UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering, Science and Mathematics

School of Chemistry

The Study of Conformational Motions using Enhanced Sampling Techniques

by Sarah L Williams

A thesis submitted for the qualification of

Doctor of Philosophy at the University of Southampton

June 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

SCHOOL OF CHEMISTRY

Doctor of Philosophy

THE STUDY OF CONFORMATIONAL MOTIONS USING ENHANCED SAMPLING TECHNIQUES

by Sarah L Williams

Computational methods are used for the study of conformational change in biological molecules. Conventional molecular dynamics (MD) has been shown to frequently fail to sample the long timescales over which many biological processes occur, and numerous methods have been developed to overcome this problem. In this thesis, two such methods, in addition to MD, are used to study the conformations of biological molecules. The enhanced sampling method of parallel tempering (PT) is used to investigate the *trans-cis* isomerism of peptide bonds in two small cyclic peptide systems, and reversible digitally filtered molecular dynamics (RDFMD), a method which enhances conformational dynamics by selectively enhancing or suppressing the vibrational motion of specific regions of interest in molecules, is used to investigate the effects on conformation as a result of mutation in the HIV-1 enzymes, HIV-1 protease (HIV-1 PR) and HIV-1 integrase (HIV-1 IN) in the apo state. Additionally, the method has been used to investigate the effect of mutation on the conformational dynamics when HIV-1 PR is bound to inhibitors.

In simulations where the enhanced sampling methods of PT and RDFMD have been applied in this thesis, increased conformational sampling compared with MD simulations has been observed. PT simulations are found to be computationally expensive and perhaps unsuitable for application to large protein systems. The RDFMD simulations are able to efficiently accelerate infrequent large-scale conformational changes, thereby revealing new conformations which were not present in simulations carried out using conventional MD, and provide a greater understanding of the mechanism and the effect of mutation on the HIV-1 PR and HIV-1 IN enzymes.

Acknowledgements

Firstly, I would like to take the opportunity to thank Jonathan Essex for all his help, support and enthusiasm throughout this research. Additionally, many thanks goes to the entire Essex group, my family and friends, all of whom have assisted me in some way throughout these years of study.

A special mention goes to the following past and present group members; Adrian Wiley, for his invaluable help, especially through the early stages, Chris Woods, for his advice and for rewriting the RDFMD code into something far more user-friendly! Thanks also to the group system administration team, Juan Carmona-Fernandez, Mishal Patel and Julien Michel for keeping the group computers in check, and Ivan Wolton and all the others who maintain and support the university computer clusters, which I have used extensively over the years.

Finally, a special thanks goes to my husband Paul for his love, patience and support.

Contents

Chapter 1

Introduction

The study of the conformational change of biological molecules using computational methods is of great interest [1]. Proteins are not static structures and conformational change is vital to their function [2]. A number of experimental techniques exist which are able to study the structure of a protein, such as X-ray crystallography, which is able to provide snapshots of protein structures, while NMR can provide a time-resolved dynamic structure, with the ability to resolve flexible regions, which may not be possible using X-ray crystallography. However, detailed experimental data can be difficult to obtain, especially for extreme conditions, whereas computer simulations permit the study of complex, dynamic processes that occur in biological systems on an atomic scale.

A drawback of using the conventional computational methods, such as molecular dynamics (MD), is the timescales which can be simulated. The timescale of motions of biological systems can occur on a millisecond scale, and despite the advances in computers, conventional techniques fail as they are unable to simulate these systems for a period long enough to observe large-scale conformational changes. Typically, the potential energy surface of a protein is rugged and low temperature simulations become trapped in local energy minima. A number of enhanced sampling methods have been developed in an attempt to overcome this problem, two of which are presented in this thesis, parallel tempering (PT) [3] and reversible digitally filtered molecular dynamics (RDFMD)[4][5][6].

In this thesis, the enhanced sampling technique of RDFMD is applied to Human Immunodeficiency Virus-1 proteins (HIV-1), and PT is applied to cyclic peptides. The results are compared with those generated by classical molecular dynamics and show that improved sampling is obtained with the use of these enhanced sampling techniques.

The aim of the PT method is to perform several concurrent independent simulations of the same system at a range of temperatures and periodically attempt to swap configurations between neighbouring temperatures. The higher temperature replicas allow simulations to overcome the energy barriers, whilst the lower temperatures allow extensive sampling of the energy minima. This method has been applied to two cyclic peptides to study the *trans-cis* isomerism of peptide bonds. On comparison of the PT results with experimental data, varied levels of success were demonstrated owing to deficiencies in the force field used. In addition, the simulation was found to be computationally expensive, even for the small size of the cyclic peptides studied here, perhaps indicating this method to be unsuitable for the simulation of larger protein systems.

The second method, RDFMD, is a non-equilibrium method where energy is added to specific regions of the protein of study. This method has previously been successfully used in the study of the dynamics of protein systems *Escherichia coli* dihydrofolate reductase (EcDHFR) [6] and T4 lysosyme (T4L) [7]. In this thesis, the method has been applied to two HIV-1 enzymes, HIV-1 protease (HIV-1 PR) and HIV-1 integrase (HIV-1 IN). In both enzymes, the presence of mutations have been reported which affect the binding of substrate and/or inhibitors. The mechanism by which some of these mutations operate has been reported to be through alteration of the conformational dynamics of structural regions of the enzymes which are of importance to the catalytic mechanism. RDFMD has been used to investigate the conformational dynamics of the wild-type (WT) and two mutant (M46I and V82F/I84V) forms of apo HIV-1 PR and the WT and one mutant (G140A/G149A) of the HIV-1 IN enzyme. The results reveal several major biologically relevant conformational changes and demonstrate significant differences in the conformational sampling between the WT and mutant forms of these apo-enzymes, thus suggesting possible mechanisms for these mutants.

Additionally, the inhibitor-bound (saquinavir and indinavir) WT and mutant HIV-1 PR systems have been studied with the use of the RDFMD technique. Owing to the

increased bonding between the inhibitor and the flaps, significant conformational change is not observed and as a consequence, significant differences between the conformational sampling of the WT and mutants are not detected, although differences between the dynamics of HIV-1 PR when bound to different inhibitors is shown.

Chapter 2

Computer Simulation

Protein dynamics plays an important role in the majority of biological processes [2], and the ability of proteins to change conformation is essential for their function. Therefore, a thorough knowledge of the principles governing the dynamics of proteins would greatly facilitate the understanding of these processes. Additionally, it would enhance the possibilities of modifying the dynamic processes of proteins by mutation for industrial or medical purposes.

The majority of information concerning the structure and dynamics of proteins is obtained from experimental techniques, and owing to advances in crystallography, the three-dimensional structure of many proteins have been revealed in atomic detail [8]. However, proteins are not static and it is difficult to track all the different viable conformations experimentally, as it is difficult to isolate proteins in their high energy states long enough to acquire an accurate structure. Some experimental techniques, e.g. NMR can study dynamics, but the process is limited by the size of the system due to the complexity of the spectra and poor sensitivity [9]. Since these marginally stable non-native conformations are difficult to determine experimentally, computer simulation has proven to be useful, giving detailed atomic level representation and high resolution on a picosecond to microsecond timescale.

Computer simulation has become an increasingly important tool in the study of biological systems as computers have become more powerful and available more cheaply [10]. The most popular methods of computer simulation are molecular mechanics (MM) and

quantum mechanics (QM). MM refers to the use of Newton's laws to model molecular systems, where the potential energy of systems is calculated through the use of force fields. The lowest level of study is the atom, or groups of atoms, which is in contrast to QM, where electrons are studied explicitly. MM thus has the ability to consider a greater number of particles compared with QM, and is routinely used to investigate the structure, dynamics and thermodynamics of biological systems [11].

2.1 Force Fields

In molecular mechanics, a force field is used to describe molecular structures and their properties. It refers to the functional form and the parameter sets used to describe the potential energy of a system. All-atom force fields use parameters for every atom in the system, whereas united-atom force fields do not include unimportant atoms, usually the apolar hydrogen atoms. The neglected atoms are accounted for by increasing the size of the atoms to which they are bonded, usually by increasing the van der Waals radii. This type of representation is useful when the intermolecular motion of the system is more important than the intramolecular motion and is also capable of modelling simple packing effects i.e. van der Waals interactions [12]. However, owing to the absence of the hydrogen atoms, a simplified representation of the electrostatic effects results. Further reduced representation of systems can be achieved with coarse-graining force fields, which are useful in the simulation of large systems by increasing computational efficiency.

There are various different types of force field, among the most popular used being AMBER [13] (Assisted Model Building and Energy Refinement), CHARMM (Chemistry at Harvard Macromolecular Mechanics, united-atom, CHARMM19 [14], protein all-atom, CHARMM22 [15], DNA, RNA and lipids: CHARMM27 [16]) and GROMOS (GROningen MOlecular Simulation Software) [17] [18] [19]. The energy terms incorporated vary in different force fields, but most generally include terms which are used to describe the various intra- and intermolecular forces (equation 2.1), and some additionally include other energy terms. Since all force fields are based on numerous approximations and are derived from different experimental data, parameters from different force fields cannot be combined with each other.

$$U(r^N) = \sum_{bonds} \frac{k_i}{2}(\mathrm{l}_i\text{-}\ l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\omega - \gamma)) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^6] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}))$$

(2.1)

The first two terms in equation 2.1 are based on Hooke's Law, a harmonic potential, which describes the stretching and bending energy of the bond respectively. The k_i parameter describes the stiffness of the bond/angle, whereas $l_{i,0}$ and $\theta_{i,0}$ define the equilibrium length/angle; penalties are incorporated for deviations (denoted by l_i for bonds and θ_i for angles in equation 2.1) from the equilibrium bond lengths and angles. A description of how the energy varies as the bonds are rotated, the torsion energy, is calculated in the third term. The torsion potential is described by the Fourier series where the sum is taken over all the dihedral angles. It is modelled as a simple periodic function, with V_n as a constant reference energy, n representing the periodicity, ω being the torsion angle spanned by the 4 atoms, and γ is the phase shift along the rotation axis.

The fourth term is the non-bonded term, which is modelled using a Lennard-Jones potential for van der Waals interactions (first bracketed section) and a Coulomb potential for the electrostatic contribution (second bracketed section).

Van der Waals interactions are short-range and die off rapidly as the atoms move away from each other, and repulsion occurs when the distance between the interacting atoms (i and j) becomes marginally less than the sum of their contact radii (σ). This is modelled by the $1/r^{12}$ dependency, which is designed to rapidly increase at close distances between atoms. The Lennard-Jones potential (6-12 equation) is used to describe the

attraction/repulsion for the smooth transition between the states, the attractive force is proportional to r^{-6} and the repulsion varies as r^{-12}, with ε related to the well-depth.

The electrostatic contribution is modelled using a Coulombic potential. This is an effective pair potential that describes the interaction between two point charges q_i and q_j, where r is the distance between two charges and ε_0 relates to the electronic permittivity of free space. Two opposite charges give a negative potential energy, whereas ions of the same charge sign are repulsive and give a positive potential energy. The electrostatic term is harder to calculate as it does not drop off rapidly over distance, as do the van der Waals attractions. Additionally, long-range electrostatics are often of interest in simulations. A basic method to deal with the calculation is to apply a cutoff, but this results in an abrupt discontinuity. An improved method is switching, where the electrostatics are smoothed between the inside and outside of the cutoff radii. More improved and complex methods also exist e.g. Particle Mesh Ewald (PME) [20] where the calculation of electrostatic forces and energies are separated into short- and long-range interactions. The short-range interactions are calculated explicitly, whereas the long-range interactions are summed in Fourier space.

The parameters of atoms (includes atomic mass, partial charges, equilibrium values for bond lengths and angles), together with the above potential energy terms constitute the force field. For the more commonly used force fields (e.g. CHARMM [21], AMBER [13], GROMOS [17] [18] [19]), parameters are often calibrated to experimental results (e.g. interaction energies can often be obtained by mass spectrometry) and quantum mechanical calculations of small organic model molecules (e.g. calculations of conformational energies and hydrogen bonding [22]), and their ability to reproduce physical properties measurable by experiment is tested. The calculation of the free energies of solvation is one such suitable test for a force field, since accurate experimental values are available [22]. However, many simulations involve biological macromolecules e.g proteins, so the parameterisation is not fully accurate. Alternative force fields have been developed, which are derived from other types of experimental data, e.g. enthalpy of sublimation and enthalpy of vapourisation (e.g. OPLS [23] [24]).

Limitations in current empirical force fields exist, for example, they are unable to model bond formation or breakage, thus making these force fields inadequate for modelling

chemical reactions. Additionally, since the atom parameters are considered as constants, electronic polarisability is typically excluded. To address these limitations, mixed quantum mechanical - molecular mechanical force fields (e.g. reactive force fields [25] [26]) are currently under development.

In the work of this thesis, the all-atom force fields of CHARMM22 [15] and AMBER99 [27] have been used.

2.2 Monte Carlo Simulations (MC)

Monte Carlo methods are stochastic techniques, using random numbers and probability statistics. There are two basic approaches to MC methods, direct sampling and Markov-chain sampling [28]. Usually, direct sampling is not feasible and the Markov-chain sampling approach is employed. The method introduces large changes to a system, allowing random moves from one configuration to another, then determining whether the new configuration is energetically feasible or not by determining whether they satisfy the detailed-balance. The Metropolis criteria [29] is used to ensure detailed-balance is satisfied, where the generated Markov-chain must satisfy the following conditions: the outcome of each trial move must only depend on the preceding state, not on any other, and each move belongs to a finite set of outcomes [30].

The method enables the crossing of energy barriers, as only the energy of the new configuration relative to the previous is considered when accepting or rejecting a move.

A disadvantage of MC methods is the absence of a time variable, meaning time-dependent properties cannot be obtained. However, this method can be more efficient compared with molecular dynamics in the calculation of the thermodynamic properties, as molecular dynamics simulations can become trapped in local energy minima, failing to sample a significant amount of phase-space.

2.3 Molecular Dynamics (MD)

In contrast to Monte Carlo techniques, molecular dynamics is able to calculate the time-dependent behavior of a molecular system. Once the positions and velocities of the

atoms are known, the state of the system at any time in the past or future can be predicted by integrating the equations of motion of a set of interacting atoms. These simulations allow the detailed study of processes of complex systems e.g. their dynamic conformational change and protein folding.

In MD, the method is based on classical molecular mechanics, most notably Newton's second law, where the force exerted on each atom, i, residing in a system consisting of N atoms equals the mass of the atom, m_i multiplied by the acceleration of the atom, a (equation 2.2).

$$\underline{\mathrm{F}}_i = m_i \underline{\mathrm{a}}_i \tag{2.2}$$

The force can also be expressed as the gradient of the potential energy (equation 2.3), and the combination of equations 2.2 and 2.3 results in equation 2.4, from which the positions of atoms in time can be related to the derivative of the potential energy (V) by Newton's equations of motion.

$$\underline{\mathrm{F}}_i = -\nabla V \tag{2.3}$$

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \tag{2.4}$$

Therefore, from a knowledge of the force, it is possible to determine the acceleration of each atom in the system, by calculating the gradient of the potential energy function. Positions and velocities can then be determined as demonstrated in equation 2.5, where acceleration (a) is shown to be the first derivative of velocity (v) with respect to time, and the second derivative of position with respect to time.

$$\underline{\mathrm{a}}_i = \frac{d\underline{v}_i}{dt} = \frac{d^2 r_i}{dt^2} \tag{2.5}$$

This integration of the equations of motion results in a trajectory which describes the positions, accelerations and velocities of the atoms as a function of time. It is from this trajectory that the average properties can be determined.

The force acting on each atom changes whenever the position of the atom or other atoms present in the system change. All the atoms are subject to Newton's law, but the motions of each of the atoms are coupled, so the equations of motion must be calculated simultaneously, which is not analytically soluble for three atoms, known as the three-body problem. Since the equations of motion cannot be solved analytically, they are solved using a finite difference method, which assumes that positions and dynamic properties can be calculated using a Taylor series.

One of the simplest finite difference methods is the Verlet algorithm [31]. However, the use of this algorithm has disadvantages as it does not involve the use of velocities and requires the positions of the current and previous timesteps to calculate the next set of atom positions (i.e. it is not a self-starting algorithm).

Several other algorithms have been developed which improve on the Verlet algorithm [31], including the leap-frog [32], Beeman [33] and velocity Verlet algorithms. The velocity Verlet algorithm [34] is the integrator used in the simulations presented in this thesis. This algorithm calculates the velocities, positions and accelerations at time t, and is time-reversible, which is an essential requirement of the reversible digitally filtered molecular dynamics (RDFMD) [4] [5] [6] enhanced sampling method used in this thesis. The steps involved in the calculation are shown in equations 2.6 to 2.10.

$$\underline{a}(t) = -(\frac{1}{m})\nabla V(\underline{r}(t)) \tag{2.6}$$

$$\underline{r}(t+\delta t) = \underline{r}(t) + \underline{v}(t)\delta t + \frac{1}{2}\underline{a}(t)\delta t^2 \tag{2.7}$$

$$\underline{v}(t+\frac{\delta t}{2}) = \underline{v}(t) + (\frac{1}{2})\underline{a}(t)\delta t \tag{2.8}$$

$$\underline{a}(t+\delta t) = -(\frac{1}{m})\nabla V(\underline{r}(t+\delta t)) \tag{2.9}$$

$$\underline{v}(t+\delta t) = \underline{v}(t+\frac{\delta t}{2}) + (\frac{1}{2})\underline{a}(t+\delta t)\delta t \tag{2.10}$$

When choosing an integration algorithm, the following considerations should be taken into account: its stability (i.e. conservation of the energy of the system), its accuracy, speed and computer efficiency. The stability of the integrator algorithm is dependent on the timestep used. The biologically relevant behaviour of systems takes place mostly on relatively large time scales, often in the range of seconds, but time steps of a few femtoseconds are usually used to conserve the energy of the system. The small time step means that the trajectory covers an insufficient proportion of the phase space, whereas, the use of a large time step leads to forces changing too quickly, instabilities in the integration algorithm and degradation in energy conservation due to high energy overlaps between atoms. Therefore, there is a trade-off between accuracy and economy.

A solution is to dampen the high-frequency vibrations by constraining the bonds involved to their equilibrium values, while still allowing the rest of the degrees of freedom to vary under the intramolecular and intermolecular forces present. The most commonly used method is the SHAKE algorithm [35], where the equations of motion are solved, whilst simultaneously satisfying the imposed constraints.

2.4 Physical Properties

A variety of thermodynamic properties can be calculated from simulations, the values of which can be compared to experimental values as a method of determining the accuracy of the simulation. In addition, values for which no experimental data exists, or is difficult to attain, can be calculated from the simulation. Some of the properties which are routinely calculated include energy, heat capacity, temperature and pressure.

The thermodynamic state of a system is usually defined by a number of parameters which include the temperature, volume, pressure and number of particles which make

up the system. These conditions characterise a particular thermodynamic state, called an ensemble.

A number of different ensembles are used, the microcanonical ensemble (NVE), where the number of atoms (N), the volume (V) and the total energy (E) are fixed. A simulation using these parameters is considered as an adiabatic process with no heat exchange [36], with the generated trajectories being considered as exchanges in kinetic and potential energy, with no change in the total energy. In molecular dynamics simulations, it is of more interest to simulate in the canonical ensemble, where the temperature (T), number of number of atoms (N) and volume (V) are kept constant, or isobaric-isothermal ensemble (NPT), where it is the pressure that is kept fixed, rather than the volume as in the NVT ensemble.

The integrated equations of motion generate trajectories in the microcanonical (NVE) ensemble; therefore, adjustments to the equations are required to simulate in the NVT and NPT ensembles.

In simulating the canonical ensemble, there are different approaches to controlling the temperature which add and remove energy from the simulation in a realistic manner. One method is to rescale velocities, as the temperature of the system is associated with the kinetic energy of the atoms e.g. the instantaneous kinetic energy is calculated as shown in equation 2.11 where m is the mass of the atoms, i.

$$E_{kin} = \sum_{i=1}^{N} m_i \underline{v}_i^2 / 2 \tag{2.11}$$

and the velocity ($\underline{v}$) of each atom is rescaled by the same factor to satisfy $E_{kin} = 3/2 k_B T$, where k_B is the Boltzmann constant, T is temperature.

An alternative simple method of controlling temperature and pressure is the weak coupling system [37], where the particles of the system are coupled to a "bath" of constant temperature via some suitable coupling parameters. The advantage of these methods is that they are simple and cause little perturbation of the system. However, these methods do not generate a rigorous statistical mechanical ensemble, although whether this is significant for most structural properties is debatable.

More rigorous methods exist, such as the use of the Langevin or Nosé-Hoover thermostats [38] [39] [40] to control the temperature. The Langevin thermostat [41] follows the Langevin equations of motion [42] rather than Newton's equations of motion.

$$m\underline{a} = -\zeta \underline{v} + \underline{F} + \underline{F}' \tag{2.12}$$

Equation 2.12 explains the application of the Langevin thermostat to control temperature, where m is the mass of the atom, $\underline{a}$ is acceleration, $\underline{F}$ is the conservative force acting on the atom, and $\underline{v}$ is the velocity of the atom, ζ is a frictional constant and $\underline{F}'$ is a random force. The frictional force decreases the temperature because ζ is a fixed positive value. The random force is randomly selected from a Gaussian distribution and adds kinetic energy into the particles of the system, with its variance being the function of the selected temperature and timestep. This results in the random force being balanced with the frictional force and maintains the system temperature at the desired value.

Pressure control requires the use of periodic boundary conditions in simulations. Many methods which are used for controlling the temperature can be modified to control the pressure e.g. Berendsen [43], where the system is coupled to a pressure bath, and the pressure is controlled by dynamically adjusting the size of the unit cell and rescaling the coordinates of non-fixed atoms during the simulation. These algorithms can be applied isotropically, where the unit cell maintains its shape or anisotropically, where all three dimensions of the unit cell are allowed to fluctuate.

2.5 Periodic Boundary Conditions (PBC)

In simulations where it is the solute which is of concern, calculating a large number of solvent molecules is of little interest and computationally expensive, and therefore, periodic boundary conditions (PBC) are used.

By use of PBC, simulations can be carried out using a relatively small number of particles in such a way that the particles experience forces as if they were in bulk fluid. The cubic

box is the easiest to simulate, although, different shapes may be considered to be more appropriate for a simulation. During a simulation, the protein and solvent are contained in the box which is replicated through space. Only the coordinates of the central box being stored, since all the boxes are identical. This eliminates any unwanted surface effects; if a molecule leaves the box, one of its images enters through the opposite side. This way, no surface, and thus no effects, exist.

2.6 Solvent Models

Modelling solvated systems requires models for the interaction of the solvent with the solute, and for this purpose, both implicit and explicit solvent models have been developed.

There are several types of explicit water models, which vary according to their complexity.

The simple water models comprise three particles possessing rigid bonds and angles and have between 3 and 5 interaction points, with the greater number of interaction points resulting in a greater computational expense. In models with three interaction points (e.g. TIP3P [44] and SPC [45]), the interaction between two water molecules is calculated using a Lennard-Jones function at just one interaction point per molecule centered on the oxygen atom, with no van der Waals interactions involving the hydrogen atom.

Other, more complex water models exist, including those which possess flexibility and allow for polarisability e.g. TIP4P [46] and TIP5P [47].

