25.B	25.A		6
P1	P1'	P2	25.A		6

Table A.31. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in the best members of the largest cluster of MD simulated N88D mutant p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				316
P1	27.A	25.A			118
P1	P2	25.A			89
P1	P1'	25.A			44
P1	25.B	25.A			37
P1	P3	27.A	25.A		29
P1	P3	P2	25.A		23
P1	27.A	28.A	25.A		20
P1	27.B	25.A			20
P1	P2	27.A	25.A		19
P1	27.A	26.A	25.A		14
P1	28.B	25.A			12
P1	P3	25.A			10
P1	P2	84.A	25.A		10
P1	P2	P1'	25.A		10
P1	P2	28.A	25.A		9
P1	P1'	27.B	25.A		9
P1	25.B	26.A	25.A		8
P1	P2	P3	25.A		8
P1	P1'	84.A	25.A		8
P1	25.B	27.B	25.A		7
P1	84.B	25.A			7
P1	27.A	25.B	25.A		6
P1	27.B	25.B	25.A		6

Table A.32. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in the best members of the largest cluster of MD simulated D30N-N88D mutant p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				541
P1	P2	25.A			136
P1	28.A	25.A			58
P1	25.B	25.A			53
P1	27.B	25.A			52
P1	P1’	25.A			51
P1	27.A	25.A			46
P1	P2	28.A	25.A		19
P1	P2	84.A	25.A		17
P1	27.A	26.A	25.A		16
P1	28.A	27.A	25.A		15
P1	84.A	25.A			14
P1	25.B	27.B	25.A		9
P1	P1’	P2	25.A		9
P1	P1’	P2’	25.A		9
P1	25.B	26.B	25.A		8
P1	27.B	26.B	25.A		8
P1	25.B	27.A	25.A		7
P1	28.B	27.B	25.A		7
P1	P3	P2	25.A		7
P1	28.A	27.A	26.A	25.A	6
P1	23.B	27.A	25.A		6
P1	25.B	26.A	25.A		6
P1	P1’	27.B	25.A		6

Table A.33. Dominant pathways between the P1 cleavage site and the active site residue 25 of protease monomer A in the best members of the largest cluster of MD simulated D30N-N88D-LP1'F mutant p1-p6 complex structure

Starting point	Step 1	Step 2	Step 3	Step 4	Frequency
P1	25.A				311
P1	P2	25.A			156
P1	27.A	25.A			103
P1	P1'	25.A			47
P1	P2	P1'	25.A		37
P1	25.B	25.A			34
P1	P2	27.A	25.A		32
P1	27.A	28.A	25.A		19
P1	27.A	26.A	25.A		18
P1	P2	28.A	25.A		15
P1	P3	P2	25.A		12
P1	P2	P3	25.A		10
P1	29.A	28.A	25.A		9
P1	P2	P3	27.A	25.A	9
P1	25.B	27.B	25.A		8
P1	28.B	25.A			8
P1	23.B	25.B	25.A		7
P1	P2	27.A	28.A	25.A	7
P1	P1'	27.B	25.A		7
P1	29.A	87.A	25.A		6
P1	P3	P2	27.A	25.A	6
P1	27.A	28.A	26.A	25.A	5
P1	27.A	25.B	25.A		5
P1	27.A	P2	25.A		5

REFERENCES

Alpaydin, E., 2004, *Introduction to Machine Learning*, The MIT Press, Cambridge, Massachusetts.

Altman, M. D., E. A. Nalivaika, M. Prabu-Jeyabalan, C. A. Schiffer and B. Tidor, 2008, "Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease", *Proteins: Structure, Function and Genetics*, Vol. 70, pp. 678-694.

Altuvia, Y., O. Schueler and H. Margalit, 1995, "Ranking potential binding peptides to MHC molecules by a computational threading approach", *Journal of Molecular Biology*, Vol. 249, pp. 244-250.

Altuvia, Y., A. Sette, J. Sidney, S. Southwood and H. Margalit, 1997, "A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets", *Human Immunology*, Vol. 58, pp. 1-11.

Amadei, A., A. B. Linssen and H. J. C. Berendsen, 1993, "Essential dynamics of proteins", *Proteins*, Vol. 17, pp. 412-425.

Amadei, A., M. A. Ceruso and A. Di Nola, 1999, "On the Convergence of the Conformational Coordinates Basis Set Obtained by the Essential Dynamics Analysis of Proteins' Molecular Dynamics Simulations", *Proteins: Structure, Function, and Genetics*, Vol. 36, pp. 419-424.

Amitai, G., A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger and S. Pietrokovski, 2004, "Network analysis of protein structures identifies functional residues", *Journal of Molecular Biology*, Vol. 344, pp. 1135-1146.

Atilgan, A. R., D. Turgut and C. Atilgan, 2007, "Screened non-bonded interactions in native proteins manipulate optimal paths for robust residue communication",

Biophysical Journal BioFAST, doi:10.1529/biophysj.106.099440.

Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin and I. Bahar, 2001, "Anisotropy of fluctuation dynamics of proteins with an elastic network model", *Biophysical Journal*, Vol. 80, pp. 505-515.

Bahar, I., 1999, "Dynamics of proteins and biomolecular complexes: inferring functional motions from structure", *Reviews in Chemical Engineering*, Vol. 15, pp. 319-349.

Bahar, I. and R. L. Jernigan, 1996, "Coordination geometry of non-bonded residues in globular proteins", *Folding and Design*, Vol. 1, pp. 357-370.

Bahar, I. and R. L. Jernigan, 1997, "Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separations", *Journal of Molecular Biology*, Vol. 266, pp. 195-214.

Bahar, I. and R. L. Jernigan, 1998, "Vibrational dynamics of transfer RNAs: Comparison of the free and synthetase-bound forms", *Journal of Molecular Biology*, Vol. 281, pp. 871-884.

Bahar, I. and R. L. Jernigan, 1999, "Cooperative fluctuations and subunit communication in tryptophan synthase", *Biochemistry*, Vol. 38, pp. 3478-3490.

Bahar, I., A. R. Atilgan and B. Erman, 1997a, "Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential", *Folding and Design*, Vol. 2, pp. 173-181.

Bahar, I., A. R. Atilgan, M. C. Demirel and B. Erman, 1998a, "Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability", *Physical Review Letters*, Vol. 80, pp. 2733-2736.

Bahar I., A. Wallquist, D. G. Covell and R. L. Jernigan, 1998b, "Correlation between native state hydrogen exchange and cooperative residue fluctuations from a simple

model", *Biochemistry*, Vol. 37, pp. 1067-1075.

Bahar I., B. Erman, T. Haliloglu, and R. L. Jernigan, 1997b, "Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations", *Biochemistry*, Vol. 36, pp. 13512-13532.

Bahar, I., C. Chennubhotla and D. Tobi, 2007, "Intrinsic Enzyme Dynamics in the Unbound State and Relation to Allosteric Regulation", *Current Opinion in Structural Biology*, Vol. 17, pp. 633-640.

Bahar, I., M. Kaplan and R. L. Jernigan, 1997c, "Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches", *Proteins*, Vol. 29, pp. 292-308.

Batagelj, V. and A. Mrvar, 1998, "Pajek - Program for large network analysis", *Connections*, Vol. 21, pp. 47-57.

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.

Bernstein, E. E., T. F. Koetzle, G. J. B. Williams, J. E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. J. Tasumi, 1977, "The Protein Data Bank: a computer-based archival file for macromolecular structures", *Journal of Molecular Biology*, Vol. 117, pp. 535-542.

Bryant, S. H. and C. E. Lawrance, 1993, "An empirical energy function for threading protein sequences through the folding motif", *Proteins*, Vol. 16, pp. 92-112.

Bowman, M. J., S. Byrne and J. Chmielewski, 2005, "Switching between allosteric and dimerization inhibition of HIV-1 protease", *Chemistry & Biology*, Vol. 12, 439-444.

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*, University

of California, San Francisco.

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

Changeux, J. and S. J. Edelstein, 2005, "Allosteric Mechanisms of Signal Transduction", *Science*, Vol. 308, pp. 1424-1428.

Chellappan, S., G. S. Kiran Kumar Reddy, 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.

Chelli, R., F. L. Gervasio, P. Procacci and V. Schettino, 2004, "Inter-residue and solvent-residue interactions in proteins: a statistical study on experimental structures", *Proteins*, Vol. 55, pp. 139-151.

Chen, Z., Y. Li, E. Chen, D. L. Hall, P. L. Darke, C. Culberson, J. A. Shafer and L. C. Kuo, 1994, "Crystal structure at 1.9-A resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases", *Journal of Biological Chemistry*, Vol. 269, pp. 26344-26348.

Chennubhotla, C., A. J. Rader, L. Yang and I. Bahar, 2005, "Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies", *Physical Biology*, Vol. 2, pp. S173-S180.