Explicitly solvated systems contain hundreds of solvent molecules, for which, solvent-solvent interactions are considered, causing the calculations to converge slowly. Owing to this associated computational expense of explicit models, especially for systems with a large number of degrees of freedom, there is a significant interest in the development of implicit solvent models.

Implicit solvent models treat the solvent as a continuous medium, surrounding the solute beginning at the van der Waals surface and having the average properties of the actual

solvent [48]. With the use of these models, viscosity is not incorporated as it is with explicit models, meaning that the molecule can sample conformational space in shorter time. Therefore, simulations carried out in implicit and explicit solvent possess different energy landscapes. Additionally, the implicit model value is for an infinite volume of solvent, thus removing any errors associated with the use of periodic conditions.

The most physically correct implicit solvent model represents the solvent as a highly polarisable dielectric continuum and the electrostatic potential is obtained by solving the Poisson Boltzmann (PB) equation [49][50][51]. The PB equation gives the exact theoretical treatment of the electrostatic interactions in continuous media. This equation describes the electrostatic interactions between molecules in ionic solutions. The use of these equations has been rigourously tested for their use in calculation of continuum electrostatics, however, they are computationally expensive to compute and so other methods have been developed e.g. Generalised Born (GB) models [52] [53] [54].

The GB approximation is based on the PB theory, but the solution to the electrostatic potential is replaced by an approximation of the solvent-induced reaction field energy. The GB model comprises of a system of particles with effective Born radii, a_i, (derived using a complex procedure for each atom which has a charge or partial charge) and charges, q_i. The total electrostatic free energy (G_{elec}) of such systems (equation 2.13) is a combination of the sum of the Coulomb energies and the Born free energy of solvation in a medium of relative permittivity, ε. The $(1 - \frac{1}{\varepsilon})$ term represents the change in dielectric constant, going from vacuum to solvent (dielectric medium).

$$G_{elec} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{\varepsilon r_{ij}} - \frac{1}{2}(1 - \frac{1}{\varepsilon}) \sum_{i=1}^{N} \frac{q^2 i}{a_i} \tag{2.13}$$

Explicit solvation is the more accurate method of modelling solvent as it realistically models the interaction of the solute with the solvent [55] [56], but has the limitation of computational expense, thus making implicit solvation methods an attractive alternative, especially in the simulation of large systems. However, owing to the absence of explicit solvent in using this approximation, the assumption is made that the solvent molecules reach equilibrium at a much faster time-scale than any major structural reorganisation of the solute and any specific solute-solvent interactions are neglected.

2.7 Analysis Methods

In the study of the protein systems presented in this thesis, a number of analysis tools have been used. The main methods of analysis include principal component analysis (PCA), Root Mean Squared Fluctuations (RMSF), Root Mean Squared Deviations (RMSD) and distance/angle measurements.

Principal Component Analysis (PCA)

This is a method of statistical analysis which simplifies a dataset by reducing its dimensions to the representative principal components, hopefully without losing any of the important characteristics. It captures the variance of the dataset in terms of principal components, with the first principal component (eigenvector with the highest associated eigenvalue) contributing the most to the variance in the dataset. However, the relative shortness of MD simulations compared with the time scales on which motion occurs imposes constraints on the PCA method, as it is not able to extrapolate motions to a timescale at which the conformational change is believed to occur.

The first step in the PCA process is the generation of the dataset. In the analysis of the simulations in this thesis, the dataset comprises the trajectory coordinates (x, y, z) for each of the atoms to be studied. The average structure is then subtracted across each dimension (number of dimensions is equal to the number of atoms multiplied by 3), generating a dataset whose mean is zero. Following this, the covariance matrix is calculated.

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{Y})}{(n-1),} \tag{2.14}$$

Covariance ($cov(X,Y)$) is always measured between two dimensions (equation 2.14, where $\bar{X}$ and $\bar{Y}$ are the average structures of dimensions X and Y and n is the number of dimensions). With more than two dimensions, there is more than one covariance measurement to be calculated. For example, in the case of three dimensions, covariance calculations for (x,y), (x,z) and (y,z) are carried out. For n dimensions, the number of

covariance calculations to be carried out equals the solution (N_{calc}) to equation 2.15. To obtain all the covariance values between all the dimensions, they are entered into one matrix where each entry into the matrix is the result of calculating the covariance between two separate dimensions. The resultant square matrix comprises of n rows and columns. Along the diagonal of the matrix is the covariance between itself and a dimension, giving the variance for that dimension.

$$N_{calc} = \frac{n!}{(n-2)!.2} \tag{2.15}$$

From the covariance matrix, the eigenvectors and eigenvalues are calculated. Their determination is equivalent to matrix diagonalisation where each eigenvector is paired with a corresponding eigenvalue.

Eigenvalues are closely related to eigenvectors and are associated with a particular eigenvector, with n eigenvectors existing for a dataset with n dimensions.

Once the eigenvectors have been generated, they can be ordered by eigenvalue, from highest to lowest, giving their significance to the total motion of the protein. The eigenvector with the highest eigenvalue is the first principal component of the dataset.

Only the first few of the total number of eigenvectors are studied, as the majority are of little importance. The number of eigenvectors to study is an arbitrary decision, but some criteria do exist, one being the Kaiser criterion [57]. This procedure only retains eigenvectors which have eigenvalues greater than 1. An alternative method is the Cattell Scree test [58], which uses simple line plot representations of the eigenvalues displaying the relative importance of each component in fitting the data to major conformational reorganisation. Cattell suggests that the point at which the line smooths out is selected as the cutoff for the eigenvector choice, disregarding principal components to the right of this point as 'factorial scree'. Generally, it has been found that the former method retains too many factors and the latter, too few, but both are found to be acceptable for use under normal conditions where there are relatively few factors. However, these are only guidelines and in practice, the extent to which a solution is interpretable, giving sensible results, is additionally taken into account.

The similarity between the eigenvectors can be quantified by calculation of the inner-products [59][60][61][62]. The averaged conformation of each of the two simulations for which the eigenvectors are to be compared are taken and bought into best-fit positions, reorientating the eigenvectors accordingly. The inner-product is then calculated as in equation 2.16, where i and j represent the eigenvectors to be compared of vector sets w and u respectively.

$$C_{ij} = w_i.u_j \tag{2.16}$$

If the two eigenvectors are identical, the inner-product must be unity due to normality. A value of 0 indicates no similarity between the eigenvectors.

The sum of the squared inner-products of two sets of N vectors can be compared by calculation of the root mean squared inner-product (RMSIP), as shown in equation 2.17.

$$RMSIP = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}(w_i.u_j)^2}, \tag{2.17}$$

where w_i represents the i^{th} eigenvector associated with data set w and u_j represents the j^{th} eigenvector associated with data set u. This value can be used to evaluate the overlap of the conformational space spanned by the different systems.

In this thesis, PCA has been performed using the g_covar and g_anaeig modules of GROMACS 3.2.1 [63].

Root-Mean Squared Deviation (RMSD)

This is a measure of the average distance between selected atoms and a reference structure, where δ is the distance between N atoms (usually α-carbons, or other atoms of the backbone) (equation 2.18).

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2} \tag{2.18}$$

Root Mean Squared Fluctuations (RMSF)

This is a measure of the deviation of the position of the particle x_i at time t_j and some reference point $\tilde{x}_i$, where the positions for each particle are averaged over time (T) (equation 2.19).

$$RMSF = \frac{1}{T}\sum_{t_j=1}^{T}(x_i(t_j) - \tilde{x}_i)^2 \tag{2.19}$$

Distance and Angle Measurements

Distance and angle measurements performed in this thesis have been carried out using the VMD software [64].

Chapter 3

Conformational Dynamics of Proteins

Proteins may change their conformation in order to undergo their biological function e.g. catalysis and the regulation of activity, and the mechanism for their conformational change may be a result of a number of reasons e.g. binding of a ligand, pH, temperature. For example, haemoglobin (Hb) is a oxygen-transport metalloprotein, which recognises and binds to specific diatomic molecules as a result of binding to heme iron. Subsequently, a conformational change occurs and a functional reaction takes place [65].

Both experimental (e.g. X-ray crystallography) and theoretical (e.g. molecular dynamics studies) approaches exist, and are used in an attempt to acquire a more detailed knowledge of the conformations undertaken by proteins and subsequently assist in the understanding of the mechanism of function of biological systems.

3.1 Experimental Techniques

A number of experimental techniques exist which determine the structures of proteins. One of the most powerful techniques is X-ray crystallography [66]. The vast majority of structures contained in the protein databank [8] have been resolved using X-ray crystallography, the next most popular technique being NMR [67], and a few structures by electron microscopy [68].

3.1.1 X-ray Crystallography

In X-ray crystallography, the proteins of study must be obtained in large quantities and be of a high purity for the process to be successful. Many methods exist which are used to crystallise proteins, the main techniques involving the use of a buffer to control the pH to the target value, a precipitating agent and co-factors or other additives which may help in the process [69]. A general rule is that crystals must be of 4 Å resolution or lower, in order to gain any useful information.

X-rays are diffracted by interaction with electrons within the structure, and the resultant diffraction pattern is recorded, from which the coordinates and B-factors can be determined. A disadvantage of this method is that it only provides a static model of the protein, and although this is useful, the many conformations which may exist for the protein may be critical in the functioning of the protein, and not captured by X-ray diffraction.

3.1.2 Nuclear Magnetic Resonance (NMR)

Nuclear Magnetic Resonance (NMR) is the second most commonly used technique in protein structure determination. This method is able to determine structures at a high resolution, but does require a large amount of material of high-purity [67] .

Protein NMR uses multidimensional NMR experiments to obtain information about the protein. In an ideal case, each different nucleus in the molecule experiences a distinct chemical environment in the presence of the external magnetic field, and thus has a distinct chemical shift (a function of the nucleus and its environment, defined as nuclear shielding/applied magnetic field) by which it can be identified. However, in large molecules such as proteins, the number of resonances can be very large and prone to error in their interpretation owing to incidental overlaps. In these cases, multidimensional NMR experiments are performed which are able to correlate the frequencies of different nuclei. These experiments decrease the chance of overlap and provide a greater amount of information concerning the structure since they correlate nuclei frequencies within a specific part of the protein.

A number of NMR experiments exist. In cases where the protein is isotopically labelled, a 2D heteronuclear single quantum correlation (HSQC) spectrum [70] [71] is performed. In an N-15 labelled protein, amino acids typically give rise to a single peak in the spectrum, with the exception of proline which has no α-hydrogen due to the cyclic nature of its backbone. Tryptophan and certain other residues with N-containing sidechains also result in additional signals. When the protein is labelled with carbon-13 and nitrogen-15 it is possible to connect different spin systems (spin system includes the chemical and coupling constants). Several different experiments are able to do this, all of which consist of a HSQC plane expanded with a carbon dimension [72].

When the protein has not been labelled, the typical procedure is to obtain results from of a set of 2D homonuclear NMR experiments using correlation spectroscopy (COSY), of which several types exist e.g. conventional correlation spectroscopy, total correlation spectroscopy (TOCSY) and nuclear Overhauser effect spectroscopy (NOESY)[73].

In addition to using chemical-shift patterns, hydrogen-deuterium exchange can be measured, since NMR spectroscopy is nuclei specific. The amide protons in the protein exchange readily with the solvent, and if the solvent contains a different isotope, typically deuterium, the reaction can be monitored by NMR spectroscopy. The amide exchange rates can give information on the protein e.g. which parts of the protein are buried. Comparison of the exchange of a apo-form and a bound-form of a protein can provide useful information e.g. the amides that become protected in the bound-form, are assumed to be in the interaction interface.

Further to structural determination, NMR can provide information concerning the dynamics of proteins. This usually involves measuring relaxation times (consequence of the nuclei returning to equilibrium after magnetic excitation) to determine order parameters, correlation times, and chemical exchange rates. The technique is able to measure movements over a broad range of timescales, ranging from 10 ps to ms, which capture the fast and slower motions. The overall rotational diffusion can also be measured (5 -50 ns), as well as protein domain movements which take place on a very slow timescale (ms - days)[74].

In the study of protein dynamics, the nitrogen-15 isotope is the preferred nucleus because its relaxation times are relatively easier to correlate to motions. However, nitrogen atoms

are mainly found in the backbone of proteins, and as a consequence, the results focus on the motions of the backbone, which may not be representative of the dynamics of the whole protein. Therefore, techniques involving the relaxation measurements of carbon-13 and deuterium have been developed, which enable the study of motions of the amino acid side chains in proteins [75].

Until more recently, NMR spectroscopy has been limited to relatively small proteins or parts of proteins. This is partly due to problem mentioned earlier concerning the complex spectra generated by large proteins. However, a more major problem exists owing to the fast relaxation times observed for large proteins, resulting in less time to detect the signal. This consequently results in poor sensitivity as the peaks become broader and weaker, and eventually disappear, although recently techniques have been developed to overcome this [9].

Other techniques of structure determination also exist, including IR spectroscopy and powder diffraction. These methods have been used in the resolution of a few of the structures contained in the protein databank.

Additional experimental methods which have the capability to observe protein dynamics include the atomic force microscope [76], second-harmonic generation technique [77] and picosecond circular dichroism [78].

3.2 Computational Techniques

A variety of computational techniques exist, which are used to study the conformations of biological systems.

3.2.1 Non-dynamical Methods of Conformation Study

Non-dynamical methods of sampling provide conformations which are not part of a time-series of events. A popular method in the study of the collective movements of proteins is normal mode analysis (NMA) [79][80][81]. The definition of the method is the study of potential wells by analytical means. The theory is, that in the vicinity of a stationary point, the potential energy of a protein can be approximated, since the fluctuations of

the atoms are assumed to be very small. Within these approximations, the equations of motion can be solved analytically. Thus, by modelling the interatomic bonds as springs and treating the protein as a set of harmonic oscillators which are coupled, the frequency of periodic motion can be calculated for each normal mode. The mode with low frequencies have been found to usually be associated with the major conformational changes observed in proteins [81].

However, this method is limited to the study of the protein dynamics in a single potential well, which does not make it a good technique to use to study conformational change. Also, much criticism has been directed at the use of the approximation beyond its theoretical capability [81].

Simulated annealing [82], a method which improves on conventional methods in obtaining low-energy configurations by initially carrying out MC steps or MD at high temperatures, which allows the simulation to overcome energy barriers. The temperature is then reduced until it reaches 0 K.

Principal component analysis (PCA) [83] is a technique which reduces the complexity of a dataset by the reduction of multidimensional data to fewer dimensions. The method can be used to extract the large significant motions of a protein, with the first principal component corresponding to the greatest variance, filtering out the smaller, insignificant motions, thus facilitating the study of long time dynamics. This method has been used as a method of analysis for HIV-1 enzymes in this thesis and has been described in greater detail in chapter 2.

The CONCOORD method [84] (from CONstraints to COORDinates) is used for the generation of conformations based on geometric restrictions from an initial structure. There are two stages to the method, the first being to identify all the interatomic interactions in the given structure and classify them according to their strengths. The second stage utilises this information to produce feasible conformations based on the constraint information using an iterative procedure.

3.2.2 Enhanced Sampling Methods

Simulations carried out using conventional MD and MC methods can become trapped in local energy minima, separated by large energy barriers, resulting in these simulations failing to sample all the available conformations, causing bias and inaccuracy in the calculation of average properties of the simulations.

A large number of sampling methods now exist which overcome the shortcomings of the conventional methods. Methods of enhanced conformational sampling exist which can enhance the sampling along a particular reaction coordinate, steering the simulation in a desired direction, whilst other methods focus on the escape from energy minima without the use of any final assumptions on the final state of the system. Several reviews [85] [86] [87] [88] [57] have been produced over the last few years, and a summary of some of the methods are included here.

3.2.2.1 Methods which Explore a Desired Pathway.

The majority of these methods employ the use of forces to direct the simulation in a desired direction.

The use of activated molecular dynamics [89] is useful where active processes occur in the system, which is usually the case for biological systems. Active processes are those which require energy to bring about a change e.g. where a conformational change occurs on binding of a substrate and a large energy barrier separates the initial and final states. A number of simulations are initially carried out, where the trajectories are forced along a particular pathway by the use of imposed constraints on the coordinates, in order to locate the energy barrier. An example where this method has been successfully applied is, in the study of the HIV-1 PR enzyme, where biologically relevant conformations of the flap structures of the enzyme, which control access to the active site cavity were observed [90].

Steered molecular dynamics (SMD) [91] [92] imposes restraints e.g. a harmonic potential on particular atoms, which directs the simulation along particular degrees of freedom, steering them on a prescribed path which accelerates processes that would otherwise be too slow to capture in a simulation using conventional methods e.g. in the study

of the release of ADP from cAMP-dependent-protein-kinase A (PKA), where the long timescale of ADP release would be difficult to study using conventional MD [93]

Targeted molecular dynamics method (TMD) [94] establishes a distance constraint between two structures, the initial and final, such that the sum over the distances of all atoms between these two structure has to be reduced in every step of the simulation. A successful application of the TMD method is demonstrated in a study which used the method to aid the determination of a conformational transition pathway in the allosteric process of human glucokinase [95].

The nudged elastic band method (NEB) [96] is a technique for finding transition paths and the corresponding energy barriers between known initial and final states, e.g. the method has been used to determine the saddle points connecting stable geometries between the reactant conformation and the transition state conformation of cAMP-dependent protein kinases [97]. The method works by simultaneously optimising the configuration of a chain of images of the systems along the reaction pathway, which are connected by springs. These are then relaxed, converging towards the minimum energy path (MEP).

3.2.2.2 Methods which Explore all of Phase Space

These include a variety of methods, such as accelerated MD [98], conformational flooding [99] [100], hyperdynamics [101] [102], puddle-jumping [103], and potential smoothing [104].

Generally, these methods reduce the amount of time a simulation spends in energy minima, thus forcing additional sampling of the rest of phase space.

A simple method of achieving this is high-temperature molecular dynamics [105], where simulations are carried out at high temperatures, thus overcoming the energy-barriers. However, there are several criticisms of this method, one being that force field are not parameterised for these higher than physiological temperatures, structures often result in having a proportion of cis peptides, and even after minimisation of the high-energy structures, the relevant low-energy structures are often not obtained.

Multiple time-step MD is based on integration schemes which allow for timesteps of different lengths depending to how rapidly a given type of interaction evolves in time.

Several algorithms exist, one of the more successful being that of Tuckerman, Martyna and Berne [106], the reversible reference propagation algorithm (r-RESPA), where the forces are classified as either hard (bond stretching and angle bending), medium (dihedral angle and non-bonded interactions) and soft (remaining non-bonded terms). The algorithm has been tested by Watanabe *et al.* [107] on the bovine pancreatic trypsin inhibitor protein, in addition to smaller peptides, demonstrating a reduction in computer time and agreement with MD simulations.

Local-enhanced sampling (LES) [108][109] [110] increases the conformational sampling of an area of interest within a system. This method is used where the system comprises of a small sub-system (region of interest) and a larger sub-system which interacts with the smaller one. This sub-system of interest is replaced with multiple non-interacting replicas, which are free to move apart and explore different regions of conformational space. The remainder of the system experiences the mean field of the replicated components. This method has successfully been used to increase the conformational sampling in the study of the motion of the catalytic loop of the core domain of HIV-1 integrase [111], with the loop as the sub-system, interacting with the rest of the enzyme.

Self-guided molecular dynamics (SGMD) [112] employs a force to guide the simulation. This guiding force is a constantly adjusted time average of the force of the current simulation, leading to improved search efficiency by aiding the simulation over energy barriers e.g. the method has successfully been used in the revelation of the β-hairpin folding mechanism of a nine residue peptide in explicit solvent [113]. However, the force guiding the simulation may not be guiding towards the large conformation changes of interest which are the slow motions of low-frequency since the forces of the high-frequency motions occur with a far greater magnitude

Modifications to this method include the enhanced hybrid Monte Carlo method (MEHMC). This method overcomes this problem by using average momentum instead of average force to bias the initial momentum within a hybrid MD/MC procedure, which is believed to give accurate canonical averages [114].

Conformational flooding [99] [100] incorporates the use of PCA. The initial protein system is destabilised through the addition of an adverse potential, and the method then

predicts the significant conformational events which take place using PCA. This information is incorporated in the application of a gaussian potential to force the simulation along the essential degrees of freedom.

Hyperdynamics [101] [102] and accelerated molecular dynamics [98] advance simulations by decreasing the amount of time which is spent in energy minima. The method alters the potential energy surface through the addition of a bias potential to the original potential, causing the surfaces near the minima to be raised, whilst leaving those near energy barriers unaffected.

Methods such as DFMD [115] selectively enhance or suppress motions on the basis of frequency, thus enabling the low-frequency motions, normally associated with large conformational change to be amplified. A time reversible version has also been developed, reversible digitally filtered molecular dynamics (RDFMD) [4]. The RDFMD technique, developed within the group has been used for the study of several protein systems [4] [5] [6], including the HIV-1 PR and HIV-1 IN enzymes studied in this thesis. The method is explained in more detail in chapter 4.

Another method of increasing conformational sampling is by the use of a coarse-grained model. Models of this type represent a system with a reduced number of degrees of freedom compared with the corresponding all-atom system. Owing to this reduction in the degrees of freedom and elimination of fine interaction details, the simulation of a coarse-grained system is less computationally expensive and increases the timescales that can be achieved. The flaps of the HIV-1 PR enzyme cover the active site and control access to the active site cavity. Experimental techniques suggest these flaps to undergo significant conformational change in order to allow this access, motions which are hard to capture using conventional MD. However, a coarse-grained approach has seen the flaps move from a semi-open to open conformation [116].

Multi-body dynamics (MBOND) [117] combines rigid body dynamics with multiple time steps. A system comprises of bodies of atoms, and other atoms which are not part of the bodies. This way, the number of variables of the system decreases, thus increasing the efficiency of the simulation. Since the highest frequency harmonic motions are removed whilst retaining the low-frequency anharmonic motions, relatively large timesteps are

able to be used. This method has been demonstrated to reproduce global essential dynamics properties of protein and nucleic acid systems [117].

The majority of the methods described so far (with the exception of coarse-graining) use a bias to influence the scaling of energy barriers in order to obtain improved conformational sampling. In most cases, these methods may not sample an accurate distribution of the conformations, resulting in inaccuracies in the calculation of average properties i.e. they are non-equilibrium.

Generalised Ensembles

The use of generalised-ensembles allows a free random walk in potential energy space, overcoming the energy barriers, thus sampling the full-range of conformations that exist [118]. In simulations in the canonical ensemble, the probability of energy barrier crossing (energy barrier height, ΔE) is proportional to $e^{-\Delta E/k_B T}$, where k_B is the Boltzmann constant, and T is temperature. Thus, at low temperatures, simulations can become stuck in energy minima, therefore limiting the amount of conformational phase space that can be explored.

When using these techniques, simulations are performed in an artificial ensemble where convergence to the equilibrium distribution is fast. Thermodynamic quantities are then calculated through re-weighting to the canonical ensemble [119]. These types of methods are characterised by the condition that the MD or MC simulation leads to a uniform non-canonical distribution of some pre-chosen physical property e.g. in the multi-canonical simulations, a uniform distribution of energy is obtained [118]. This is carried out through the replacement of the canonical weights, which suppress barrier-crossing, with non-canonical weights which allow the crossing. Each state is reweighted by an artificial non-Boltzmann probability weight factor, which are not a *priori* known, except in the case of parallel tempering (PT) [3] (where the weights are simply a product of Boltzmann factors, and so are known). This stage is not a trivial procedure, especially for systems with a large number of degrees of freedom. Time-consuming iterations of short trial simulations must be carried out (using e.g. histogram reweighting technique [119]).

The most widely known varieties of generalised-ensembles are, the multicanonical algorithm (MUCA) [120] [121]((also known as the random walk algorithm, entropic sampling,

adaptive umbrella sampling of the potential energy and density of states Monte Carlo)), simulated tempering (ST) [122] and replica-exchange methods (REM) [123] [124].

While a simulation in the multicanonical ensemble performs a random walk in potential energy space, in simulated tempering a random walk in temperature space is performed. This random walk, in turn, induces a random walk in potential energy space.

In REM, the replicas of the system, which are identical, except for some chosen physical property (temperature in the case of parallel tempering [3]) are run simultaneously as independent non-interacting simulations using conventional MD or MC. Then, at periodic intervals, the replicas are attempted to swap with their neighbour using an MC test. All the replicas together are considered as a single system which satisfies equilibrium thermodynamics.

Disadvantages to this method are the number of replicas and the computational expense, as the number of replicas required significantly increases with the number of degrees of freedom in a system, if the acceptance rate for the replica exchange moves are to be accepted with an adequate probability.

Algorithms have been developed in an attempt to overcome the computational expense of the REM method and the problem of the weight determination in generalised ensemble methods. These newer algorithms are essentially a mix of the algorithms previously mentioned. Some are based on the idea of obtaining the weight factor through a short replica-exchange simulation followed by running a longer simulation in a MUCA (REMUCA [125] [126]) or ST (REST [127]) ensemble.

Other modified versions of the REM technique have also been developed e.g. multiplexed-replica exchange molecular dynamics (MREMD) [128]. REM is only really efficient on a homogenous computer cluster, where the processors are comparable, as the method relies on synchronisation of the machines to facilitate the replica swapping, and thus the slowest replica determines the speed of the overall simulation. Additionally, the methods only scales to tens of processors i.e. to one replica per processer. The MREMD method overcomes these problems by the use of multiple copies for each temperature level and exchanges between these are also attempted, therefore eliminating the requirement for synchronisation.

In this thesis, the generalised ensemble method, parallel tempering has been used to study the conformations of cyclic peptides, and the methodology is explained in more detail in chapter 4.

Chapter 4

Enhanced Sampling Techniques: Parallel Tempering and Reversible Digitally Filtered Molecular Dynamics

This chapter describes the two enhanced sampling methods employed in this thesis, reversible digitally molecular dynamics (RDFMD) and parallel tempering (PT). Other enhanced sampling methods are summarised in chapter 3.

4.1 Reversible Digitally Filtered Molecular Dynamics (RDFMD)

Reversible digitally filtered molecular dynamics (RDFMD)[4] [5] [6] samples conformational change through the use of digital filters. These digital filters are applied to the internal velocities within a simulation to selectively enhance or suppress vibrational motion of selected structural regions of the protein.

Since RDFMD works by amplifying the motions of specific regions, some prior knowledge of the system is helpful to assist in the selection of the residues to target.

Frequency Range

To enhance conformational change, the filters are designed to extract certain frequencies.

The frequencies of interest are obtained through MD simulation, where the motion is analysed using Fourier transform methods or the Hilbert-Huang transform [129][130][131]. The more useful frequencies which produce increased conformational sampling, have been shown to be the low-frequency motions of the system [4] [5]. Further investigation into suitable frequency ranges has been carried out in the work of Wiley *et al.* [6], where the RDFMD methodology has been applied to the YPGDV peptide and the dihydrofolate reductase (DHFR) protein. Analysis of the simulations was carried out using the Empirical Mode Decomposition (EMD) method [132]. EMD is a signal processing method which, for a complicated signal, can decompose it to a number of high and low frequency components, known as intrinsic mode functions (IMFs). Each IMF can describe the evolution of frequency present in the input signal, and they are broadly ordered with frequency, with the lowest frequencies described by the last IMFs present in the system. The Hilbert-Huang Transform (HHT) is then applied to the IMFs to construct a time-frequency representation of the original signal. The analysis of simulations showed the frequency range of 0-25 cm^{-1} to include the vibrations associated with conformational change itself. To promote conformational change, the study suggested targeting the 0-100 cm^{-1} frequency range, which additionally includes the dihedral motions (25-100 cm^{-1}), as this was found to be more efficient if the conformational change of interest, involved escape from local energy minima. By choosing this frequency range, large scale motions associated with timescales which can be in the order of milliseconds can be targeted, whilst discarding uninteresting high-frequency motions which occur on fs timescales. Further work by Wiley *et al.* has employed this frequency range in the study of the conformational dynamics the T4L lysosyme [7] and HIV-1 protease enzymes [133], which resulted in the successful enhancement of conformational change.

To target these low-frequency motions, a filter is used, a filter being a list of coefficients (c_i) which will weight the vector input $(\boldsymbol{x}_i)$, and cause a desired frequency response $(\boldsymbol{y})$, affecting the amplitude and phase of the input signal (equation 4.1). The filter is generated using the *fircls* function in MATLAB [134], and as figure 4.1 shows, the greater the number of coefficients (m) in the designed filter, the closer the intended response is

to the target response of one. A filter of this type causes a frequency response of one over the frequency range to be amplified, and zero elsewhere [4].

$$\mathbf{y} = \sum_{t=-m}^{m} c_i \mathbf{x}_i \tag{4.1}$$

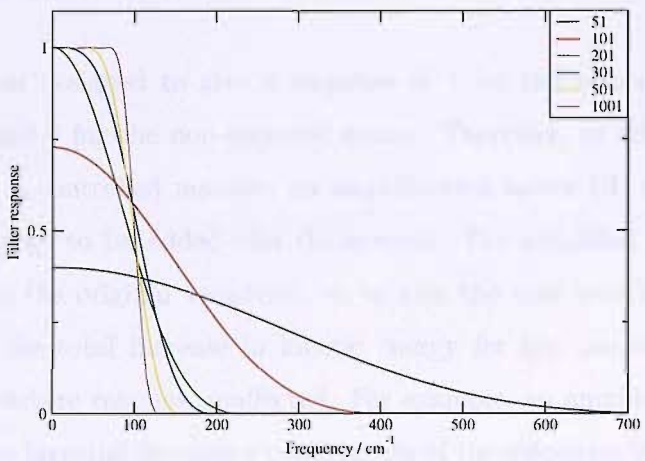


FIGURE 4.1: Filter response of 0-100 cm^{-1} filters using different numbers of coefficients, with a target filter response of 1.

In the application of the filter procedure, first, a buffer is filled with velocities, v, generated by running MD in the NVE ensemble. The number of steps in this stage must equal the number of coefficients, according to equation 4.2. Therefore, a compromise between the accuracy of the frequency response and the number of coefficients has to be made, as a large number of coefficients would result in a lengthy RDFMD simulation. The work carried out by Wiley *et al.* [6] found a 201 coefficient filter to be a good compromise. Following this stage, the rotational and translational components, $\mathbf{v}_{ext}$ of the velocities are removed to give the internal velocities, $\mathbf{v}_{int}$ (equation 4.3), as amplification of these motions ($\mathbf{v}_{ext}$) is of no use in enhancing conformational change. The next step is the

filter application (equation 4.2), where the filter is applied to the internal velocities, $\mathbf{v}_{int}$ to give filtered velocities, $\mathbf{v}_{filt}$.

$$\mathbf{v}_{filt} = \sum_{i=-m}^{m} c_i \mathbf{v}_{int,i} \tag{4.2}$$

$$\mathbf{v}_{int} = \mathbf{v} - \mathbf{v}_{ext} \tag{4.3}$$

Amplification of Internal Velocities

The filter has been designed to give a response of 1 for the velocities of the atoms to be amplified, and 0 for the non-targeted atoms. Therefore, to add energy into the selected region in a controlled manner, an amplification factor (A) is used to specify the amount of energy to be added into the system. The amplified velocities, $A\mathbf{v}_{filt}$, are added back to the original velocities, $\mathbf{v}_0$ to give the new velocities, $\mathbf{v}$' (equation 4.4). This gives the total increase in kinetic energy for the targeted regions, while the rest of the structure remains unaffected. For example, an amplification factor of 2 would multiply the targeted frequency components of the velocities by 2 and add them back to the original velocities, resulting in a 3-fold increase in the kinetic energy of the targeted frequencies, without changing the other frequencies. This amplification value is system dependent and values ranging from 1.4 to 2.4 have been used in this thesis. An amplification which is too large would cause the energy to be added too quickly into the system and risk the denaturation of the structure.

The new velocity set appropriate for the buffer centre, along with the coordinates are saved and the simulation is restarted from here.

$$\mathbf{v}' = \mathbf{v}_0 + A\mathbf{v}_{filt} \tag{4.4}$$

4.1.0.3 Stages of RDFMD Simulation

Figure 4.2 shows the overall stages of an RDFMD simulation. The first stage involves running a section of MD in the NVT or NPT ensemble. This section allows time for the

simulation to return to the selected simulation temperature. For the systems studied in this thesis, 4 ps has been found to be sufficient for this purpose, and is based on previous RDFMD work [6].

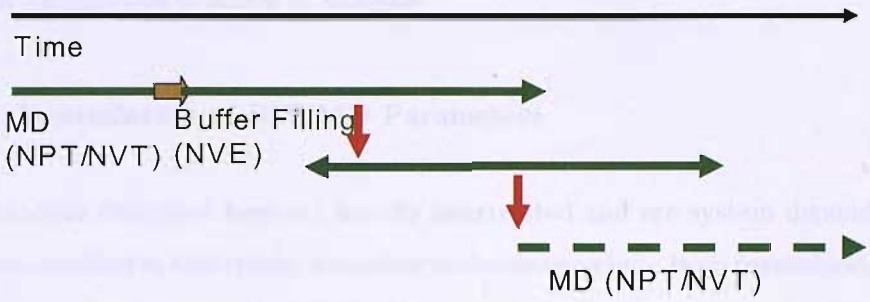


FIGURE 4.2: Application of filters in RDFMD. The dotted line represents the continuation of the process.

Next, the buffer is filled and the first filter application occurs, as described earlier. Figure 4.2 shows how further applications of filters are applied. Repeated applications of filters have been shown to be more effective than the use of a single filter application [4] as it allows the energy to build up gradually.

The second and subsequent filters can be positioned so they are slightly forward in time relative to the previous filter, known as the filter delay (measured in steps). The filter delay controls the kinetic energy dissipation and allows energy to be added to the system gradually, which permits the system to explore areas of the potential energy surface before the conformational change occurs. To enable user specification of this filter delay, the second and subsequent buffers are generated by running the MD simulations in both the backwards and forwards directions, the backwards simulation achieved by making the timestep negative. This results in a buffer with a first part filled with the buffer from the backward simulation, and the rest filled with the buffer from the forward simulation. The filter is then applied in the usual way. This application of filters is continued until the kinetic energy of the system rises above a specified limit, known as the temperature cap, or a specified number of filters has been applied, known as the filter cap.

Once the temperature cap is reached, another 4 ps of MD in the NPT or NVT ensemble is run. This process of MD and filter application stages is continued for a user-specified time. For the systems studied in this thesis, a total of 100 MD sections for each RDFMD

simulation has been carried out. Continuing the simulation for further increases the risk of denaturing the protein system.

Analysis of the RDFMD simulations has been carried out on the concatenated sections of MD (totalling 400 ps). The trajectories produced from the filter application stages have been disregarded in terms of analysis.

4.1.0.4 Interrelation of RDFMD Parameters

The parameters described here are heavily interrelated and are system dependent. For each system studied in this thesis, a number of simulations have been carried out, varying the parameters in a systematic way for their optimisation. The parameters which have been used in this thesis are similar to those chosen in previous work carried out by Wiley *et al.* [6] [7] [133] and are given in the main text.

4.2 Parallel Tempering (PT)

Parallel tempering (PT) [3] is a replica-exchange method which is one of the generalised ensembles that performs a random walk in energy space due to a random walk in temperature space (see chapter 3 for generalised ensemble methods).

The basis of PT is, that a given number of alternative starting conformations of the same system (replicas) are run as concurrent simulations at different temperatures. All of the replicas together are considered as a single system, which satisfies equilibrium thermodynamics.

When interpreting the results of PT simulations, the moves are unphysical, so information concerning dynamics cannot be obtained, and the method is seen only as a form of sampling.

Each replica substate (x) of the total system state (X) can be described by it's coordinates $q^{[i]}$ and momenta $p^{[i]}$ of N atoms in replica i at temperature T_m (equation 4.5)

$$x_m^{[i]} \equiv (q^{[i]}, p^{[i]})_m \tag{4.5}$$

In the canonical ensemble, each replica is weighted by the Boltzmann factor. The assumption is made that the probability of configuration x, $(P(x))$ of the m^{th} replica abides by the Boltzmann distribution at temperature m, as shown in equation 4.6.

$$P(x) = \frac{1}{Z_m} exp[-\beta_m E(x)] \tag{4.6}$$

where β_m is the inverse temperature, $(k_B T_m)^{-1}$, $E(x)$ is the Hamiltonian (represents the energy of the physical system), and Z_m is the partition function (encodes the statistical properties of a system in thermodynamic equilibrium).

Since the replicas of the total system are run independently, the weight factor for the state of total system X is a product of the Boltzmann factors for each replica, as shown in equation 4.7 for a M replica system.

$$P(X) = \prod_i^M P_i(X_i) \tag{4.7}$$

When attempting a swap between neighbouring replicas, the transition probability is represented as $W(X,\beta_m;X',\beta_n)$, with the probability of the reverse given as $W(X',\beta_m;X,\beta_n)$, where subscript m and n are the temperatures and X and X' are the configurations to be swapped.

This leads to equation 4.8, which imposes the detailed balance condition on the swap occurring (the forward and backward transitions occur with equal probability), allowing the process to converge towards an equilibrium distribution.

$$\frac{W(X, \beta_m; X', \beta_n)}{W(X', \beta_m; X, \beta_n)} = exp(-\Delta), \tag{4.8}$$

where,

$$\Delta = (\beta_m - \beta_n)\{E(X') - E(X)\} \tag{4.9}$$

To determine whether a swap of configurations between adjacent replicas is viable, the Metropolis-MC type criteria [29], as described in equation 4.10 is used.

$$W(X, \beta_m; X', \beta_n) = \begin{cases} 1 & \text{for } \Delta \leq 0, \\ \exp(-\Delta) & \text{for } \Delta > 0, \end{cases} \tag{4.10}$$

In MD, the momenta, $\underline{p}$, have to be dealt with after a swap has occurred. Sugita and Okamoto proposed a PT MD method where, after an exchange, the new momenta for the replica should be calculated as shown in equation 4.11 [135]

$$\begin{cases} p^{[i]'} \equiv \sqrt{\frac{T_n}{T_m}} p^{[i]} \\ p^{[j]'} \equiv \sqrt{\frac{T_m}{T_n}} p^{[j]} \end{cases} \tag{4.11}$$

where $p^{[i]}$ and $p^{[j]}$ are the old momenta and $p^{[i]'}$ and $p^{[j]'}$ are the new momenta for replicas i and j respectively. T_n and T_m are the temperatures of the replica after and before the swap respectively for replicas i and j. This allows the velocities to be rescaled by the square-root of the ratio of the two temperatures.

The PT algorithm results in enhanced sampling along the reaction coordinate. Exchange at low temperatures enhances sampling in the local minima (where energy dominates). Exchange at high temperatures assists in barrier crossing (where entropy dominates).

The set of temperatures for a PT simulation are not known *a priori* and several strategies are reported in the literature for their determination. [136][137][138][139].

The highest temperature must be high enough to ensure there is no trapping in local energy minima, and the number of replicas must be large enough for sufficient swapping between the adjacent replicas. There are several suggestions for the calculation of the optimal determination of replicas.

Energy histograms generated by simulations (e.g. at temperatures T_1 and T_2) separated by ΔT would see an overlap in the two sets of energy distributions if the temperature difference was sufficiently small, with an overlap approaching 1 for ΔT values of 0. This represents the likelihood of configurations observed in the simulation at T_1 being observed in simulation carried out at T_2. Therefore, in order for swaps to be accepted, the energy histograms must overlap to an extent. Sugita *et al.* and Kofke [135] [140] [141] proposed that, in order to ensure all replicas spent the same amount of time at each temperature, the probability profile should be uniform.

The simplest methods of achieving this assume a geometric distribution of temperatures to give a uniform distribution of replica temperatures [136]. However, this is only the case where the specific heat of the system is independent of temperature, which is not the case in protein systems, near transition temperatures (e.g. the heat capacity changes for conformationally flexible systems). At these temperatures, the specific heat rises and the exchange between replicas at the temperatures determined by this method prove to be inefficient [138].

An approach devised by Sanbonmatsu *et al.* [142] improved on the previous method in the consideration of the thermodynamic behaviour of the system. A number of short MD simulations are carried out and a polynomial fit obtained for the average energy as a function of time. The information is then used to iteratively solve the equation 4.12.

$$A_{target} = exp[\triangle\beta\triangle E], \tag{4.12}$$

where A_{target} is the target acceptance ratio, $\triangle E$ is the difference in the average energy of the systems adjacent in temperature (β being the inverse temperature, $(k_B T)^{-1}$).

Rathore *et al.* [138] found the method of Sanbonmatsu *et al.* to give consistently higher probabilities than the target, and improved on the method by additionally using the Gaussian distribution widths. From a number of short MD simulations, the average energy and the gaussian distribution width (σ(T)) can be estimated for the first temperature of the temperature range of interest. Therefore, the corresponding ΔE/σ can be obtained and used in the iterative calculation of the temperatures (T_i) as shown in equation 4.13.

$$\left.\frac{\Delta E}{\sigma_m}\right|_{T_i} = \left[\frac{\Delta E}{\sigma}\right]_{target}, \quad (4.13)$$

where $\sigma_m = [\sigma(\mathrm{T}_i)+\sigma(\mathrm{T}_i - 1]/2$ (average deviation of the energies of the two systems at adjacent temperatures (T_i and $\mathrm{T}_i - 1$)). $\Delta\mathrm{E} = \mathrm{E}(\mathrm{T}_i)\text{-}E(T_{i-1})$ (the difference between the energies in the two systems at adjacent temperatures (T_i and T_i-1)).

From this method, the optimum numbers of replicas used for PT simulations was determined when an acceptance probability of 20 % was required. This has also found to be the case in more recent work of Kone and Kofke [143], who calculate replica temperatures over sections of temperature intervals, making the assumption that C_v remains constant over these temperature intervals. Their focus is to maximise the mean squared displacement (σ^2) of the system with respect to the acceptance probability, as σ^2 is proportional to the number of swaps and $ln(\beta_j/\beta_i)^2$, where β is the inverse temperature of each of the replicas i and j. Although this assumption may not be true, they claim it to be reasonable and outweighs the iterative methods, in that this method is less computationally expensive.

The method of determination of replica number and distribution used in the PT studies in this thesis is similar to that of Rathore *et al.* and is described in chapter 5. This method was developed and used successfully in PT studies performed by Wiley [7].

Chapter 5

Application of Parallel Tempering to Small Cyclic Peptides

5.1 Aims

The free energy landscapes of proteins can be very rugged due to the comparatively close packing of the atoms in the native structure. MD simulations often become trapped in the energy minima, thus preventing the sampling of the whole of phase space. A number of enhanced sampling methods have been developed to overcome this problem, some of which are described in chapter 3. One such method is parallel tempering (PT), which has been applied in several protein-folding studies [144] [145] [146].

The aim of this chapter is to investigate the conformational sampling obtained with use of the PT methodology and compare the results with those obtained by experimental studies and conventional MD. As a test-case, the PT methodology has been applied to the study of the *trans-cis* isomerisation of proline-containing peptide bonds in small cyclic peptide systems.

The time-scale of the isomerism event is long (typically $10 - 10^3$ s), corresponding to a barrier height of 15-20 kcal/mol [147] [148] [149]. Therefore, this event proves difficult to study experimentally and virtually impossible to study by conventional MD and MC.

Two cyclic peptides have been selected, cPro-Ala-Pro-Ala-Ala (cPAPAA) and cPhe-Phe-Aib-Leu-Pro (cFFAiLP). cPAPAA has been designed for use as a rigid structural template, which is a common feature in strategies for protein design, and cFFAiLP contains the Pro-Phe-Phe sequence, which has been proposed to be responsible for the cytoprotective ability of antamanide and cyclolinopeptides. Antamanide is a decapeptide which exists as a natural component in the poisonous mushroom, *Amanetta phalloides*, acting as a antidote against the poisonous phallotoxins present [150], and cyclinopeptides are naturally occurring peptides which play a protective role in plant seeds [151].

Both these cyclic peptides have been studied previously, by experimental and theoretical means.

5.2 Background to Cyclic Peptides

Peptides are natural constituents of living systems, with an increasing number being found to play an important role in biologically relevant processes [152].

The development of pharmaceutically acceptable drugs is hindered by the drugs limited stability and absence of oral activity [153]. In solution, the majority of peptides lack fixed three-dimensional structures, adopting an ensemble of conformations without a clearly defined global minimum.

The cyclisation of peptides reduces the conformational freedom of inherently flexible linear molecules and often results in higher receptor binding affinities. This is a consequence of reduced unfavourable entropic effects and enhanced biological specificity, activity and metabolic stability against the action of proteolytic enzymes, resulting in super-potent analogues if locked in the active conformation [152]. Furthermore, peptides often display lower toxicity compared to small molecules as they do not accumulate in the body owing to their short half-lives [154].

The ability to predict conformations of cyclic peptides as inhibitor candidates prior to synthesis is important, and their properties have resulted in numerous potential therapeutic applications in the pharmaceutical industry [155].

Owing to the restraints in the main-chain backbone, low energy states of a cyclic peptide can be separated by high energy barriers. The presence of such barriers can strongly reduce the efficiency of conventional molecular dynamics and Monte Carlo simulations to search for possible low energy conformational states.

5.3 The *Cis* Peptide Bond

The peptide bond (figure 5.1) is shorter than a single bond and has a partial double bond character which results in the planar arrangement of the backbone atoms due to delocalisation of the carbonyl π electrons and the nitrogen lone pair [156]. The dihedral angle associated with the peptides is defined by four atoms of the backbone, C^{α}-C' - N-C^{α} (where C' is the carbonyl carbon), and there are two conformations which are energetically and sterically favourable, the *cis* conformation with an ω angle of 0^{o}, and the *trans* conformation, which has an ω angle of 180^{o} (figure 5.2).

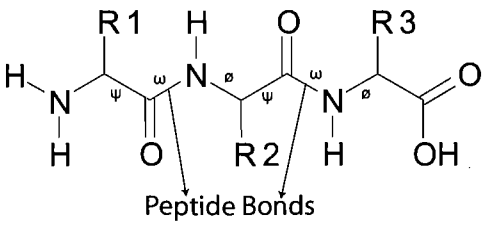


FIGURE 5.1: The peptide bond (two peptide bonds linking three amino acid residues are marked).