Chennubhotla, C. and I. Bahar, 2006, "Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES", *Molecular Systems Biology*, Vol.2, pp. 36-48.

Chennubhotla, C. and I. Bahar, 2007a, "Markov methods for hierarchical coarse-graining of large protein dynamics", *Journal of Computational Biology*, Vol.14, pp. 765-776.

Chennubhotla, C. and I. Bahar, 2007b, "Signal propagation in proteins and relation to equilibrium fluctuations", *PLOS Computational Biology*, Vol.3, pp. 1716-1726.

Chennubhotla, C., Z. Yang and I. Bahar, 2008, "Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL", *Molecular Biosystems*, Vol. 4, pp. 287-292.

Chou, K.C., 1996, "Prediction of human immunodeficiency virus protease cleavage sites in proteins", *Analytical Biochemistry*, Vol. 233, pp. 1-14.

Covell, D. G. and R. L. Jernigan, 1990, "Conformations of folded proteins in restricted spaces", *Biochemistry*, Vol. 29, pp. 3287-3294.

Crooks, G. E., J. Wolfe and S. E. Brenner, 2004, "Measurements of protein sequence-structure correlations", *Proteins*, Vol. 57, pp. 804-810.

Cui, Q. and I. Bahar, 2005, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, CRC Press, Boca Raton, FL.

deGroot, B. L., S. Hayward, D. M. F. vanAalten, A. Amadei and H. J. C. Berendsen, 1998, "Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data", *Proteins*, Vol. 31, pp. 116-127.

Delano, W. L., 2002, *The PYMOL Molecular Graphics System*, DeLano Scientific, San Carlos, CA, USA.

del Sol, A., H. Fujihashi, D. Amoros and R. Nussinov, 2006a, "Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families", *Protein Science*, Vol.15, pp. 2120-2128.

del Sol, A., H. Fujihashi, D. Amoros and R. Nussinov, 2006b, "Residues crucial for maintaining short paths in network communication mediate signaling in proteins", *Molecular Systems Biology*, Vol.2, 2006.0019. doi:10.1038/msb4100063.

del Sol, A. and P. O'Meara, 2004, "Small-world network approach to identify key residues in protein-protein interaction", *Proteins*, Vol. 58, pp. 672-682.

Doruker, P., A. R. Atilgan and I. Bahar, 2000, "Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to -amylase inhibitor", *Proteins: Structure, Function and Genetics*, Vol. 40, pp. 512-524.

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

Feig, M., J. Karanicolas and C. L. Brooks, 2004, "MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology", *Journal of Molecular Graphicas & Modelling*, Vol. 22, pp. 377-395.

Flexner, C., 1998, "HIV-1 protease inhibitors", *The New England Journal of Medicine*, Vol. 338, pp. 1281-1292.

Flory, P. J., 1969, *Statistical Mechanics of Chain Molecules*, Interscience, New York.

Flory, P. J., 1976, "Statistical thermodynamics of random networks", *Proceedings of the Royal Society of London A*, Vol. 351, pp. 351-380.

Go, N., T. Noguti and T. Nishikawa, 1983, "Dynamics of a small protein in terms of low-frequency vibrational modes", *PNAS*, Vol. 80, pp. 3696-3700.

Goodsell, D. S., 2000, "Molecule of the month, HIV-1 protease", *PDB Newsletter*, Vol. 6.

Gordon, D. B., S. A. Marshall and S. L. Mayo, 1999, "Energy functions for protein design", *Current Opinion in Structural Biology*, Vol. 9, pp. 509-513.

Haliloglu, T., 1999, "Characterization of internal motions of Escherichia coli Ribonuclease H by Monte Carlo simulation", *Proteins*, Vol. 34, pp. 533-539.

Haliloglu, T. and I. Bahar, 1998, "Coarse-grained simulations of conformational dynamics of proteins: application to apomyoglobin", *Proteins*, Vol. 31, pp. 271-281.

Haliloglu, T. and I. Bahar, 1999, "Structure-based analysis of protein dynamics: Comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data", *Proteins*, Vol. 37, pp. 654-667.

Haliloglu, T., I. Bahar and B. Erman, 1997, "Gaussian dynamics of folded proteins", *Physical Review Letters*, Vol. 79, pp. 3090-3093.

Hamacher, K. and J. A. McCammon, 2006, "Computing the amino acid specificity of fluctuations in biomolecular systems", *Journal of Chemical Theory and Computation*, Vol. 2, pp. 873-878.