In proteins, generally only low populations of the *cis* conformation of the peptide bond exist due to the intrinsic non-equivalence of the rotational barrier heights between the *cis* and *trans* states. Greater steric hinderance exists between the functional groups attached to the α-carbon atoms in the *cis* conformation compared with the *trans*, which results in this higher energy. However, in peptide bonds containing proline residues, the *cis* conformation has an increased probability of existence due to the cyclic nature of the side chains (figure 5.2).

Proline is an imino acid, rather than an amino acid, yielded by the primary amine on the α-carbon of glutamate semi-aldehyde forming a Schiff base with the aldehyde which is then reduced. This residue has a higher propensity for the *cis* conformation, which is explained by the smaller energy difference of 2 kcal/mol between the *cis* and *trans* conformations, the *cis* conformation being the higher in energy [156].

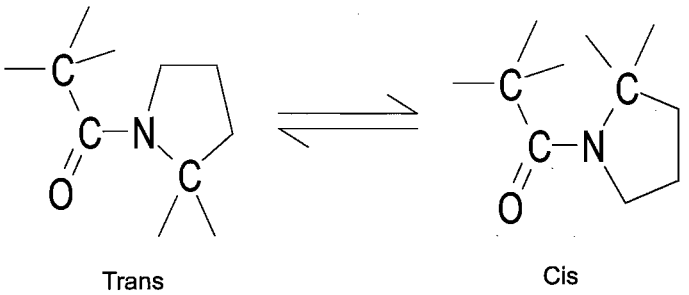


FIGURE 5.2: The isomerism of *trans* to *cis* conformations of the proline ω angle

There is functional evidence for the *cis-trans* equilibrium of the prolyl peptide bond through the existence of peptidyl-prolyl isomerases (PPIs), whose function is to catalyse the *cis-trans* isomerisation of peptide bonds N-terminal to proline residues. The purpose of these enzymes is thought to be crucial for the correct functioning of biological processes, such as protein folding of newly synthesised proteins [157]. These naturally occurring enzymes are able to catalyse the *cis-trans* isomerisation through lowering the activation energy of the process. There are two such mechanisms by which they operate, both of which result in changes that favour the single-bonded transition state, where the ω angle is 90^o. This is achieved by placing the peptide group in a hydrophobic environment or donating a hydrogen bond to the nitrogen atom of an X-Pro group [158].

5.4 Prior Experimental and Theoretical Studies

Nuclear Magnetic Resonance (NMR) experiments have detected the *cis-trans* isomerism rate in solution of prolyl peptide bonds in proteins [159] and cyclic peptides [160] and

determine it to often be the rate determining step in protein folding, with an energy barrier of 19 kcal/mol [161], resulting in the natural equilibration time for cyclic peptides to be in the timescale of 10 - 100 seconds [162].

NMR experiments have shown the *cis-trans* isomerism ratio to depend on the amino acid adjacent to the proline. A correlation has been found between the isomerism rate and the bulkiness of the side-chain of the residue preceding the proline residue; the isomerism rate decreases as the bulkiness of the side-chain increases [163]. Further NMR studies have shown that the *cis-trans* ratio is influenced by the nature of the succeeding amino acid, with positively charged side chains appearing to destabilise the *cis* conformation relative to *trans*, whereas aspartate, asparagine and glycine stabilise the *cis* form [164].

A study of the role of proline in defining β-turn conformations in cyclic hexa- and pentapeptides using NMR spectroscopy, including cPAPAA has been carried out by Müller *et al.* [165]. In addition to the experimental study, the authors carried out restrained MD simulations *in vacuo* and in the same solvent (DMSO) in which the NMR experiments were performed. This restrained MD used the interproton distance information from the forcefield, with a harmonic potential scaled by a force constant. Analysis of the MD simulation *in vacuo* was carried out on a total of 60 ps, starting from a conformation derived from a high temperature MD simulation in vacuum, in order to generate a conformation which agrees with experimental results. The simulation in DMSO solvent started from a conformation generated by the simulation *in vacuo*, and was carried out for 100 ps using restrained MD with force constants of 1000 kJ mol^{-1} nm^{-2} for the first 70 ps and 500 kJ mol^{-1} nm^{-2} for a further 30 ps. An additional 50 ps were carried out without restraints to check the stability of the conformation obtained from the restrained MD simulation.

This study concluded that the cPAPAA peptide adopts just a single conformation. Similar cyclic peptides, e.g. cPAAAA also investigated in this study were reported to adopt at least three conformations due to the less restricted alanine residue compared with the proline. The single conformation of cPAPAA was shown to possess two *cis* peptide bonds in the location of the proline residues, whilst the other peptide bonds were in the *trans* conformation.

Zanotti *et al.* studied the conformations of cFFAiLP using X-ray diffraction and NMR experiments [166]. Results derived from the solid-state structure obtained from the X-ray diffraction analysis describe the molecule as possessing a *cis* peptide bond between the Leu^4 and Pro^5 residues whilst all other peptide bonds remain in a *trans* conformation, with the existence of a γ-turn on the opposite side of the structure to the *cis* prolyl bond. NMR studies were carried out in three different solvents (chloroform, acetonitrile and methanol) and demonstrated this cyclic peptide to be conformationally heterogenous. Analysis of the peptide in acetonitrile produced the most defined spectra, demonstrating the presence of two almost equally populated conformers at 240 K, the two conformers differing in the *cis-trans* conformations of the prolyl-bond between residues Leu^4 and Pro^5.

As a consequence of the timescales required for the isomerism of the peptide bonds, conventional MD and Monte Carlo simulations are unable to sample the full range of conformations in the time available [167]. Studies using enhanced sampling methods have been carried out to study the *cis-trans* isomerism of peptide bonds; one such method being the analytical rebridging scheme [168]. This method incorporates three types of move; rebridging moves, semi-look-ahead (SLA) moves for side-chains, and swapping moves which use the parallel tempering algorithm. In this method, the bond lengths and angles are held at their equilibrium values and the molecule is considered as a set of rigid units. When a peptide rebridging move is applied, a local conformational change within the molecule results, leaving the rest of the molecule fixed. This move involves rotation of the angles of the residues concerned, breaking the connectivity of the molecule. Following this, all solutions which allow the units to be reinserted in a valid way are determined. For a resulting plausible geometry of the peptide to be obtained, only six units can be feasibly modified in one move. In adding back the side-chains, the SLA move is used, regrowing a randomly selected side chain back onto the backbone unit by unit. The swapping moves occur through the use of the parallel tempering technique at high temperatures, which aids equilibration by allowing the simulation to escape from local energy minima of the system.

The application of this method to the cyclic peptides, cPAPAA and cFFAiLP shows effective equilibration of the conformations, which agree well with experimental NMR

results [165] [166]. In both these studies, cPAPAA is shown to adopt a single conformation in DMSO solvent which involves the two prolyl peptide bonds in a *cis* conformation, with the remaining bonds in *trans* conformations. In the case of cFFAiLP, two interconverting conformers in acetonitrile were identified, with the *cis* and *trans* conformations of the prolyl peptide bond occurring in approximately equal proportions, whilst the non-proline containing bonds remained in *trans* conformations.

However, this method is limited in its application, as only two discrete states, $\omega = 0^o$ and $\omega = 180^o$ are assigned, so intermediate states are not taken into account. Additionally, solvent is only considered as an implicit model.

A further method which is able to enhance the conformational sampling of cyclic peptides is by the reversible scaling of dihedral angle barriers [169]. In this approach, a potential scaling protocol has been developed which lowers the energy barriers for main-chain dihedral angles in cyclic peptides at the start of the simulation. These are then smoothly rescaled throughout the duration of the simulation until the original potential of the forcefield is obtained. This method has been applied using both an implicit and explicit solvent model, displaying efficient equilibration of cyclic peptide systems, with results in good correlation with those obtained experimentally.

5.5 Simulation Details

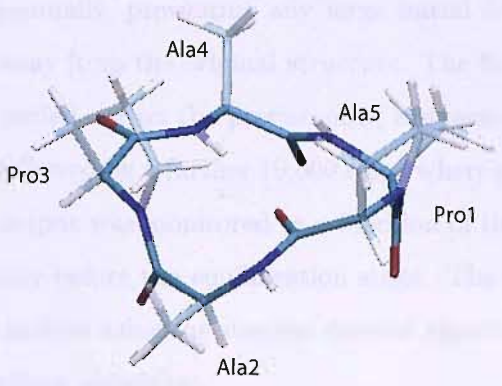
Both conventional MD and PT simulations have been used to study the conformations of cyclic peptides cPAPAA and cFFAiLP. All simulations have been carried out using the NAMD molecular dynamics code [170] with the AMBER99 forcefield [27]. SHAKE [35] was applied to all bonds containing hydrogen, using a tolerance of 10^{-8} Å.

5.5.1 Preparation of Cyclic Peptide Systems

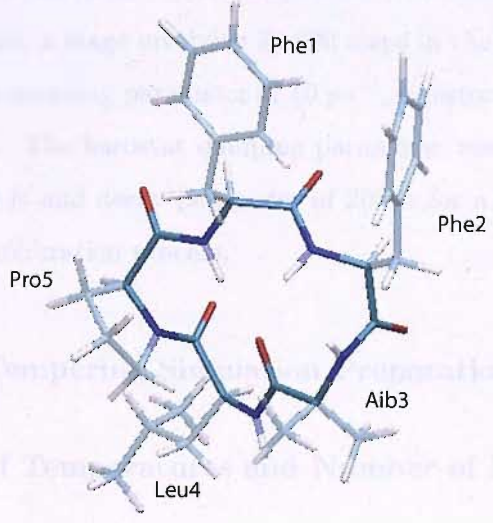
Construction

In construction of both the cPAPAA and cFFAiLP structures (figure 5.3), the appropriate amino acid units were joined together and polar hydrogens added using the WHATIF software [171]. Solvation of the peptide systems was carried out using the amber

utility, XLEAP [172], where, in each case, the solvent was added at a distance of 10 Å from the peptide. In the case of cPAPAA, the solvent used was dimethyl sulphoxide (DMSO), added as a pre-equilibrated box of 160 solvent molecules [173], and for cF-FAiLP, 283 molecules of acetonitrile (MeCN) solvent [174] was added. Additionally, XLEAP was used to add all other hydrogens required for the systems.



(a) cPAPAA



(b) cFFAiLP

FIGURE 5.3: The all-*trans* starting conformations of cPAPAA and cFFAiLP in liquorice representation. The backbone is shown in bold, with the residue sidechains faded out.

Minimisation and Equilibration

For both cyclic peptide systems, minimisation was carried out in stages, which allows the system to adjust gradually, preventing any large initial forces which may lead to artefactual movement away from the original structure. The first stage involved 10,000 steps of minimisation carried out on the protein only, restraining the atom positions of the rest of the system, followed by a further 10,000 steps where all atoms were unrestrained. The total energy output was monitored as a function of time to ensure the system was minimised adequately before the equilibration stage. The minimisation procedure used employs an initial section using the steepest descent algorithm followed by a period using the conjugate gradient algorithm.

A cutoff of 12 Å was used, with a switching distance of 9 Å. Electrostatic charges were calculated using Particle Mesh Ewald (PME) [175], using a interpolation order of 6 with a gridsize of 32 x 32 x 32 Å, similar dimensions to those of the boxsizes of x: 26.3, y: 28.7, z: 26.0 Å for cPAPAA and x:29.4, y: 35.2, z: 27.0 Å for cFFAiLP.

Equilibration of the systems was carried out gradually; the first stage was carried out in the NVT ensemble at 300 K for 50,000 steps with a thermostat damping parameter of 5 ps^{-1}. Following this, a stage involving 50,000 steps in the NPT ensemble was carried out with a Langevin damping parameter of 10 ps^{-1}, a piston period of 200 fs and decay parameter of 100 fs. The barostat damping parameter was reduced to 1 ps^{-1} with a piston period of 400 fs and decay parameter of 200 fs for a further 50,000 steps as the final stage in the equilibration process.

5.5.2 Parallel Tempering Simulation Preparation

Determination of Temperatures and Number of Replicas

The PT simulations were performed in the NVT ensemble using the NAMD [170] package and the CHARMM22 [15] forcefield.

As described in the parallel tempering methodology (see chapter 4), it is desirable to achieve the best possible sampling with the minimum amount of computational effort,

which is taken into account in the determination of the number and temperature range of the replicas.

The maximum temperature of the replicas must be sufficiently high as to ensure no replica becomes trapped in local energy minima and the number of replicas must be large enough to ensure swapping occurs between all adjacent replicas.

The maximum temperature replica was determined as a result of a number of short MD simulations carried out in the NVT ensemble at a range of temperatures. The temperature at which no barriers exist between the *cis* and *trans* isomers was identified as the approximate value for the maximum temperature replica. For both cPAPAA and cFFAiLP systems, this temperature was determined to be 1500 K.

Since the higher temperature simulations will have an effect on the stability of the integration algorithm, energy conservation has been tested across the temperature range of the replicas and the timestep altered accordingly. For both cyclic peptide systems, a timestep of 2 fs has been found to be appropriate for temperatures of 500 K and below, 1 fs for temperatures of 500 K to 1200 K, and 0.5 fs for temperatures of 1200 K, up to the value of maximum temperature replica.

The number and temperatures of the other replicas were determined so a swap between neighbouring replicas had a target acceptance probability of 20 %. The replica temperatures were calculated using the average and the variance of the energy values, obtained from the short simulations. This information is sufficient to calculate the energy distribution of a given temperature and thus able to predict the probability of accepting an exchange move at a particular temperature [176].

It has been proposed that the acceptance probability should be made uniform across all the replicas in an attempt to guarantee that each one spends an equal amount of simulation time at every temperature [135] [140] [141]. To determine the number and temperatures of replicas which give a uniform acceptance probability, an iterative scheme has been used, which uses the resultant means and standard deviations taken from the short MD simulations. In this case, the temperature of interest is 300.0 K, so the mean and standard deviation values obtained from the short MD simulation carried out at 300.0 K have been used to initiate the calculation.

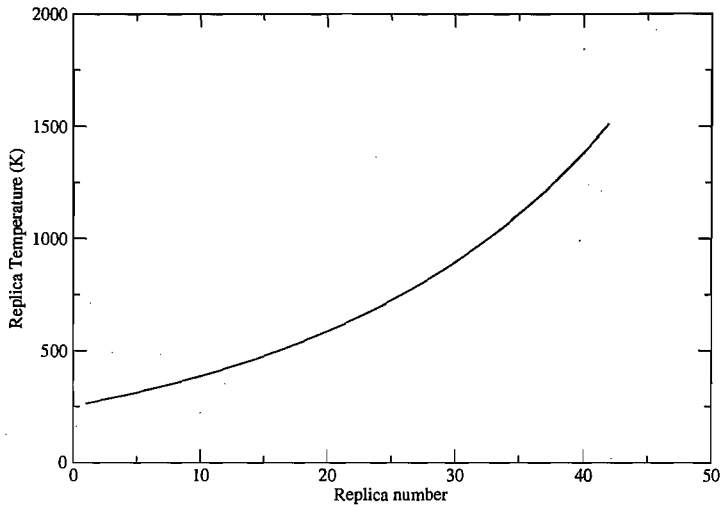
As previously mentioned, energy is not sufficiently conserved at the higher temperatures, and the timestep has been reduced (at 500 and 1200 K to 1 fs and 0.5 fs respectively). Reduction of the timestep has an effect on the variance of the energy distribution (i.e. results in a smaller energy histogram width). The variance of the energy using a smaller timestep is smaller than using a larger timestep at the same temperature, i.e. energy is better conserved using a smaller timestep. Therefore, the calculation used to determine the replica temperatures has been carried out over the temperature intervals where the timestep remains constant. Additionally, in order to calculate the replicas required below 300.0 K, the calculation has been carried out backwards, starting from 300.0 K.

This calculation of the replica temperatures for cPAPAA resulted in 42 replicas (highest temperature: 1524.5 K, lowest: 264.0 K) and for cFFAiLP, 41 replicas (highest temperature: 1524.5 K, lowest temperature: 264.0 K). Figure 5.4 shows the resulting temperature distribution of the replicas. As the replica temperature increases, their energy histograms become broader, allowing increasingly wider temperature intervals between the replicas.

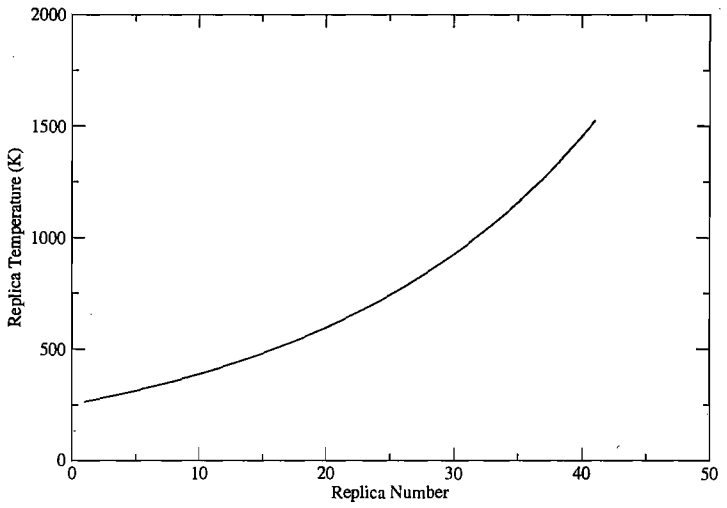
5.5.3 Production Simulation Parameters

The MD simulations presented in this chapter were performed in the NVT ensemble using the NAMD [170] package and the CHARMM22 [15] forcefield.

Production simulations using PT and conventional MD have been carried out in the NVT ensemble using box dimensions of 26.3, 28.7 and 26.0 Å for cPAPAA and 29.4, 35.2 and 27.0 Å for cFFAiLP. PME Electrostatic charges were calculated using Particle Mesh Ewald (PME) [175], using a interpolation order of 6 with a gridsize of 32 x 32 x 32 Å, similar values to the boxsizes. A cutoff of 12 Å and a switching distance of 9 Å, with a Langevin damping parameter of 5 ps^{-1} has been employed. Conventional MD simulations have been carried out at a range of temperatures which span the temperature range of the replicas in the parallel tempering simulation (300.0, 611.8, 933.9 and 1511.7 K for cPAPAA, and 300.0, 501.9, 846.5 and 1524.5 K for cFFAiLP). MD simulations



(a) cPAPAA



(b) cFFAiLP

FIGURE 5.4: Temperature distribution of replicas for cPAPAA and cFFAiLP PT simulations

at 300.0 K were performed for 20 ns, whereas simulations at other temperatures were carried out for 10 ns.

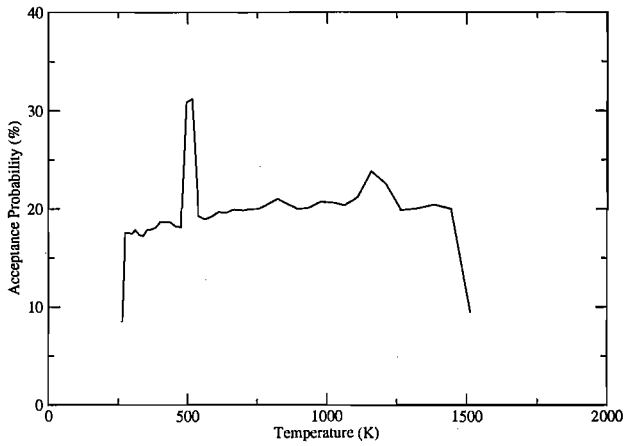
5.6 Results

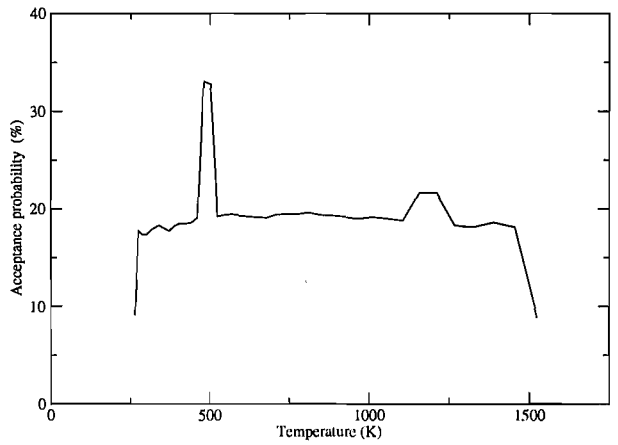
5.6.1 PT: Determination of Replica Temperature and Number

The PT simulation of cPAPAA in 160 DMSO molecules ran for a total of 10 ns for each of the 42 replicas, totalling 0.42 μs of total simulation time. For cFFAiLP in 283 MeCN molecules, 25 ns for each replica was completed, generating a total of 1.025 μs of total simulation time for all the replicas.

The average acceptance probability for each of the replicas exchanging with their immediate neighbours has been calculated for the duration of the simulation (figure 5.5). The first and last replicas display acceptance probabilities which are half the value of the target probability as they are only tested half the amount of time due to having only one neighbour with which to swap.

The acceptance probability of exchange for both cyclic peptides is shown to deviate above the anticipated 20 % at temperatures where the timestep has been decreased in order to ensure energy conservation in the integration algorithm. The probable cause of the non-uniform probability profiles is the method used in calculating the replica temperatures. The results suggest that where the timestep has changed (where the calculation is restarted), the temperature of the replicas in this region have been assigned inaccurately. The increase in the acceptance probability indicates that the temperature intervals are too close in these regions.

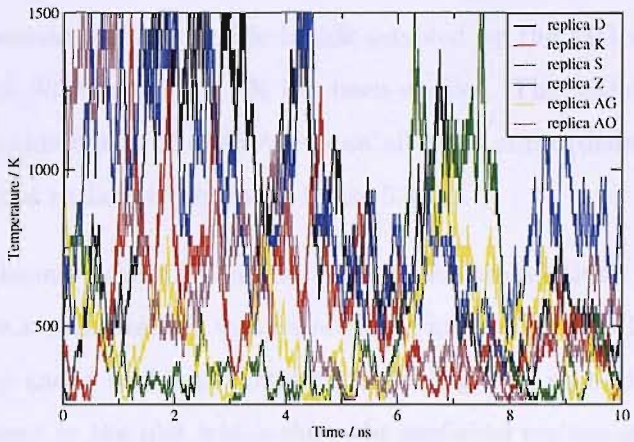




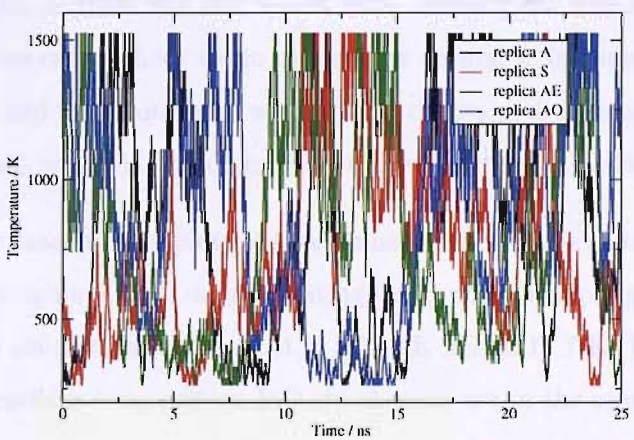
(b) cFFAiLP

FIGURE 5.5: Acceptance probability profile of replica swapping averaged over the length of PT simulations of cPAPAA and cFFAiLP.

These acceptance probability profiles are not ideal, since they may result in the replicas spending an unequal amount of time at each of the temperatures. However, the deviation is above the target 20 % acceptance probability (and not below) and mobility across the temperature range is demonstrated by the selection of evenly-spaced replicas reaching both top and bottom temperatures for the PT simulations of both cyclic peptide systems (figure 5.6).



(a) cPAPAA



(b) cFFAiLP

FIGURE 5.6: Mobility of evenly-spaced replicas over the length of the PT simulations of [a]: cPAPAA (10 ns) and [b]: cFFAiLP (25 ns)

5.6.2 Conformational Sampling

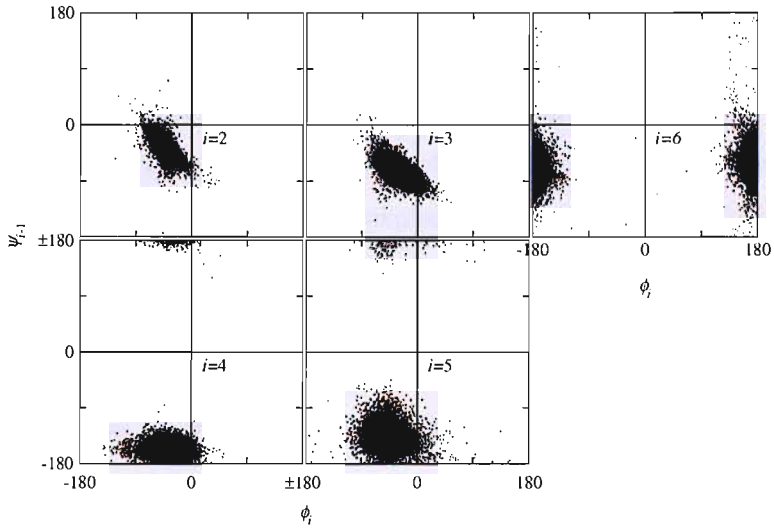
5.6.2.1 cPAPAA

MD Simulations

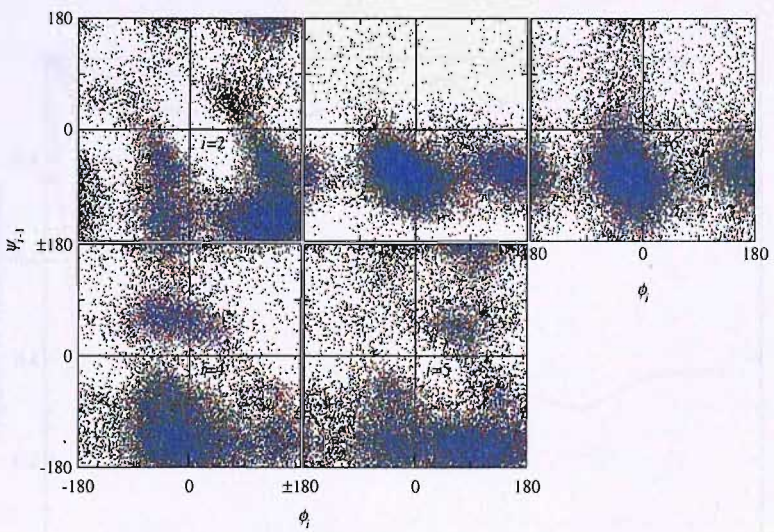
The *trans-cis* isomerism of the peptide bonds sampled by the MD simulations carried out at 300.0, 611.8, 933.9 and 1511.7 K has been studied. The MD simulations started with all of the peptide bonds of cPAPAA in an all-*trans* state (defined in terms of the ω angle, as described earlier) as shown in figure 5.3 (a).

In polypeptides, the main chain N-Cα and C-Cα bonds are relatively free to rotate, and these rotations are represented by torsion angles ψ and ϕ (figure 5.1). Ramachandran plots display the ψ and ϕ backbone conformational angles of each of the residues. The core regions are areas in the plot which show the preferred regions of sampling for ψ-ϕ pairs for the residues. The entire area of the ramachandran plots cannot be sampled as some conformations are disallowed, where the atoms of the polypeptide chain come closer than the sum of their van der Waals radii, causing the regions to be sterically disallowed. In these cases where steric hinderance occurs, it is between the side chain methylene groups and the main chain atoms. This occurs with all amino acids, with the exception of glycine, which is unique in that it does not possess a sidechain.

Figure 5.7 shows ramachandran plots of (ψ-ϕ values of each of the residues (residue numbering as shown in figure 5.3(a)) demonstrating the ψ and ϕ sampling over the duration of each of the MD simulations carried out at 300.0 K and 1511.7 K. The ramachandran plots of the intermediate temperature MD simulations are in the appendix.



(a) 300.0 K



(b) 1511.7 K

FIGURE 5.7: Backbone dihedral space sampled at 300.0 K and 1511.7 K over the 20 ns of MD simulation. Sampling measured in blocks of 5 ns, black 0 - 5 ns, red: 5 - 10 ns, green: 10 - 15 ns and blue: 15 - 20 ns. Values of i correspond to residues labelled in figure 5.3(a).

The ramachandran plots demonstrate that as the temperature increases, the defined regions of sampling seen at 300.0 K become more diffuse, sampling a larger range of ψ-ϕ values. At the higher temperatures, each of the ψ-ϕ torsion angle plots display sampling of additional conformations, further to those observed at 300.0 K. Additionally, sampling between the denser clusters is evident, indicating that the higher temperature MD simulations are able to overcome the energy barriers separating these conformations, thus demonstrating that the simulations explore a greater amount of conformational phase space. This is also demonstrated in figure 5.8, where the proportion of *trans* peptide bonds is measured over the duration of the MD simulations at 300.0 and 1511.7 K. At 300.0 K, all of the peptide bonds of cPAPAA are shown to have remained in the all *trans* state, whereas at 1511.7 K, the two prolyl-containing peptide bonds (ω_2 and ω_5) display the ability to isomerise to the *cis* conformation. By the end of the 20 ns of MD simulation at 1511.7 K, these two peptide bonds are shown to converge with approximate populations of 32 % and 27 % for peptide bonds defined as ω_2 and ω_5 *trans* respectively.

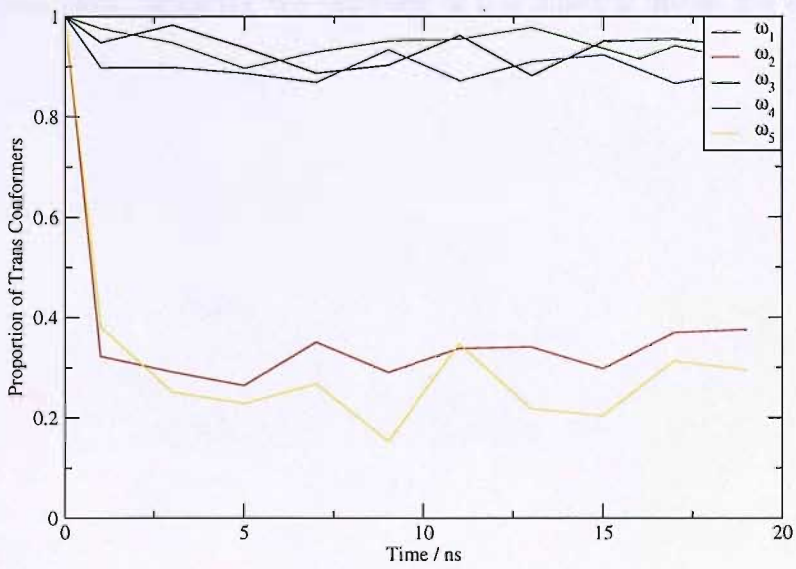


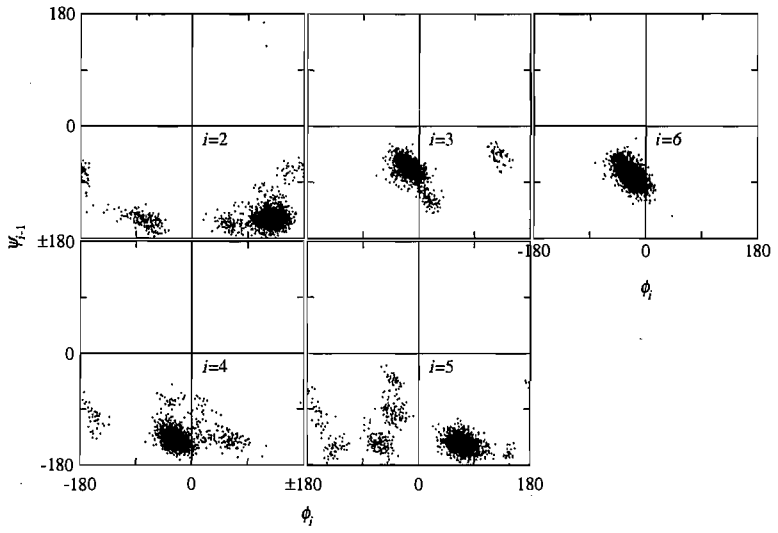
FIGURE 5.8: The proportion of *trans*-prolyl conformers sampled at 1511.7 K over the length of the 20 ns MD simulations. The average value has been calculated in 2 ns blocks over the duration of the simulation.

Furthermore, the ramachandran plot of the ψ/ϕ torsion angles observed at 300.0 K shows the conformational sampling does not increase throughout the simulation. The different coloured regions displayed on the plots represent the sampling in 5 ns sections over the length of the MD simulations. For 300.0 K, the sampling of each of the torsion angles at the start and end of the 20 ns simulation overlay on each other, sampling the same conformations. This indicates the simulation to have become trapped and unable to scale the energy barriers, forcing the simulation to sample only the local energy minima and therefore sample the same conformations throughout the simulation. This demonstrates that MD is not a suitable technique for the study of the *trans-cis* isomerisation of the prolyl peptide bonds, and so the method of PT has been used in an attempt to overcome this problem.

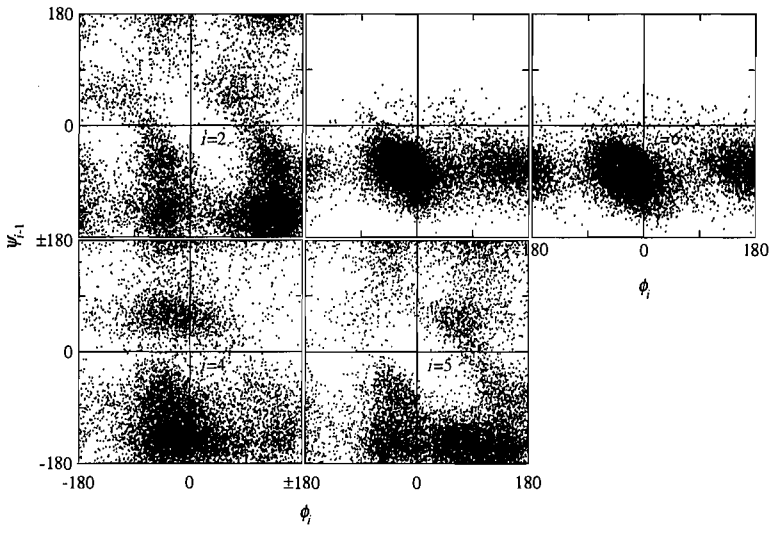
PT Simulations

As with the MD simulations, the *trans-cis* isomerism of peptide bonds in cPAPAA have been studied.

Ramachandran plots displaying the sampling of ψ-ϕ dihedral angles are shown figure 5.9.



(a) 300 K



(b) 1511.7 K

FIGURE 5.9: Backbone dihedral space sampled by replicas at 300.0 K and 1511.7 K over the 10 ns of PT simulation. Values of i correspond to residues labelled in figure 5.3 (a)

As observed with the MD simulations, the regions of sampling at lower temperatures are shown to be more defined and cover a smaller area of the ramachandran plot compared with sampling at 1511.7 K (top temperature replica), where the sampling is notably more diffuse. This indicates the use of the higher temperature replicas in overcoming the energy barriers, while the lower temperatures sample energy minima more extensively.

Comparison of the ramachandran plots displaying the sampling of replicas reaching 300.0 K in the PT simulation and the 300.0 K MD simulation demonstrates differences in the conformations sampled. In the ψ/ϕ plots corresponding to the torsion angles of ψ_1, ϕ_2 and ψ_5, ϕ_6, the PT simulation samples different regions of conformational space compared to the MD simulation at 300.0 K. However, the MD at 1511.7 K does observe these conformations, demonstrating the PT simulation to have overcome energy barriers and to sample minima not accessible by the MD at 300.0 K.

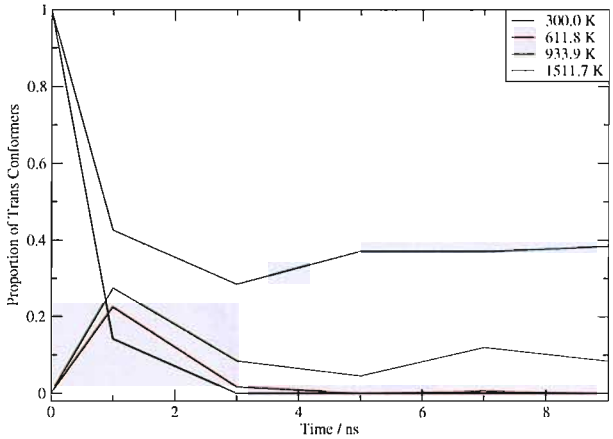
Also, in the ramachandran plots of the PT simulation at 300.0 K, several of the clusters of sampling are shown to linked, which is not observed in the plots generated by the MD simulation at 300.0 K. This sampling corresponds to sampling of conformations which cross the energy barriers between the energy minima.

Sampling by replicas reaching 1511.7 K in the PT simulation shows a notable increase in the area of the ramachandran plots sampled compared with the lower temperatures. Additionally, an increase in sampling is observed between the dense clouds of dots connecting the major conformations, demonstrating the heightened ability to overcome the energy barriers. The regions sampled are similar to those sampled by the MD simulation at 1511.7 K with sampling of torsion angles ϕ_3, ψ_2 and ϕ_6, ψ_5 (figure 5.9 (b)) showing considerably more white space compared with the sampling of the other dihedral angles. This white space represents the disallowed conformations due to steric hinderance between the sidechain groups and main-chain atoms of the corresponding residues.

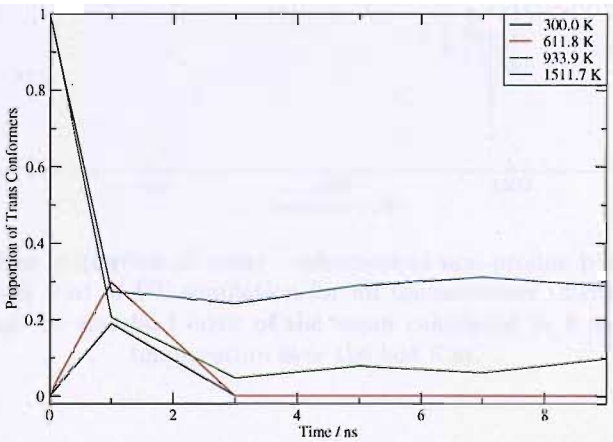
Figure 5.10 shows the average proportions of *trans* prolyl peptide bonds, ω_2 and ω_5 *trans* over the 10 ns of PT simulation at a range of temperatures. The simulations at the higher temperatures generate the *cis* conformations of the peptide bonds which then use PT swapping to drop down and populate the lower temperatures. Thus, at higher temperatures, the population of *cis* conformers is dependent on the rate of transition,

whereas the population at lower temperatures is reliant on the efficiency of the replica swapping. As figure 5.10 shows, at the higher temperatures, the simulation converges to increasingly higher proportions of *trans* peptide bonds conformations of the prolyl-containing peptide bonds. As entropy is known to dominate at higher temperatures (enthalpy dominates at lower temperatures), it suggests that the *trans* form of prolyl bonds of cPAPAA are entropically more favourable, whereas *cis* states of the prolyl-peptide bonds are more favourable at the lower temperatures. At 300.0 K, the simulation is shown to converge to 100 % of *cis* conformers for ω_2 and ω_5 prolyl peptide bonds, which is consistent with the NMR [165] and analytical rebridging study [159].

The convergence of the non-prolyl peptide bonds (figure 5.11) are harder to determine as the energy barrier to the *trans* to *cis* isomerism is far higher. No *cis* isomers of these bonds are seen at the lower temperatures (below $\sim$ 600 K). Above this temperature, isomerism of these bonds is observed, although significantly fewer transitions occur compared with the prolyl containing peptide bonds. Therefore, it is harder to judge whether the simulation, in terms of the non-proline containing peptide bonds has converged. NMR [165] and previous theoretical data [168] report only *trans* conformers of these peptide bonds at 300.0 K, which is consistent with the analysis of the PT simulations carried out in this chapter.



(a) ω_2



(b) ω_5

FIGURE 5.10: The proportion of *trans*-prolyl conformers reaching a range of temperatures over the length of the PT simulation. Values have been calculated as averages over 2 ns blocks of simulation time.

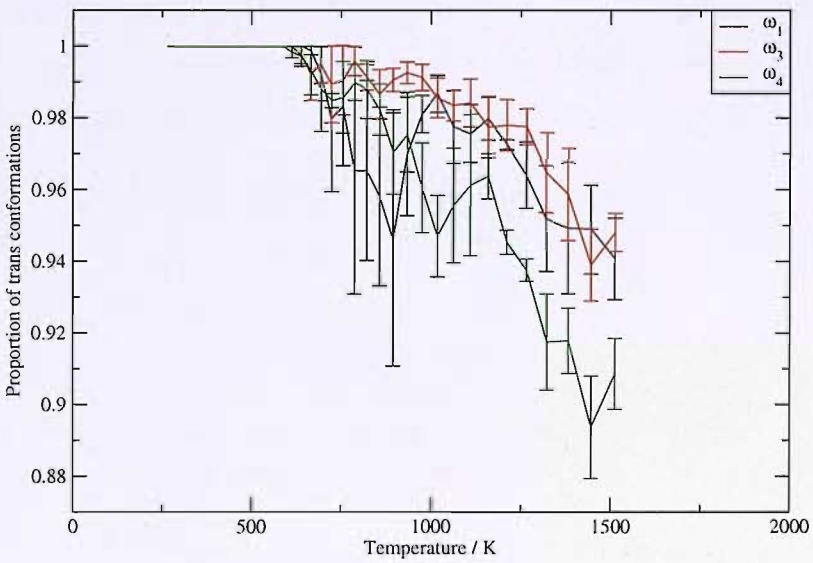
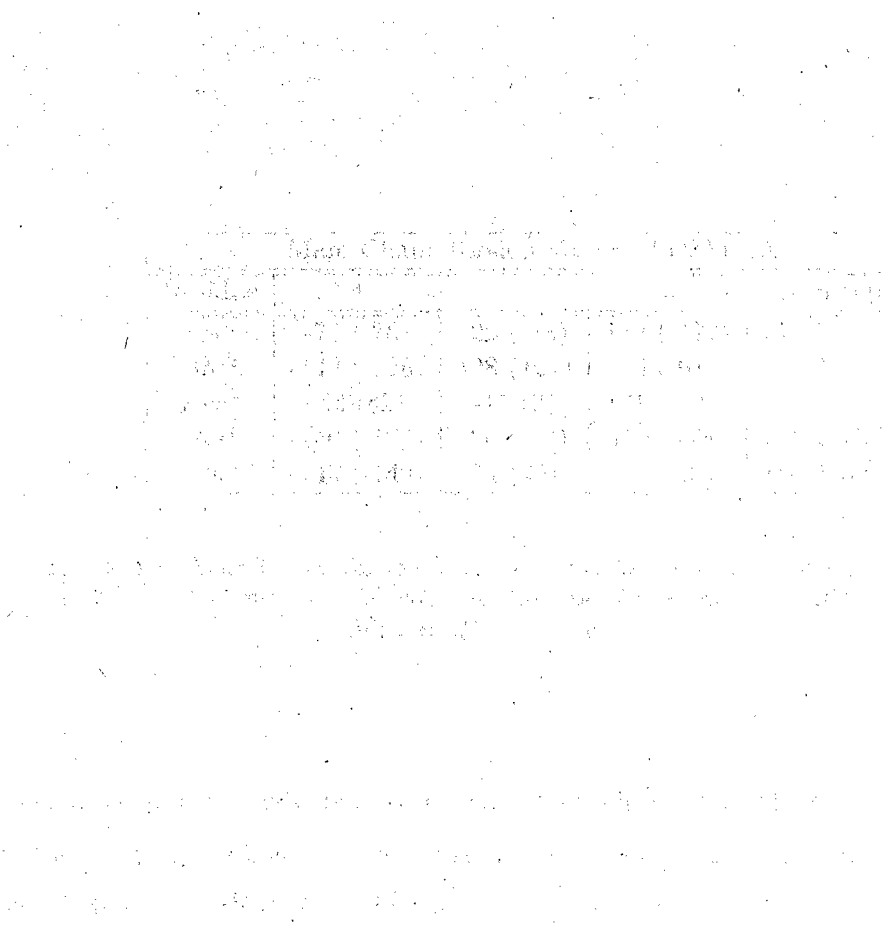


FIGURE 5.11: The proportion of *trans* conformers of non-proline peptide bonds averaged over the last 5 ns of PT simulation for all temperatures. Error bars represent the calculation of the standard error of the mean calculated in 1 ns blocks for each temperature over the last 5 ns.

The number of transitions throughout the 10 ns parallel tempering simulation shows the escalation in the rate of isomerism with an increase in temperature. This is expected as the higher temperatures will be able to overcome the energy barriers with greater ease compared with the lower temperatures. Additionally, the rate of isomerism is higher for the proline-containing peptide bonds owing to the relatively lower energy barrier to the event compared with the non-prolyl peptide bonds.



The main chain torsion angles have been reported in the literature, as shown in table 5.1. These values were obtained from a restrained MD simulation carried out in addition to NMR studies by Müller *et al* [165]. Values have also been calculated for the PT simulation. Average ψ and ϕ values have been calculated for the last 5 ns of the PT simulation at 300.0 K (experimental temperature) for each of the residues. For each of the residues numbered in the table, the peptide bonds of which the ω angles are measured, are between itself and the proceeding numbered residue. In the case of Ala^5, the ω angle between itself and Pro^1 is measured.

The ω angle values for the proline residues were shown to fluctuate between the *trans* states of approximately + and - 180°, therefore 180° was added to the negative values before averaging took place. Therefore, the sign of the value for these values is not taken into account, but does show the prolyl peptide bonds to be in the *trans* state.

Main Chain Torsion Values of cPAPAA				
Residue	ϕ	ψ	ω	Turn type
Pro^1	-78 (-76)	-20 (-28)	172 (-177)	$i+2$ βVIb
Ala^2	-148 (-126)	128 (127)	0 (-9)	$i+1$ βVIb
Pro^3	-62 (82)	-17 (2)	181 (-178)	
Ala^4	-140 (-123)	-26 (-56)	176 (178)	$i+2$ βVIb
Ala^5	-143 (-140)	74 (69)	-2 (-18)	$i+1$ βVIb

TABLE 5.1: Main Chain Torsion Values of cPAPAA. Average values calculated from last 5 ns of PT simulation at 300.0 K. Results of the restrained MD simulations of Müller *et al.* are shown in brackets

Comparison of the values obtained by the PT simulations and the literature values [165] for the main chain torsion angles are shown to be similar, both indicating the prolyl *cis* peptide bonds of cPAPAA are in the central part of two VIb β-type turns.