Hayward, S., A. Kitao and G. Nobuhiro, 1994, "Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and principal component analysis", *Protein Science*, Vol. 3, pp. 936-943.

Hinsen, K., 1998, "Analysis of domain motions by approximate normal mode calculations", *Proteins*, Vol. 33, pp. 417-429.

Hoffman, N. G., C. A. Schiffer and R. Swanstrom, 2003, "Covariation of amino acid positions in HIV-1 protease", *Virology*, Vol. 314, pp. 536-548.

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, "Improves survival among HIV-infected individuals following initiation of antiretroviral therapy", *Journal of the American Medical Association*, Vol. 279, pp. 450-454.

Hornak, V. and C. Simmerling, 2007, "Targeting structural flexibility in HIV-1 protease inhibitor binding", *Drug Discovery Today*, Vol. 12, pp. 132-138.

Hou, T., W. A. McLaughlin and W. Wang, 2008, "Evaluating the potency of HIV-1 protease drugs to combat resistance", *Proteins: Structure, Function and Genetics*, Vol. 71, pp. 1163-1174.

Jernigan, R.L. and I. Bahar, 1996, "Structure-derived potentials and protein simulations", *Current Opinion in Structural Biology*, Vol. 6, pp. 195-209.

Jones, D. T. and J. M. Thornton, 1992, "Protein fold recognition", *Journal of Computer Aided Molecular Design*, Vol. 7, pp. 439-456.

Jones, D. T. and J. M. Thornton, 1996, "Potential energy functions for threading", *Current Opinion in Structural Biology*, Vol.6, pp. 210-216.

Jones D. T., W. R. Taylor and J. M. Thornton, 1992, "A new approach to Protein Fold Recognition", *Nature*, Vol. 358, pp. 86-89.

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.

Kaldor, S.W., V. J. Kalish, J. F. Davies 2nd., B. V. Shetty, J. E. Fritz, K. Appelt, J. A. Burgess, K. M. Campanale, N. Y. Chirgadze, D. K. Clawson, B. A. Dressman, S. D. Hatch, D. A. Khalil, M. B. Kosa, P. P. Lubbehusen, M. A. Muesing, A. K. Patick, S. H. Reich, K. S. Su and J. H. Tatlock, 1997, "Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease", *Journal of Medicinal Chemistry*, Vol. 40, pp. 3979-3985.

Kempf, D. J., K. C. Marsh, J. F. Denissen, E. McDonald, S. Vasavanonda, C. A. Flentge, B. E. Green, L. Fino, C. H. Park, X. P. Kong, et al., 1995, "ABT-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral

bioavailability in humans", *PNAS USA*, Vol. 92, pp. 2484-2488.

Keskin, O. and I. Bahar, 1998, "Packing of sidechains in low-resolution models for proteins", *Folding and Design*, Vol. 3, pp. 469-479.

Keskin, O., R. L. Jernigan and I. Bahar, 2000, "Proteins with similar architecture exhibit similar large-scale dynamic behavior", *Biophysical Journal*, Vol. 78, pp. 2093-2196.

Keskin, O., S. R. Durell, I. Bahar, R. L. Jernigan and D. G. Covell, 2002, "Relating molecular flexibility to function: A case study of Tubulin", *Biophysical Journal*, Vol. 83, pp. 663-680.

Kim, E. E., C. T. Baker, M. D. Dwyer, M. A. Murcko, B. G. Rao, R. D. Tung and M. A. Navia, 1995, "Crystal structure of HIV-1 protease in complex with Vx-478, a potent and orally bioavailable inhibitor of the enzyme", *Journal of American Chemical Society*, Vol. 117, pp. 1181-1182.

King, N., M. Prabu-Jeyabalan, E. A. Nalivaika and C. A. Schiffer, 2004, "Combating susceptibility to drug resistance", *Chemistry & Biology*, Vol. 11, pp. 1333-1338.

Koehl, P. and M. Levitt, 1999a, "De novo protein design. I. In search of stability and specificity", *Journal of Molecular Biology*, Vol. 293, pp.1161-1181.

Koehl, P. and M. Levitt, 1999b, "De novo protein design. II. Plasticity in sequence space", *Journal of Molecular Biology*, Vol. 293, pp. 1183-1193.

Kolli, N., 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.

Krohn, A., S. Redshaw, J. C. Ritchie, B. J. Graves and M. H. Hatada, 1991, "Novel binding mode of highly potent HIV-proteinase inhibitors incorporating the (R)-hydroxyethylamine isostere", *Journal of Medicinal Chemistry*, Vol. 34, pp. 3340-3342.