Overall, PT simulation has concluded results in agreement with prior experimental and theoretical studies. In solution, cPAPAA exists in a single conformation, with two *cis* prolyl peptide bonds, with other bonds remaining in the *trans* conformation (figure 5.13).

MD simulations at 300.0 K were not able to sample the isomerism of the prolyl-peptide bonds, and they remained in their initial all-*trans* state.

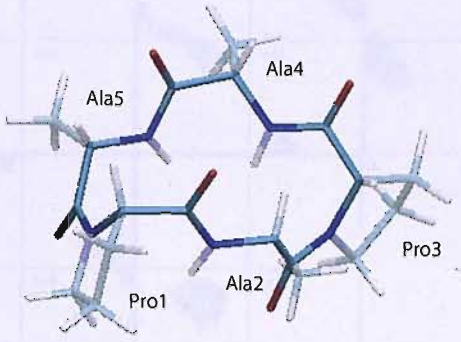


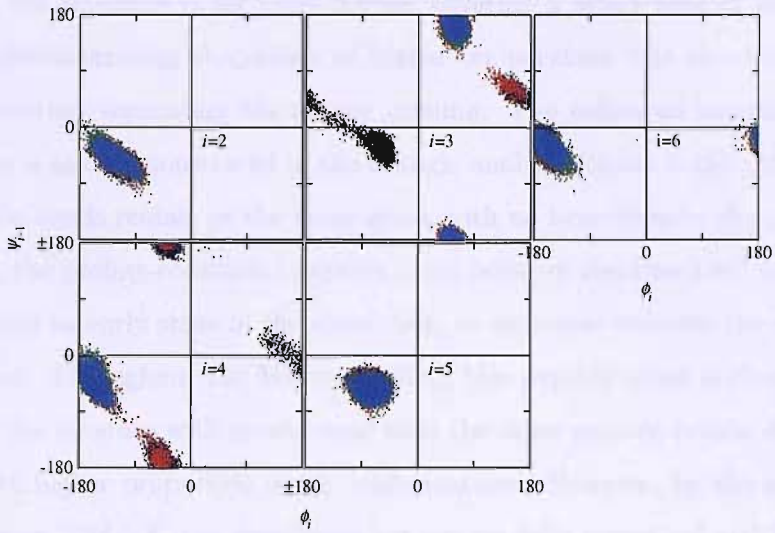
FIGURE 5.13: Single conformation of cPAPAA present in DMSO solvent. Prolyl peptide bonds are *cis* and non-prolyl bonds are *trans*.

5.6.2.2 cFFAiLP

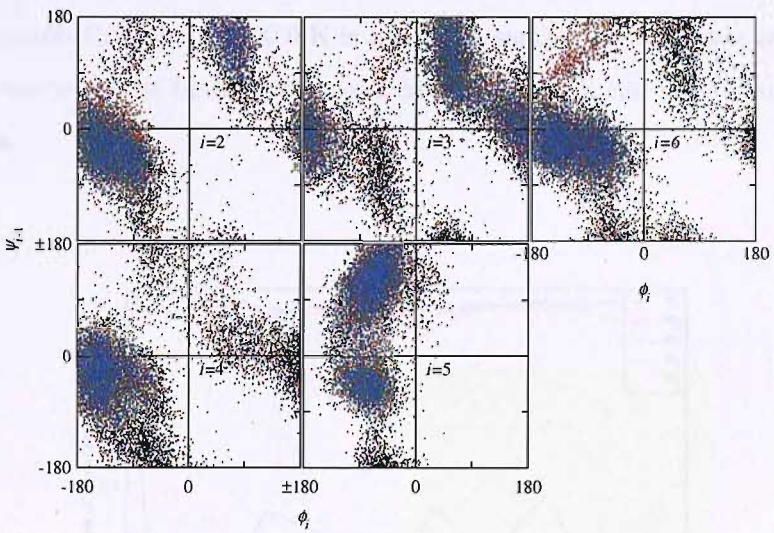
MD Simulations

The *trans-cis* isomerism of the peptide bonds sampled by the MD simulations carried out at 300.0, 501.9, 846.5 and 1524.5 K have been studied. The MD simulations started with all *trans* peptide bonds of cFFAiLP (figure 5.3 (b)).

Ramachandran plots display the sampling of the ψ-ϕ torsion angles over the duration of the 20 ns MD simulation at 300.0 K and 1524.5 K (top temperature replica of PT) (figure 5.14). Plots showing the sampling of intermediate temperatures (501.9 K and 846.5 K) are in the appendix.



(a) 300.0 K



(b) 1524.5 K

FIGURE 5.14: Backbone dihedral space sampled at 300.0 K and 1524.5 K over the 20 ns of MD simulation. Sampling measured in blocks of 5 ns, black 0 - 5 ns, red: 5 - 10 ns, green: 10 - 15 ns and blue: 15 - 20 ns. Values of i correspond to residues labelled in figure 5.3 (b).

As seen in the simulations of cPAPAA, sampling at 300.0 K for cFFAiLP results in defined clusters of ψ-ϕ angle pairs, whereas at 1524.5 K (temperature of top replica in PT simulation), the sampling is far more diffuse, covering a larger area of the ramachandran plots, demonstrating the ability of higher temperature MD simulations to scale the energy barriers separating the energy minima. The enhanced sampling at higher temperatures is also demonstrated in the ω angle analysis (figure 5.15). At 300.0 K, all of the peptide bonds remain in the *trans* state, with no isomerism to the *cis*. Whereas, at 1524.5 K, the proline-containing peptide bond between residues Leu^4 and Pro^5 (ω_4) are shown from an early stage of the simulation, to isomerise between the *cis* and *trans* conformations. Throughout the MD simulation, this peptide bond is clearly shown to isomerise to the *cis* state with greater ease than the other peptide bonds, demonstrated by its relative higher proportion of *cis* conformations. However, by the end of the 20 ns simulation at 1524.5 K, the simulation has not yet fully converged and large fluctuations in the proportions of the *trans* conformation of the prolyl-peptide bond (ω_4) are evident. The non-prolyl peptide bonds are shown to predominantly remain in the *trans* conformation throughout the whole 20 ns of MD simulation.

This demonstrates that MD at 300.0 K is unable to sample the *cis-trans* isomerism and high-energy barriers exist between the conformations, which can be overcome at higher temperatures.

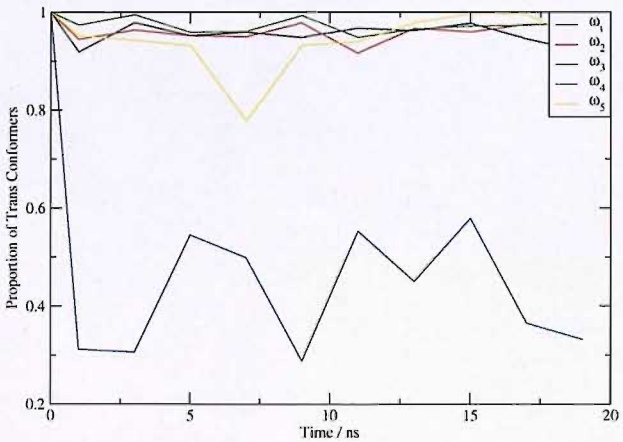
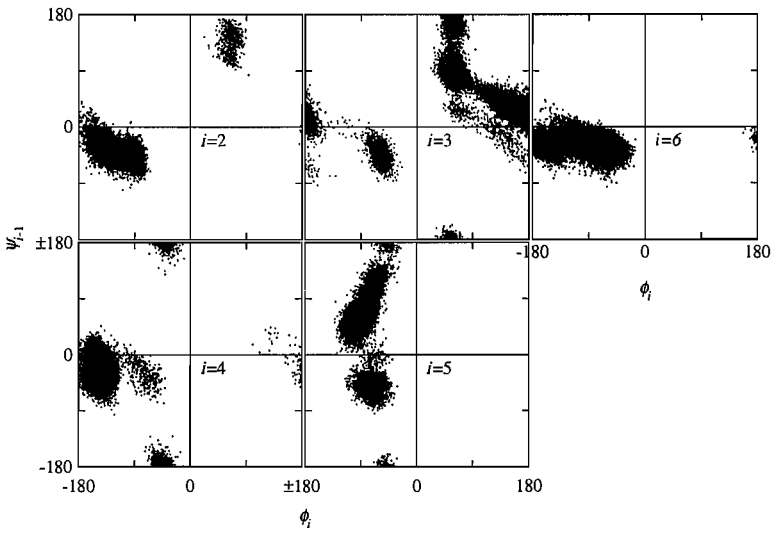


FIGURE 5.15: The proportion of *trans*-prolyl conformers reaching 1524.5 K over the length of the 20 ns MD simulations. The average value has been calculated in 2 ns blocks over the duration of the simulation.

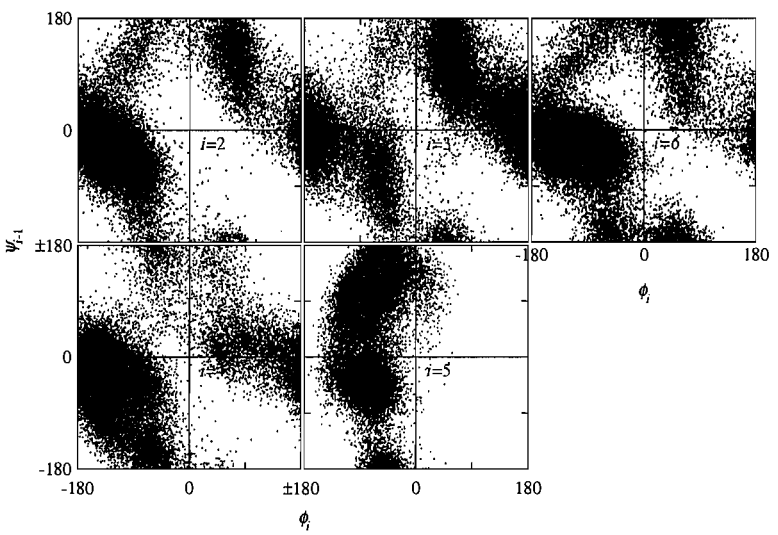
PT Simulations

A total PT simulation time of 1.025 μs was carried out for 41 replicas, equalling 25 ns of PT simulation for each replica.

Ramachandran plots demonstrate the ψ-ϕ torsion angles sampled at 300.0 K and 1524.5 K (top replica temperature) (figures 5.16) of the PT simulation. As observed in the results of the MD simulations, an increase in temperature results in an enhancement in the number of conformations sampled, indicated by the significantly larger area of the ramachandran plot sampled at 1524.5 K compared with 300.0 K. As noted in the conclusions of cPAPAA simulations, this demonstrates the ability of the higher temperature replicas to be able to scale energy barriers. These conformations are then swapped to lower temperatures in order for the new energy minima to be sampled more extensively, resulting in the denser clouds observed at 300.0 K. Dots between the clouds represent the pathway of conformational sampling which separates the major conformations observed in different energy minima.



(a) 300.0 K



(b) 1524.5 K

FIGURE 5.16: Backbone dihedral space sampled by replicas at [a]: 300.0 K and [b]: 1524.5 K over the 25 ns of PT simulation. Values of i correspond to residues labelled in figure 5.3 (b)

The ability of the PT simulation to overcome energy barriers enables the *trans-cis* isomerisation to be observed. Figure 5.17 shows the *trans* to *cis* isomerism of the prolyl (ω_4) peptide bond throughout the 25 ns of PT simulation for a range of replica temperatures. The plotted points are average values over 2.5 ns blocks. All but one (1524.5 K) of the starting structures initiate the PT simulation with a *trans* prolyl peptide bond (ω_4) conformation. Figure 5.17 shows at all temperatures, isomerism to the *cis* conformation occurs from the start of the simulation with convergence at the higher temperatures occurring more quickly than at lower temperatures. This is to be expected since the isomerism rate is faster at the higher temperatures, as shown in figure 5.18.

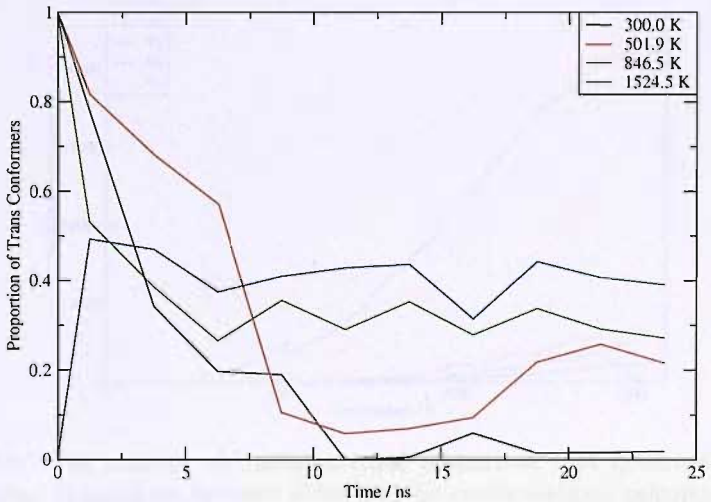


FIGURE 5.17: The proportion of *trans*-prolyl conformers reaching a range of temperatures over the length of the PT simulation. Values have been calculated as averages over 2.5 ns blocks of simulation time.

As the temperature increases, the relative proportion of *trans* conformers also increases, demonstrating the *trans* conformers of the prolyl peptide bonds to be entropically the more favoured state, as also noted in the case of cPAPAA. At 300.0 K, the proportion of *trans* conformers of prolyl peptide bond (ω_4) has converged to 0.050 ± 0.009, calculated from the average of the last 5 2.5 ns blocks. This value is not consistent with those obtained by the NMR [166] and analytical rebridging theoretical study [168], where the *cis* and *trans* conformations are observed in approximately equal proportions. The cause of this is not clear, a possible reason being a deficiency in the force field used.

The conformations of the of the non-prolyl peptide bonds have been monitored over the last 5 ns of PT simulation for each of the replica temperatures (figure 5.19). As previously mentioned, the non-prolyl peptide bonds in the starting conformations are in the *trans* conformation. At 300.0 K, the non-prolyl peptide bonds are shown to remain in the *trans* conformation, with figure 5.18 showing no isomerisation to the *cis* conformation at this temperature. At the higher temperatures, isomerism to the *cis* conformations is observed. The rate of *trans-cis* isomerism is significantly lower in these non-prolyl peptide bonds than that of the prolyl-containing bonds, signifying the higher energy barrier to isomerism in the absence of proline. As a consequence, it is harder to judge the confor

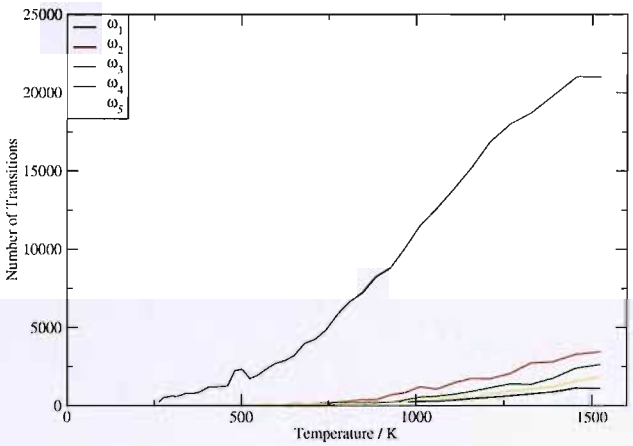


FIGURE 5.18: The number of conformational transitions that occurred over 25 ns PT simulation. Transitions between *trans* and *cis* conformations calculated from the dot product of the two planes (plane 1: atoms Cα-C'-N, plane 2: atoms C'-N-$C\alpha^2$). Dot product of 1 represents a *trans* conformation, dot product of 0 represents a *cis* conformation of the peptide bond.

Overall, the PT simulations here indicate the presence of a predominant conformation of cFFAiLP in MeCN at 300.0 K (figure 5.20 (a)), where the prolyl peptide bond is in a *cis* conformation, whilst the non-prolyl peptide bonds remain in a *trans* conformation. A second conformation of cFFAiLP is shown to exist at 300.0 K, although detected in small quantities, where all the peptide bonds are in an all-*trans* state (figure 5.20(b)). This result is not consistent with the analytical bridging study [159] or NMR results [166], which observe two major conformations, with almost equal proportions of the prolyl bond in *cis* and *trans* conformations. The reason for the discrepancy may be owing to the force field being unsuitable for this purpose, although this is not clear.

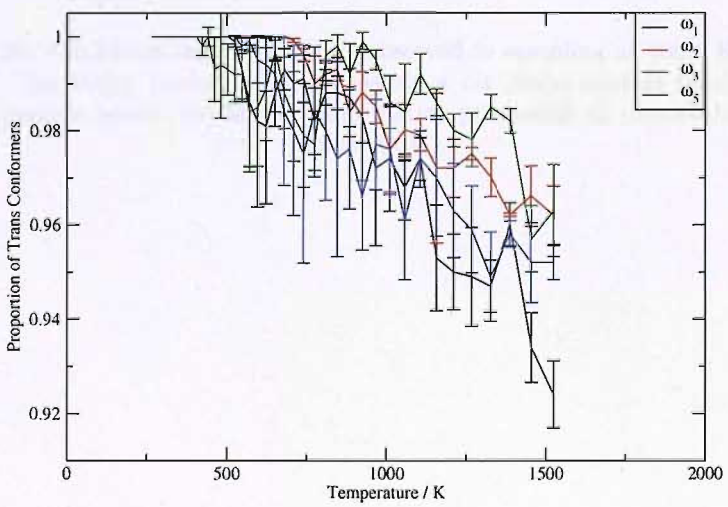
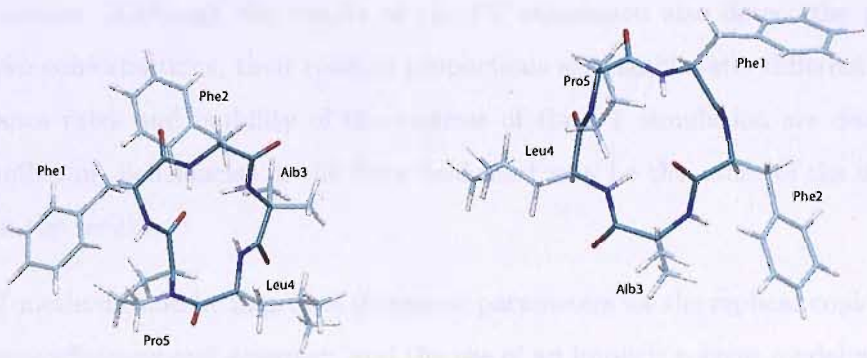


FIGURE 5.19: The proportion of *trans* ω non-prolyl peptide bonds averaged over the last 5 ns of PT simulation for all temperatures. Error bars represent the calculation of the standard error of the mean calculated in 1 ns blocks for each temperature over the last 5 ns.



(a) Major conformation of cFFAiLP (b) Minor conformation of cFFAiLP

FIGURE 5.20: Conformations of cFFAiLP observed in sampling at 300.0 K of the PT simulation. (a) Major conformation possessing a *cis* prolyl-peptide bond and trans non-prolyl peptide bonds. (b) Minor conformation possessing all peptide bonds in the *trans* state.

5.7 Conclusions

The results of the PT simulation for cPAPAA in DMSO solvent compare well with experimental [165] and a previous theoretical study [168] which employs the use of an enhanced sampling method. The cyclic peptide is shown to exist in a single conformation at 300.0 K, where the prolyl peptide bonds are in the *cis* conformation, with other bonds being in the *trans* conformation.

The results of the PT simulation of cFFAiLP in MeCN are not consistent with previous work. Prior experimental [166] and theoretical results [168] report the existence of two conformations, one where the prolyl bond is in the *cis* conformation, and one where it is in the *trans* conformation. All non-prolyl bonds are observed to be in the *trans* conformation. Although the results of the PT simulation also detect the presence of these two conformations, their relative proportions are significantly different. Since the acceptance rates and mobility of the replicas of the PT simulation are demonstrated to be sufficient, deficiencies in the force field used may be the cause of the discrepancy between the results.

The PT method could be improved if optimal parameters for the replicas could be chosen with more efficiency and accuracy, and the use of an implicit solvent model would make the method computationally less expensive. However, for the purposes of the studies carried out in this thesis, it was decided that the PT methodology was not a feasible method for the study of proteins in explicit solvent with the current resources available, and other methods were investigated (e.g. RDFMD).

Chapter 6

Human Immunodeficiency Virus (HIV): Background

Acquired immunodeficiency syndrome (AIDS) is caused by the human immunodeficiency virus type 1 and type 2 (HIV-1 and HIV-2), members of the family of retroviruses. The syndrome describes the symptoms and infections caused as a result of the specific damage to the immune system caused by the HIV virus. Two species of the virus exist; HIV-1 and HIV-2. The HIV-1 virus is the more virulent strain, which is more easily transmitted, whereas HIV-2 is less easily transmitted and is confined mostly to West Africa [177].

The HIV virus is transmitted through direct contact of a mucous membrane or the bloodstream with a bodily fluid. If the virus is left uninterrupted, it spreads throughout the body and results in the destruction of the body's immune system. Although a number of drugs exist which are able to slow down the progress of the virus, no cure currently exists.

6.1 Life-Cycle of HIV-1

Figure 6.1 outlines the life-cycle of the HIV virus. The HIV genome encodes for three enzymes which are required for both the infection and replication of the virus. These enzymes are HIV-1 protease (HIV-1 PR), integrase (HIV-1 IN) and reverse transcriptase

(HIV-1 RT). Both the reverse transcriptase and protease enzymes have been extensively studied, but the mechanism of integrase is less understood.

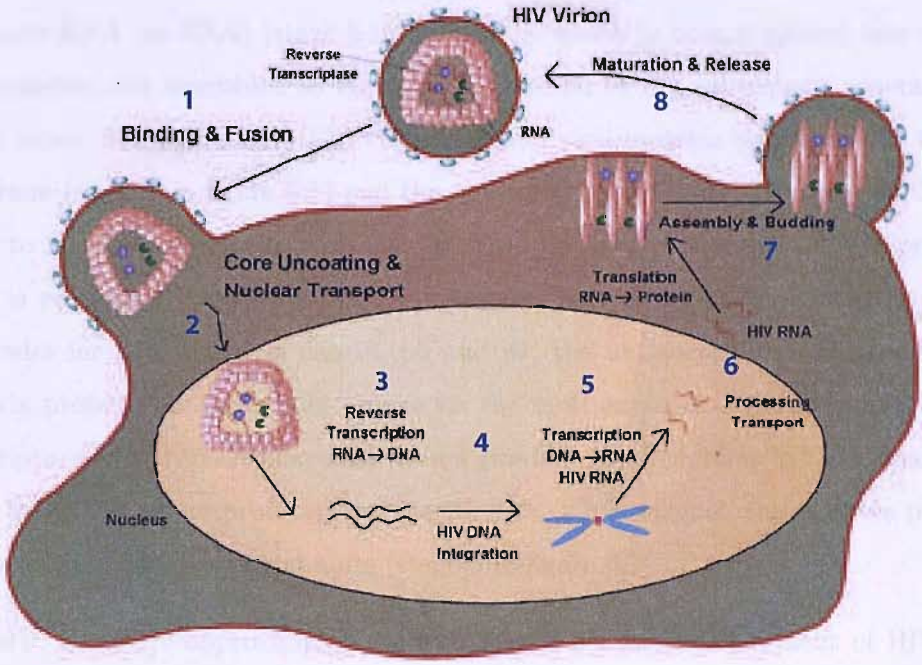


FIGURE 6.1: Life-cycle of the HIV-1 virus [178]

HIV activates infection of a vulnerable host cell by binding to the CD4 receptors (stage 1 in figure 6.1) which are present on the surface of many lymphocytes that are involved in the immune response. Recent evidence suggests that a co-receptor is required for HIV to enter the cell [179], which has created opportunities for the design of drugs which prevent HIV infection through prevention of the fusion of HIV with its host cell.