Kurt, N., T. Haliloglu and C. A. Schiffer, 2003a, "Structure based prediction of potential binding and non-binding peptides to HIV-1 protease", *Biophysical Journal*, Vol. 85, pp. 853-863.

Kurt, N., W. R. Scott, C. A. Schiffer and T. Haliloglu, 2003b, "Cooperative fluctuations of unliganded and substrate-bound HIV-1 protease: a structure-based analysis on a variety of conformations from crystallography and molecular dynamics simulations", *Proteins*, Vol. 51, pp. 409-422.

Kuznetsov, I. B. and S. Rackovsky, 2002, "Discriminative ability with respect to amino acid types: assessing the performance of knowledge-based potentials without threading", *Proteins*, Vol. 49, pp. 266-284.

Lazaridis, T. and M. Karplus, 2000, "Effective energy functions for protein structure prediction", *Current Opinion in Structural Biology*, Vol. 10, pp.139-145.

Leach, A. R., 2001, *Molecular Modeling: Principles and Application*, Prentice Hall.

Levy, Y., P. G. Wolynes and J. N. Onuchic, 2004, "Protein topology determines binding mechanism", *PNAS*, Vol. 101, pp. 511-516.

Liu, X., H. A. Karimi, L. Yang and I. Bahar, 2004, "Protein Functional Motion Query and Visualization", *Proceedings of the 28th Annual International Computer Software and Applications Conference (COMPSAC'04)*, pp.86-89.

Lin, Y. C., Z. Beck, T. Lee, V. D. Le, G. M. Morris, A. J. Olson, C. H. Wong and J. H. Elder, 2000, "Alteration of substrate and inhibitor specificity of feline immunodeficiency virus protease", *Journal of Virology*, Vol. 74, pp. 4710-4720.

Lockless, S. W. and R. Ranganathan, 1999, "Evolutionarily conserved pathways of energetic connectivity in protein families", *Science*, Vol. 286, pp. 295-299.

Luque, I., M. J. Todd, J. Gomez, N. Semo and E. Freire, 1998, "Molecular basis of resistance to HIV-1 protease inhibition: A plausible hypothesis", *Biochemistry*,

Vol. 37, pp. 5791-5797.

Mammano, F., C. Petit and F. Clavel, 1998, "Resistance-associated loss of viral fitness in human immunodeficiency virus type 1: phenotypic analysis of protease and gag coevolution in protease inhibitor-treated patients", *Journal of Virology*, Vol. 72, pp. 7632-7637.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. J. Teller, 1953, "Equation of state calculations by fast computing machines", *Journal of Chemical Physics*, Vol. 21, pp. 1087-1092.

Micheletti, C., F. Cecconi, A. Flammini and A. Maritan, 2002, "Crucial stages of protein folding through a solvable model: Predicting target sites for enzyme-inhibiting drugs", *Protein Science*, Vol. 11, pp. 1878-1887.

Micheletti, C., P. Carloni and A. Maritan, 2004, "Accurate and efficient description of protein vibrational dynamics: Comparing molecular dynamics and Gaussian models", *Proteins: Structure, Function and Genetics*, Vol. 55, pp. 635-645.

Mirny, L. and E. Shakhnovich, 2001, "Protein folding theory: From lattice to all-atom models", *Annual Review of Biophysics and Biomolecular Structure*, Vol. 30, pp. 361-396.

Moore, M. L. and G. B. Dreyer, 1993, "Structure-based inhibitors of HIV-1 protease", *Perspectives in Drug Discovery and Design*, Vol. 1, pp. 85-108.

Ota, N. and D. A. Agard, 2005, "Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion", *Journal of Molecular Biology*, Vol. 351, pp. 345-354.

Ozer, N., T. Haliloglu and C. A. Schiffer, 2006, "Substrate Specificity in HIV-1 Protease by a Biased Sequence Search Method", *Proteins: Structure, Function, and Bioinformatics*, Vol. 64, pp. 444-456.

Perryman, A. L., J. Lin and J. A. McCammon, 2004, "HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs", *Protein Science*, Vol. 13, pp. 1108-1123.

Pettit, S. C., G. J. Henderson, C. A. Schiffer and R. Swanstrom, 2002, "Replacement of the P1 amino acid of human immunodeficiency virus type 1 Gag processing sites can inhibit or enhance the rate of cleavage by the viral protease", *Journal of Virology*, Vol. 76, pp. 10226-10233.