Following the fusion, HIV enters the cell (stage 2 in figure 6.1), releasing the viral RNA which undergoes reverse transcription (stage 3 in figure 6.1), resulting in the release of the single-stranded RNA from the attached viral proteins and copying to a complimentary DNA strand. This process is catalysed by HIV-1 reverse transcriptase and is very susceptible to error, and it is during this phase that mutations are likely to manifest themselves.

The viral DNA enters the host cell nucleus where it is integrated into the genetic material of the cell, catalysed by the integrase enzyme (stage 4 in figure 6.1). Once this stage is

complete, HIV may remain in a dormant state for several years. This is the main obstacle to the elimination of the HIV virus, meaning that patients must currently continue with HIV anti-viral treatment for life.

In host cells, stimulation of the T-cells results in the transcription of viral DNA into messenger RNA (m-RNA) (stage 5 in figure 6.1), which in turn is spliced into smaller viral proteins and assembled as the genetic material of the subsequent generation of viruses (stage 6 in figure 6.1). The viral RNA and viral proteins accumulate at the cell membrane (stage 7 in figure 6.1) and the new virus moves away from the cell, leaving it free to infect another cell. Amongst the viral proteins produced is HIV-1 protease, which is required to process other HIV proteins into their functional forms. HIV's gag (codes for p24, the viral capsid, p6 and p7, the nucleocapsid proteins, and p17, a matrix protein) and pol genes (codes for the viral enzymes, e.g. integrase, reverse transcriptase and protease enzymes) do not produce their proteins in their final form, but as larger aggregate proteins; the specific HIV-1 PR enzyme cleaves these proteins into their individual functional units (stage 8 in figure 6.1).

Currently, there are approximately 20 available drugs for the treatment of HIV, and with a single exception, enfuviritide (HIV fusion inhibitor which inhibits HIV entry), all of these drugs inhibit the stages involving HIV-1 RT or HIV-1 PR. Resistance that has emerged against all available antiviral drugs represents a major challenge in the therapy of HIV infection.

The enzymes, HIV-1 PR and HIV-1 IN have been chosen for study in this thesis, and literature reviews on the two HIV enzymes follow in this chapter.

6.2 HIV-1 Protease

6.2.1 Structure

Human immunodeficiency virus-1 protease (HIV-1 PR) is an aspartic protease and is one of three enzymes encoded by the viral genome. The enzyme's main role is during the course of viral maturation [180] [181] [182], where it is involved in the generation

of the important viral (gag) enzymes and structural (pol) proteins, including HIV-1 PR itself.

The HIV-1 PR enzyme (figure 6.2) comprises two identical 99 residue monomers (numbered 1-99 (A) and 100-199 (B)), with the substrate cleavage site located at the interface between them. The active site region possesses the catalytic dyad, which is typical of aspartyl protease enzymes and is covered by two identical β-hairpin loops, known as the flaps (residues 46-54 and 146-154). The flaps are maintained by hydrogen bonding and regulate substrate entry into, and exit from, the active site cleft. The regions labelled as the fulcrum (residues 11-22 and 111-122), cantilever (residues 59-75 and 159-175) and flap elbows (residues 33-42 and 133-142) in figure 6.2 are important structural regions, whose motions are believed to be concerted, linked with the opening/closing motion of the flaps and facilitating the binding of ligands [90]. X-ray crystallography studies show the flaps of the ligand-bound protease to be well ordered and to interact with inhibitors and substrates [183]. In the inhibitor-bound state, all except one of the available X-ray structures show the flaps to persist in the closed conformation, with the exception being the 1ZTZ (Protein Data Bank [8] code) crystal structure. This structure possesses a metal carborane inhibitor bound to part of the flaps, causing them to be held in an open-type conformation [184].

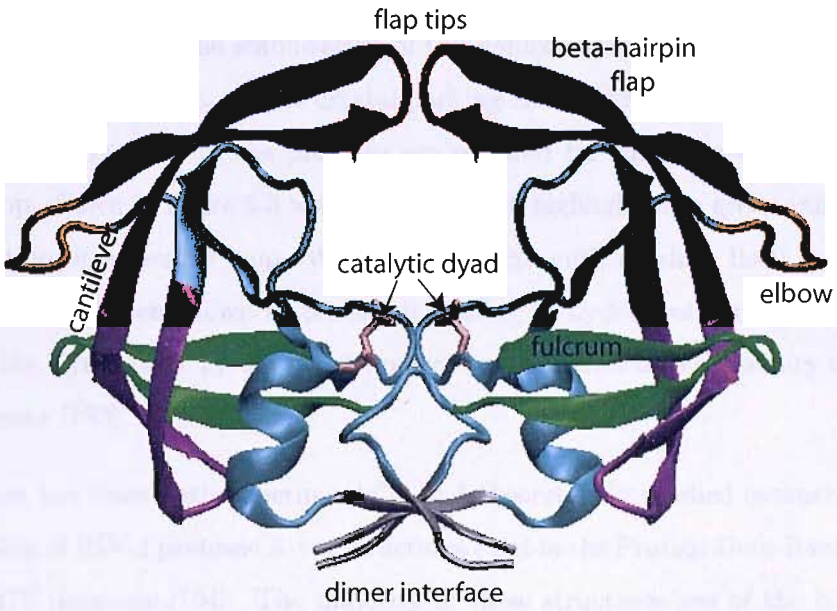


FIGURE 6.2: Side view of HIV-1 PR in semi-open conformation (1HHP [185]) with the different regions labelled.

Crystal structures of the apo-enzyme reveal more heterogenous flap structures, ranging from closed conformations such as those seen when an inhibitor is present [186], to semi-open conformations [187] [185] [188] [189] (figure 6.3) and more recently, a more open structure has been obtained with the use of mutations [190]. This crystal structure was resolved in an open conformation to 1.3 Å with the use of 10 mutations in each monomer in order to promote this open flap conformation.

As figure 6.3 shows, the flaps are shown to swap in position with respect to each other between the semi-open and closed conformations. Both the closed and semi-open conformations of the flaps prevent access to the active site [191]. There is only one structure of the apo-HIV-1 PR enzyme which shows access to the active site, resolved with the use of mutations to assist in the stabilisation of this conformation. This is due to either the opening being a rare event, and/or crystal packing effects [192], although it is presumed that open conformations of the protease are required for substrate entry and release. The P1 loop, shown in figure 6.3 with residue Tyr80 highlighted in green van der Waals representation, is a flexible loop. When the flap-tips curl, residues Ile50 and Ile150 of the flap-tips have been shown to contact a number of hydrophobic residues in the P1 loop, and the Tyr80 (and Tyr180) is postulated to play a role in the mobility of the flaps of the enzyme [193].

This system has been both experimentally and theoretically studied extensively and a large number of HIV-1 protease X-ray structures exist in the Protein Data Bank [8] (pdb) and the HIV database [194]. The majority of these structures are of the bound form since the apo-enzyme is known to be autocatalytic and as a consequence, the resolution of the unbound structures is low (2.7 - 3.0 Å).

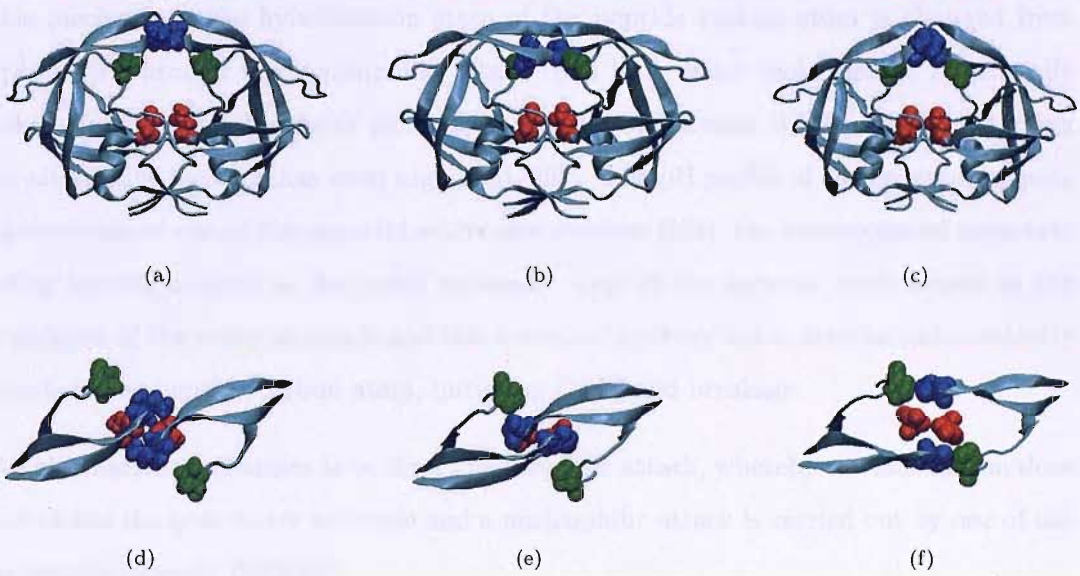


FIGURE 6.3: [a]: side view of HIV-1 PR in semi-open conformation (1HHP); [b]: side view of HIV-1 PR in closed conformation (1G6L-tethered dimer); [c]: side view of HIV-1 PR in open conformation (1TW7); [d]: top view of HIV-1 PR in semi-open conformation (1HHP); [e]: top view of HIV-1 PR in closed conformation (1G6L); [f]: top view of HIV-1 in open conformation (1TW7). Top view representations show flap (residues 44-59 and 144-159) and P1 residues (76-83 and 176-183) only. Residues coloured blue represent the flap tips (residues 50 and 150), red are active site resides (25 and 125), and green are residues of the P1 loop (residues 81 and 181).

6.2.2 Catalytic Mechanism

The catalytic mechanism of this enzyme has been studied extensively and, although it shares many features of other members of the protease family, it is still not fully understood.

On the basis of the many experimental and theoretical studies [195] [196] [197] [198] [199] [200] [201], the majority of the results are consistent with a general acid-base (GA/GB) mechanism for the peptide bond cleavage (described by Suguna [202]). In this mechanism, the hybridisation state of the peptide carbon atom is changed from sp^2 to sp^3 through the nucleophilic attack by a lytic water molecule. It is generally acknowledged that this water molecule is positioned between the aspartates, although an alternative location has been suggested [203]. The pH profile of the enzyme suggests protonation of one of the aspartyl active site residues [204], the unprotonated aspartate being located nearest to the water molecule. One of the aspartic acids assists in the ionisation of the water molecule and this activated hydroxy anion attacks and covalently bonds to the peptide carbon atom, initiating C-N bond breakage.

An alternative mechanism is by direct nucleophilic attack, whereby the mechanism does not utilise the lytic water molecule and a nucleophilic attack is carried out by one of the aspartates directly [200][205].

It is well recognised that the flaps of the HIV-1 PR enzyme play a role in inhibitor/substrate binding [183], with the flaps undergoing a large conformational change between the free and bound forms. In the one-step model, the ligand forms a collision complex with the enzyme in the open flap conformation as it enters the active site of the protease and then induces the closing of the flaps [206]. However, fluorescence and NMR studies of the kinetics of the binding mechanism of HIV-1 PR inhibitors [207] [208] [196] [209] and MD studies by Piana *et al.* [210] support the two-step binding mechanism. This model is based on the semi-open conformation being the predominant form in solution as evidenced by the existence of all crystal structures of the apo-form being in this conformation. In this mechanism, the ligand approaches the enzyme and the interactions between the flaps are disrupted from their semi-open conformation, resulting in their opening. As the ligand enters the active site, the flaps are triggered to close over

the inhibitor, positioning it in the correct orientation. The initial binding process is relatively fast, followed by the slow rate-limiting conformational rearrangements. The two models rely on different populations of the open structure, but both involve the ligand entering the active site, followed by conformational change where the flaps close over the active site and the concomitant correct positioning of the ligand in the active site. An MD study carried out by Tóth *et al.* [211] indicates the presence of a network of weakly polar interactions between the flaps in the semi-open structure. They hypothesise that this results in the persistence of this conformation based on observations in other peptides and proteins where local conformations can be attributed to the presence of weakly polar interactions [212]. Previous studies [213] show that these weak interactions can result in binding, which is suggested also to be the case for HIV-1 PR.

It is possible that the more tightly binding inhibitors appear to bind in a one-step process, but in fact involve two steps, with a faster second flap-closing event, making the lifetime of the complex short [208] [196]. This interpretation makes the timescale for flap closing inhibitor dependent.

6.2.3 Experimental Studies

NMR studies have been mainly restricted to the ligand-bound structure due to the rapid autoproteolysis of the wild-type apo-enzyme. Experimental studies by Ishima *et al.* [192] reported transverse spin relaxation rates on the apo-enzyme containing mutations to impede autoproteolysis. They revealed the flap tip residues 49-52 on each monomer to be flexible on a timescale of less than 10 ns, and another motion, involving a greater number of the flap residues occurs on a timescale of $\sim$100 μs. They suggest that the latter motion reflects an equilibrium between semi-open and open conformations, with an overall description of the flap dynamics as being an ensemble of semi-open conformations, with possibly a small population of open and closed conformations in dynamic equilibrium. The opening event is suggested to be facilitated by the interaction of the Phe53 ring with the substrate. This idea is consistent with other experimental studies [214][215], which demonstrated that the hydrophobic solvent-exposed residues Phe53 and Met46 in the crystal structures of the free protease, along with the polarisability of the Met sidechain are vital in interacting with substrates and inhibitors. Gellman [215]

explains the role of the methionine residue in the generation of binding sites, which are able to form strong interactions with non-polar surfaces of binding molecules, owing to the unusually large polarisibility of the sulphur atom in the methionine residue. Kinetic and mutagenesis studies carried out by Shao *et al.* [214] also determine Met46 and Phe53 to be hydrophobic residues which contribute to the activity of the enzyme, although the way in which they do this is clearly understood.

Freedberg *et al.* [216] used NMR to study the motions of the flaps in the apo protease. They propose that the sub-ns flap-tip movement is of biological significance in the facilitation of the release of products from the active site of the enzyme through the increase in configurational entropy.

6.2.4 Theoretical Studies of the Conformational Dynamics of the Flaps of HIV-1 PR in apo WT HIV-1 PR

The structural basis for flap separation and flap dynamics has been difficult to investigate using experimental techniques at an atomic resolution due to the short timescales associated with their dynamics, and thus theoretical studies have the potential to greatly aid the understanding of this protein system.

Studies have predicted a variety of flap motions [20] [191] [206] [211] [217] [218], and the rare opening event has been observed in a number of these simulations. In studies where a range of conformations are demonstrated, including the opening/closing events, implicit solvent models and enhanced sampling techniques have been used to overcome the timescale barrier in using conventional MD. Recently, several theoretical studies have reported the occurrence of reversible events.

Studies carried out by Rick *et al.* [191] identify a reaction path between the semi-open and closed apo X-ray crystal structures, and the free energy along this path has been calculated using umbrella sampling. The transition from the closed to the semi-open is reported to occur through a collapse and reformation of the beta structure of the flap tips. The semi-open structure is reported to have a higher potential energy and weaker interactions between the flaps than the closed, but is stabilised by entropy. The closed structure is stabilised by stronger interactions (hydrogen bonds) between the flaps. It is

hypothesised that the weaker interactions may allow the flaps to open to bind substrates more easily and rapidly than the closed.

A 10 ns MD simulation carried out by Scott and Schiffer [206], starting from the semi-open 1HHP [185] crystal structure and using explicit solvent showed the flap tips to curl inwards and for the flaps to open, making the active site accessible after just 3 ns which then remained stable for a further 7 ns. The large flap-opening motion observed in this simulation is shown by NMR experiments [192], to occur on a 100 μs timescale and is thought to be a rare event. It has been disputed that the conformational change they see is genuine. It has been suggested that the cause of the conformational change they report is actually an artefact of the use the GROMOS 43A1 force field [219] [220], which is perhaps inappropriate for this purpose [221]. Meagher *et al.* attribute the flap dynamics to an inadequate system set-up after investigating the effects of different solvation routines on HIV-1 PR [222]. They conclude that a poor solvation routine can lead to the formation of a vacuum region in the protein, which can encourage conformational deformities, which is perhaps the case in the simulations of Scott and Schiffer [206].

Additionally, Perryman *et al.* report results of MD studies of 22 ns in length, which compare the dynamics of the WT and a more flexible double mutant (V82F/I84V) in explicit solvent [221], starting from a closed apo structure. The results show both flaps of the more flexible mutant, and only one flap of the WT enzyme, to sample the semi-open conformations. Neither of the simulations display the large conformational change seen in the Scott and Schiffer [206] studies.

David *et al.* [223] have carried out simulations starting from the closed apo-enzyme, comparing three different implicit solvent models. Simulations of two of the implicit solvent models (Generalised Born [52] and finite-difference Poisson equation [224] [225] [226] [227] [228]) show similar conformational distributions for both flaps, where they remain close to the starting conformation. However, in the simulation using the distance-dependent dielectric model, the flaps of the protease are shown to collapse into the active site. David *et al.* suggest this to be an incorrect result, since the crystal structure shows the flaps to move away from the active site, rather than moving deeper into the binding-site cavity.

Several more recent studies obtain a larger range of flap conformations of the apo HIV-1 PR enzyme, several of which are shown to be reversible. Tóth *et al.* [211] carried out a number of simulations using Langevin dynamics at temperatures in the range 296 - 321 K with simulation lengths varying from 1 to 5 ns. The simulations used a continuum solvent with a low viscosity and considered only torsion angles (bond lengths and bond angles were set rigid) which enabled the increased sampling. Starting from the apo semi-open 1HHP crystal structure [185], the flaps were observed to open, revealing an accessible active site in several of the simulations. Other stable conformations, including the curling of the flap tips were also observed.

Hamelberg *et al.* [229] used accelerated molecular dynamics to study the apo-closed and inhibitor-bound closed forms of HIV-1 PR. This method allows the system to evolve in a modified energy landscape and accelerates the escape of conformations trapped in potential energy wells [167].

During these simulations, flap tip curling and sampling of closed, semi-open and open conformations in the apo form of the enzyme were observed, whereas the flaps of the inhibitor-bound structure remained in the closed conformation. Throughout the simulations, frequent *trans-cis* isomerisation of the Gly-Gly peptide bonds in the flap tips, both before and after opening events occurred. These transitions were seen in the flexible regions, mostly in the presence of a Gly-Gly motif. The timescale of *trans-cis* isomerisation of the Gly-Gly peptide bonds is typically very slow and in the order of seconds [230], similar to that of the other residues. However, simulations suggest that constraining the Gly-Gly residues in a loop increases the rate of *trans-cis* isomerisation significantly. Therefore, the β-hairpin loop at the flap tips could result in the accelerated isomerisation and increased flexibility of Gly-Gly, thus resulting in the shorter timescale for the isomerism event. This theory has been tested in simulations involving linear and cyclic peptides containing the same residue sequence present in the flap tips. The rate of isomerism has been estimated in the systems and the rate is predicted to decrease to around a microsecond to millisecond timescale due to the loop formation.

However, the isomerism events they observe have not been reported in any of the other studies, and Hornak *et al.* [90] have specifically disputed it on the basis of their own investigations. They carried out MD simulations using a implicit solvent model with

apo-closed and closed inhibitor-bound starting structures, the simulations being 42 ns and 28 ns in length respectively. The simulation involving the closed inhibitor-bound structure showed the flaps to remain in their initial closed conformations whereas in the simulation involving the apo-closed structure, a number of conformations were observed, including reversible sampling of the semi-open conformation and intermediate conformations. They noted that at no point in the simulations did the *trans-cis* isomerism of peptide bonds occur.

Wiley *et al.* [133] use the enhanced sampling technique of reversible digitally filtered molecular dynamics (RDFMD) [231][4][5][6] to study the flap dynamics of the WT apo-HIV-1 PR enzyme. They reveal the flaps to undergo three types of large-scale conformational change, flap swapping from the semi-open to closed position and motion resulting in increased access to the active site, caused by flap opening and flap-tip curling. The flap opening motion observed is reported to be similar to that seen by Scott and Schiffer [206], and in the occurrence of the flaps moving from their semi-open to closed position, as observed in the work of Rick *et al.* [191], no flap separation is seen, with the conformational change occurring through movement of the flap-tip residues only. In addition, the authors also suggest that flap-opening from the closed state is also possible in the absence of contact between the flap-tips.

Tozzini *et al.* [116] used a coarse-grained model where a single bead was used to model each residue, together with Langevin/Brownian dynamics to simulate the apo-form of HIV-1 PR. Several simulations of varying lengths were generated, performed at a range of temperatures. The simulations started from the semi-open conformation (coarse-grained model using coordinates of 1HPP crystal structure [185]) and a number of opening and closing events were observed, revealing four types of conformation, closed, open, wide-open and semi-open. They also show that in the open form, the flap-tips have van der Waals contact with the laterally located residue 80 (residue of P1 loop) of the opposite monomer. This observation agrees with the same observation seen in a crystallographic study of the more open HIV-1 PR structure (pdb code 1TW7) [190].

6.2.5 Experimental and Theoretical Studies: Inhibitor-Bound Structures of Apo and Mutant HIV-1 PR

The binding affinity of ligands is defined by the Gibbs free energy of binding, ΔG, which maybe reduced to changes in enthalpy (ΔH) and entropy (ΔS) (equation 6.1).

$$\Delta G = \Delta H - T\Delta S \tag{6.1}$$

To obtain high binding affinity, drugs are usually designed to fit the shape of the target binding site, based on the "lock and key" principle. This results in constrained ligands which undergo a low conformational entropy loss on binding. In addition, ligands are commonly designed to be hydrophobic, which also increases binding affinity to the target site.

These principles have been employed in the design of the first generation of HIV-1 PR inhibitors, whose shapes are complimentary to the binding site of the enzyme, and structures are rigid and hydrophobic (e.g. saquinavir and indinavir). The main driving force in the binding of these inhibitors is the change in entropy, and the enthalpy change on binding is usually unfavourable for this generation of inhibitors. Owing to their rigidity, they are susceptible to the effects of mutation [232] [233]. Several mutations result in the distortion of the binding site from the WT geometry, resulting in reduced interactions with the inhibitor, and consequently a reduced binding enthalpy.

One way to improve the binding is to increase the flexibility of the ligand, which has been incorporated into the more recent HIV-1 PR inhibitor designs e.g. KNI-764, but the disadvantage is the subsequent increase in the conformational entropy on binding, and therefore, a good binding affinity would require the binding enthalpy to be more favourable.