Pokala, N. and T. M. Handel, 2001, "Review: Protein Design-Where We Were, Where We Are, Where We're Going", *Journal of Structural Biology*, Vol. 134, pp. 269-281.

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 protease", *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. 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, pp. 1306-1315.

Prabu-Jeyabalan, M., E. A. 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, pp. 12446-12454.

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 Chemother, Vol. 50, pp. 1518-1521.

Russ, W. P. and R. Ranganathan, 2002, "Knowledge-based potential functions in protein design", *Current Opinion in Structural Biology*, Vol. 12, pp. 447-452.

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.

Schueler-Furman, O., Y. Altuvia, S. Alessandro and H. Margalit, 2000, "Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles", *Protein Science*, Vol. 9, pp. 1838-1846.

Scott, W. R.P. and C. A. Schiffer, 2000, "Curling of Flap Tips in HIV-1 Protease as a Mechanism for Substrate Entry and Tolerance of Drug Resistance", *Structure*, Vol. 8, pp. 1259-1265.

Sipply, M. J., 1990, "Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins", *Journal of Molecular Biology*, Vol. 213, pp. 859-883.

Stoll, V., W. Qin, K. D. Stewart, C. Jakob, C. Park, K. Walter, R. L. Simmer, R. Helfrich, D. Bussiere, J. Kao, D. Kempf, H. L. Sham and D. W. Norbeck, 2002 "X-ray crystallographic structure of ABT-378 (lopinavir) bound to HIV-1 protease", *Bioorganic & Medicinal Chemistry*, Vol. 10, pp. 2803-2806.

Surleraux, D. L., A. Tahri, W. G. Verschueren, G. M. Pille, H. A. de Kock, T. H. Jonckers, A. Peeters, S. De Meyer, H. Azijn, R. Pauwels, M. P. de Bethune, N. M. King, M. Prabu-Jeyabalan, C. A. Schiffer and P. B. Wigerinck, 2005, "Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor", *Journal of Medicinal Chemistry*, Vol. 48, pp. 1813-1822.

Sel, G. M., S. W. Lockless, M. A. Wall and R. Ranganathan, 2003, "Evolutionarily con-

served networks of residues mediate allosteric communication in proteins", *Nature Structural Biology*, Vol. 10, pp. 59-69.

Tang, S., J. Liao, A. R. Dunn, R. B. Altman, J. A. Spudich and J. P. Schmidt, 2007, "Predicting allosteric communication in myosin via a pathway of conserved residues", *Journal of Molecular Biology*, Vol. 373, pp. 1361-1373.

Tirion M. M., 1996, "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis", *Physical Review Letters*, Vol. 77, pp.1905-1908.

Trylska, J., V. Tozzini, C. A. Chang and J. A. McCammon, 2007, "HIV-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics", *Biophysical Journal*, Vol. 92, pp. 4179-4187.

Tsai, C., A. del Sol and R. Nussinov, 2008, "Allostery: Absence of a change in shape does not imply that allostery is not at play", *Journal of Molecular Biology*, Vol. 378, pp. 1-11.

UNAIDS, 2007, *AIDS epidemic update: December 2007*, WHO Library Cataloguing-in-Publication Data.

Vendruscolo, M., N. V. Dokholyan, E. Paci and M. Karplus, 2002, "Small-world view of the amino acids that play a key role in protein fold", *Physical Review E*, Vol. 65, pp. 0619101-0619104.

Wang W. and P. Kollman, 2001, "Computational study of protein specificity: The molecular basis of HIV-1 protease drug resistance", *PNAS*, Vol. 98, pp. 14937-14942.

Weber, I. T. and R. W. Harrison, 1999, "Molecular mechanics analysis of drug-resistant mutants of HIV protease", *Protein Engineering*, Vol. 12, pp. 469-474.

Weiss, R. A., 1993, "How does HIV cause AIDS?", *Science*, Vol. 260, pp. 1273-1279.

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

Wlodawer, A. and J. Vondrasek, 1998, "Inhibitors of HIV-1 protease: A major success of structure-assisted drug design", *Annual Review of Biophysics and Biomolecular Structure*, Vol. 27, pp. 249-284.

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

Yang, L., G. Song, A. Carriquiry and R. L. Jernigan, 2008, "Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes", *Structure*, Vol. 16, pp. 321-330.

Zoete, V., O. Michielin and M. Karplus, 2002, "Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility", *Journal of Molecular Biology*, Vol. 315, pp. 21-52.