A promising alternative approach is the design of allosteric protease inhibitors which have been suggested in a few theoretical studies [221] [234] [235], These studies suggest viable regions for the binding of allosteric inhibitors to be, the elbow region, interfering with the dynamics of the enzyme's flaps [221] [234], and the dimer interface, which has been shown to indirectly reduce the binding affinity of the substrate [235] (see figure 6.2

for location of regions of HIV-1 PR). There has been some development of dimerisation inhibitors, which act by dissociating the HIV-1 PR enzyme into inactive subunits. Early development of this class of inhibitors has involved the use of peptides with N- and/or C-terminal amino-acid sequences to mimic the β-sheet structure of a monomer of the enzyme [236].

Approximately twenty of the ninety-nine residues of each monomer of HIV-1 PR have been reported to mutate [237] [238] [239] to give drug resistance. The emergence of mutations can be rapid due to the selective pressure exerted by the presence of inhibitors. Some of the most commonly occurring mutations are listed in table 6.1.

The mutations are described as either primary or secondary (compensatory). Primary mutations are commonly located in the vicinity of the active site (e.g. V82F/I84V), but also have been reported to exist at locations distal to the active site (e.g. L90M). These mutations are able to reduce or even completely diminish the efficacy of current Food and Drugs Administration (FDA) [240] approved inhibitors, usually by interfering with their binding, as demonstrated by clinical studies [241] [242]. Owing to the active site being located in the substrate-binding cleft, primary mutations located here are likely to affect the processing of the natural substrates, thus conferring a fitness cost to the protease enzyme.

Compensatory mutations may be located both distal or near the active site and appear to enhance the enzymatic function of the HIV-1 PR enzyme, partially restoring some of the activity lost due to the presence of primary mutations [243] [244] [245] [246] [247] [248]. They enable sufficient catalytic activity for the enzyme to function, whilst preventing the functionality of the binding inhibitor. In many cases, these mutations involve chemically similar residues (e.g M46I and L63P) which do not change the electrostatic field significantly [249]. Thus, resistance to HIV-1 PR depends on the step-wise accumulation of several primary and secondary mutations to yield an enzyme capable of discriminating protease inhibitors (PIs) from natural substrates, in order to maintain sufficient catalytic activity required for viral replication.

HIV-1 PR Mutations									
Nelfinavir	D30N		M46I\L			V82A\T\F	I84V	N88S	L90M
Saquinavir				G48V		V82A	I84V		L90M
Indinavir			M46I\L			V82A\T\F\S	I84V		L90M
Ritonavir		L33F	M46I\L			V82A\T\F\S	I84V		L90M
Amprenavir			M46I\L	I47V	I50V		I84V		L90M
Lopinavir		L33F	M46I\L	I47V	I50V	V82A\T\F\S	I84V		L90M
Atazanavir		L33F	M46I	G48V	I50L	V82A	I84V	N88S	L90M
Tipranavir		L33F	M46I			V82A\T\F\L	I84V		L90M

TABLE 6.1: Some of the mutations associated with HIV-1 PR inhibitor resistance [250].

Interpretation of mutational data using genotypic or phenotypic assays can assist in treatment of the virus. However, the coexistence of several mutations at one time in the protease enzyme is usual and assessing their collective effect is a more complex process. This emphasizes the importance of being able to understand mutations on a molecular level.

It has been demonstrated that the HIV-1 PR enzyme must cleave the polyprotein Gag and Pol precursors at least nine different cleavage sites in the life-cycle of the virus. It has been demonstrated that there is little homology in the primary amino acid sequence of the each of the cleavage sites of the substrates [251]. Theory implies that recognition of a substrate is based on the conserved shape (i.e the secondary structure of the conserved regions) in each of the cleavage sites rather than a particular sequence [251]. Substrates position themselves in an extended conformation which provides space for water molecules, necessary for substrate recognition and/or product release [252]. In addition, structural analysis of PIs complexed with HIV-1 PR have been compared with the enzyme complexed with natural substrates [253] and show that, although PIs are chemically unique from each other, they occupy a similar volume of space within the binding cavity. They are shorter than the natural substrates and occupy different areas of the binding cavity where they may engage with the enzyme using van der Waals interactions (e.g. residues G48, I50, V82 and I84). Structural studies have shown that, at these locations, there is no interaction of the natural substrates with the enzyme, thus making them favourable locations for primary mutations, decreasing the efficacy of current inhibitors, whilst having little effect on the natural substrates [253] [254] [255]. Further evidence for this theory of PI resistance is demonstrated by the mutation profile

of amprenavir. This PI fits in the binding cavity in a more similar fashion to that of the natural substrates compared with the other PIs and, as a consequence, the mutation profile is notably different from that of the other PIs.

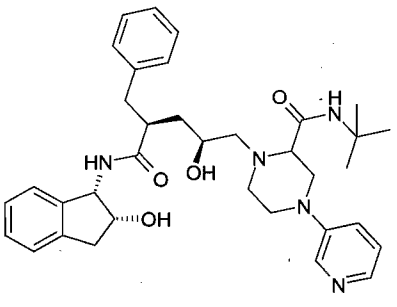
A further consideration is hydrogen bonding. Natural substrates possess hydrogen bonding between themselves and the backbone atoms of the enzyme, making the residues involved conserved. However, in the case of the majority of the PIs, bonding is formed between the side chains of the residues rather than the backbone, which can be disrupted by mutation [256] [257] [258] [259] [260] [261] [262] [263] [264].

Kinetic binding affinity experiments demonstrate differences in the ability for protease inhibitors (PIs) to bind in the presence of certain mutations, and clinical studies have shown that each PI tends to result in the occurrence of particular primary mutations and therefore subsequent secondary mutations [265] [266] [267] [268].

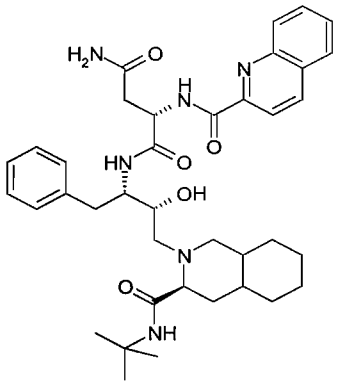
Currently, there are approximately 10 inhibitors approved by the FDA which target the HIV-1 protease enzyme, the majority of which are peptidic and all of which are competitive with the substrates for binding in the active site of the enzyme (see figures 6.4 (a)-(c) for chemical structures of such inhibitors). The main driving force for the binding of these inhibitors is the associated large positive entropy change, as a result of the burial of a large hydrophobic area [269].

Inactivation of HIV-1 PR by mutation or inhibition results in the production of immature, non-functioning virus proteins. Peptidic PIs share a structural similarity to the intermediate formed during hydrolytic cleavage of a peptide bond in the natural substrate [270] and all hydrogen bond to the flaps of the enzyme with the use of the ubiquitous water molecule. This active site water molecule links the substrates to the enzyme through hydrogen bonding to the flap-tip residues, Ile50 and Ile150, and to the CO groups of P1 and P2 binding sites on the reaction intermediate. There is current investigation into the development of non-peptidic PIs (e.g. tipranavir - see figure 6.4 (c)) which are able to bind to the flaps directly, not requiring the use of the water molecule needed by peptidic PIs and natural substrates [271]. This approach to binding is hypothesised to be entropically more favourable, thus providing an advantage to this class of inhibitors. It is also proposed to provide the inhibitor with increased flexibility,

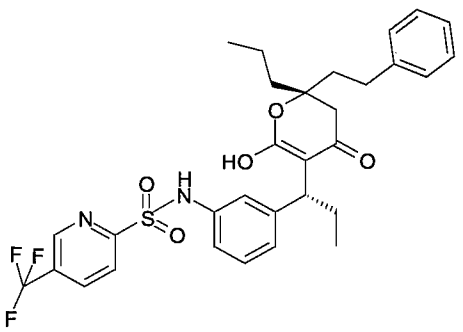
allowing adjustment to residue changes in the active site [272]. In addition, tipranavir also possesses hydrogen bonds with the active site residues of the enzyme, which includes bonding with the catalytic residues Asp29 and Asp30 and with conserved backbone residues Gly48 and Ile50 [273] [274] which puts it at an advantage over previous PIs. However, resistance to tipranavir still occurs, although an accumulation of up to ten mutations is required to achieve a high level of resistance [275].



(a) Indinavir



(b) Saquinavir



(c) Tipranavir

FIGURE 6.4: Chemical structures of (a) indinavir (b) saquinavir and (c) tipranavir.

Mutational studies involving flap residues tend to produce mutants with near WT specificity [207] and suggest that if the sequence of the flaps is not involved in the specificity of substrates then maybe the dynamics is important in altering the activity of the enzyme. It has been suggested that some mutations affect the equilibrium between the closed and open states, which in turn affects the dissociation rates and thus binding affinities of the inhibitors [276]. Kinetic experiments have been carried out on the L90M, G48V and G48V/L90M mutants and show a decreased binding affinity for inhibitors caused by increases in dissociation rates, due to increases in flap opening, decrease in flap closing or both. The association rates were not shown to be significantly affected [209].

A number of theoretical studies have demonstrated differences in the flap dynamics in the presence of mutations compared with the WT HIV-1 PR.

The double V82F/I84V mutant impedes the binding of all current FDA approved drugs by decreasing the binding affinity of the inhibitors [277] [278] [279]. A calorimetric analysis carried out by Todd *et al.* [269] finds the double mutant to cause the binding enthalpy to be more positive when binding to inhibitors nelfinavir, indinavir, saquinavir, and ritonavir. The major enthalpy effect is consistent with the idea that the introduction of the bulkier sidechains of Phe82 and Val84 distorts the binding site, disrupting the van der Waals interactions that exist between the WT enzyme and these rigid inhibitors. An additional contribution of the decreased binding affinity of the double mutant originates from the increase in energy penalty associated with the conformational change of the protease on binding.

This primary mutant is thought to cause a shift in the equilibrium which favours the semi-open conformation of the flaps. A 22 ns explicit solvent MD study carried out by Perryman *et al.* [221], compared the dynamics of the apo-closed WT with the apo-closed double mutant HIV-1 PR. They conclude that the flaps of the double-mutant to display increased flexibility and to open further, with more flap-tip curling activity compared with the WT. This would cause a detrimental effect on the binding of the inhibitor as a greater energy cost would have to be paid by the inhibitor to close the flaps in order for binding to take place. However, Hornak *et al.* dispute the theory that flap-tip curling is a trigger for flap opening in the enzyme [206]. In their simulations flap-tip curling has been observed to accompany some flap motions, however, it was not noted

to be specifically linked with the large opening events. Owing to the difference in the timescales of the occurrence of these events, reported by NMR experiments [280] [281], they disagreed that the flap tip curling and the larger motion would be correlated.

A further investigation into flap activity was carried out by Perryman *et al.* [234] with the use of six different MD simulations using distance restraints. In the earlier study, they revealed the presence of an anticorrelated relationship between the compression of the elbow of the flap to the fulcrum and the cantilever (authors refer to this as the ear-cheek region) occurred as the flaps opened, and the opposite for the closing of the flaps. Therefore, distance restraints between the ear and cheek regions were used to further investigate this effect. In the absence of an inhibitor, distance restraints causing the ear and cheek regions to be held away from each other (flap-closing restraints) were shown to result in the WT and V82F/I84V HIV-1 PR systems to remain closed. When flap-opening restraints were applied to the WT and mutant HIV-1 PR apo-systems, the V82F/I84V HIV-1 PR simulations observed more flap-opening compared to the WT simulation, supporting their hypothesis of the drug resistance caused by this mutant in an earlier study [221]. Additionally, simulations were carried out in the presence of an inhibitor with the application of the flap-closing restraints, resulted in an improved binding affinity and therefore suggest the flap-cheek region to be a possible target for allosteric inhibitors.

In addition to the MD studies, thermodynamic analysis of the WT and V82F/I84V mutant [269] suggests the mutation decreases the binding affinity in two ways. The first way is through the direct alteration of the interactions between the inhibitor and the protein, and secondly, by altering the stability of the semi-open and closed conformations, causing an increased energy penalty for the flaps to close.

In contrast to the studies which suggest the V82F/I84V mutant to affect the equilibrium of semi-open/closed flap conformations relative to the WT, an MD study by Hou *et al.* [282] contradicts this. MD simulations of 18 ns in length, using explicit solvent were performed on the apo WT and mutant V82F/I84V HIV-1 PR systems, the results of which conclude the dynamics of the two systems to be similar. Although comparing the activity of the flap-tip of monomer B of the WT and mutant HIV-1 PR does show

some differences (measurement of angle between the three flap-tip residues G48-G49-I50), with the mutant flap-tip sampling a broader range of angles. Instead, they claim the observed distortion of the geometry of the active site and the slightly improved stability of the mutant to be the reason for drug resistance and state that a change in flap dynamics does not play a significant role.

Brownian dynamics simulations in a coarse-grained model (where each residue is represented by one bead placed on the α- carbon) of HIV-1 PR are used by Chang *et al.* [283] to study the opening/closing dynamics of the flaps of three mutants G48V/V82A/I84V/L90M, G48V and L90M with three inhibitors, amprenavir, indinavir and saquinavir. They find the presence of mutants to affect the flap dynamics and the predicted association constants to agree well with experimental work. Simulations of the HIV-1 PR systems possessing the G48V and L90M mutants show the flaps' flexibility and opening frequency to be similar, but the flaps of the quadruple mutant HIV-1 PR enzyme are shown to open less frequently. This result again demonstrates the ability of mutations to affect flap dynamics and thus to contribute towards the binding affinity of inhibitors.

Binding affinity experiments involving the compensatory mutant M46I report an increased catalytic activity and have shown this mutant to not affect the overall global motion, but to cause subtle differences in the dynamics. However, this mutant is reported to have little effect on the affinity of the majority of inhibitors [254] [284] [255], indicating the probable role of other factors other than the binding affinity to be of significance. As is the case in the majority of compensatory mutations, only minor rearrangements in the structure occur which cannot account for an increase in catalytic efficiency [256] [255] [285].

A number of studies have highlighted the difference in flexibility between WT and mutant protease enzyme structures and suggest this is a prominent factor in the catalytic efficiency of the enzyme [256] [248]. A number of X-ray crystallographic studies support this theory with studies on compensatory mutants K45I [261], M46I [256], L90M [263] showing decreased flexibility as demonstrated by lower B-values. A theoretical study carried out by Collins *et al.* [218] compares the dynamics of the flaps of the apo M46I mutant with the WT structure. Simulations are carried out in vacuum and employ the use of activated molecular dynamics, a method which allowed the flaps to move freely

whilst tethering other atom positions to the target coordinates with harmonic force. Their results showed the flaps of the mutant enzyme to be less flexible, with increased stabilisation in the closed conformation compared with the WT structure.

Multi-nanosecond studies have been carried out by Piana *et al.* [286] using a L63P substrate-bound HIV-1 PR complex, a L63P WT structure containing the M46I mutant bound with a substrate and another, bound with a reaction intermediate (L63P considered as WT). The L63P-M46I complexes for both the substrate and intermediate are shown to be less flexible, although the average properties are shown to be similar to the L63P-WT when the simulation properties are averaged. Additionally, the L63P-M46I mutant is shown to induce displacement of the reaction intermediate into the active site, thus explaining the increase in catalytic efficiency.

As described previously, the flaps of the bound HIV-1 PR are swapped relative to their position in the semi-open conformation. The inhibitor-bound HIV-1 PR structures are stabilised in the closed conformation owing to the interactions between the substrate/inhibitor and the enzyme, demonstrated by NMR and crystal structures of the ligand-bound HIV-1 PR always being the closed conformation. As a consequence, the same range of conformations seen in the apo-HIV-1 PR have not been seen experimentally or theoretically for the bound HIV-1 PR.

A theoretical study carried out by Chang *et al.* [287] used a coarse-grained Brownian dynamics simulations of HIV-1 PR, where one bead represents one residue in HIV-1 PR and all-atom MD simulations in implicit solvent to study the binding of a small cyclic inhibitor (XK263) and a substrate molecule. Using the coarse-grained model, they find the flaps of HIV-1 PR to reversibly sample open and closed conformations when the ligand is away from the binding site, the dynamics being comparable to those of a previous study performed on the apo-form of HIV-1 PR [283] [116]. They find the dynamics of the flaps to change when the substrate is located in the active site. When the substrate has entered the active site and the flaps are still open, they remain open for a longer period of time than observed in the apo-HIV-1 PR. Similarly for the closed state of the flaps of the bound HIV-1 PR, the flaps remain closed for a longer period of time than seen in the apo-HIV-1 PR, demonstrating the substrate to interact with the enzyme, resulting in the observed change of dynamics. This change in dynamics was

also noted in the presence of the cyclic inhibitor, although the extent of the motion was far reduced.

Additionally, the simulations proved that the flaps are required to open for the substrate to bind, and once bound, the flaps must close in order for the processing of the substrate to take place. In cases where the flaps were closed in the vicinity of the substrate, the substrate either diffused away or remained until the flaps opened. The authors claim that, in the latter case, the substrate is able to form an interaction with one of the flaps, which may result in the asymmetric motion which is seen, resulting the weakening of the interactions between the flap-tips and causing the opening of the flaps. The same conclusion of interaction of substrate/inhibitor with the flaps has also been suggested by other studies [288].

The cyclic peptide inhibitor (XK263) is found to be able to bind in the presence of the open and closed flap conformations, as, unlike the substrate, it is able to bind from the side, in addition to entrance from the top of the active site cavity. In cases where it binds from the side, the inhibitor is found to take an extended amount of time to assume the conformation for inhibition, compared with binding from the top. All-atom simulations on the same systems were also carried out in the same study using implicit solvent. These simulations confirmed the ability of the inhibitor to enter the active site cavity from the side without the opening of the flaps, and also suggest the inhibitor to be able to induce partial flap opening to the enzyme without the flaps being in their fully open conformations. The latter observation was not seen in the coarse-grained simulations, which the authors attribute to the force-field.

Two MD simulations carried out by Tóth *et al.* [288] were performed on the HIV-1 PR with a substrate molecule. One of the simulations possessed the flaps in a closed conformation, whilst the other started with the flaps in the open conformation. The starting structure of the second MD simulation was obtained from a previous simulation using the apo-HIV PR and the substrate was superposed into the active site. Again, they observe the flaps to close in an asymmetrical manner, with substrate interaction with the hydrophobic cluster Ile50 and Pro81 of monomer B and Ile50 of monomer A. This intermediate state induced the flap of monomer B to form interactions with the

substrate and the consequent closing of the flap, which causes the final closing stage of the flap of monomer A.

This closing process of the flaps correlates with the stabilisation of the β-sheet structure of the flaps and their reduction in flexibility, and is attributed to the side chain of the aromatic ring of Phe53 flipping to the opposite strand and overlaying the backbone of Gly48 and Gly49. This stage is deemed as necessary for the formation of flap-substrate interactions in the closing of the flaps, and the same has also been observed in the opening of the flaps from the semi-open apo form of HIV-1 PR [211], suggesting this to be an intermediate step. Its purpose is to stabilise the flaps in a state which ensured the closing of the first flap, while it also initiated the closing of the second flap.

The open conformations generated in a previous study by Hornak *et al.* [90] have been used in docking studies after deletion of the inhibitor [289]. Six simulations were carried out using implicit solvent, each varying in the degree of accuracy of the docking. Five of the six simulations revealed rapid conversion of the open to closed conformations of the flaps in the presence of the inhibitor. One of the six simulations observed the flaps of the enzyme not to close over the inhibitor. These results suggest that the inhibitor presence triggers the flaps to adopt the closed conformation.

Summary of Theoretical and Experimental Studies

Structural studies show the flaps of the apo-form of HIV-1 PR to exist predominantly in the semi-open conformation, whereas the inhibitor-bound structures show the flaps to be closed over the active site, with a swapped flap conformation compared with the WT.

The flaps have been demonstrated to play a crucial role in the catalytic mechanism and thus their dynamics is of significant importance in the control of access of substrates to and from the active site.

NMR [192] [216] and theoretical studies [217] [20] [218] [191] [206] [211] of the apo HIV-1 PR show the flaps to undergo a variety of motions corresponding to the opening/closing of the flaps and flap-tip curling. On binding to a substrate or ligand, interactions between the substrate and the protein are postulated to occur [287] [288], facilitating

the opening and subsequent closing of the flaps in order to allow the processing of the substrate. Where the enzyme is inhibitor-bound, the stronger interactions prevent the flexibility of the flaps as seen in the apo-form of the enzyme, and no significant conformational changes of the flaps have been reported as of yet.

Mutations occur at many of the amino acid sites of HIV-1 PR, with their location being influenced by the presence of inhibitors, decreasing or completely diminishing drug efficacy. Primary mutations often exist near the active site cavity and significantly reduce the efficacy of drugs and the substrate processing activity of the enzyme. Secondary mutations or compensatory mutations often arise in order to restore some of the catalytic activity lost by the presence of the primary mutations to a level where processing of the substrates can occur, whilst still being inhibited by drugs. Experimental and theoretical [221] [234] [269] [283] [218] [286] studies have suggested that some of these mutations may act by affecting the conformational dynamics of the flaps, altering the equilibrium between the semi-open/open and closed conformations compared with the WT.

6.2.6 Protonation State of the Catalytic Aspartate Residues

Apo HIV-1 PR.

Determination of the protonation state of the catalytic aspartate dyad has been an area of interest in studies of HIV-1 PR and has not been fully resolved. The location of the acidic proton in the apo-enzyme is not easy to determine since proton locations are not resolvable by X-ray crystallography [202] [290].

A common feature of both the two suggested mechanisms of action which have been proposed for HIV-1 PR, the nucleophilic [291] and GA/GB mechanism [195] [200] [292], is the requirement that one of the catalytic aspartates possess a hydrogen, causing it to be neutral, whilst the other is negatively charged. Experimental calculations of the pKa values of the aspartic acids confirm this, reporting pKa values of ~3.5 and ~6.5, with a maximum stability and enzymatic activity occurring at pH ~5-6 [195].

However, there is little direct experimental evidence of any structural difference if the aspartates are deprotonated or monoprotonated.

An NMR study by Smith *et al.* [293] reveals that at pH 6, the aspartates in the apo-form are both dianionic. However, MD studies and *ab initio* calculations involving the apo-enzyme [294] [210] find the aspartates to be almost ionic, sharing one proton; and suggest that this may the cause of the conclusion of Smith *et al.* [293] for the dianionic state.

Wang *et al.* [295] also conclude a dianionic state for the catalytic aspartates. They carry out MD simulations of 120 ps in length, using an implicit solvent model. They measure the distance between the catalytic aspartate residues and conclude from this, that the distance closest to that seen in the crystal structure is when both aspartate residues are dianionic. However, only small differences in distance are seen with all protonation states investigated, and the simulation time may not be a significant length

MD studies of both Perryman *et al.* and Scott *et al.* [221, 206] use dianionic catalytic aspartates (personal communication), which, along with their protonation state of histidines in their system, is consistent with simulations at a neutral pH.

Other MD simulations investigating the aspartyl group protonation state have been carried out by Wiley [7], and reveal instability of the protease structure in the absence of the proton on the Asp125.

MD studies carried out by Kovalskyy *et al.* [296], investigate the protonation state of the catalytic aspartates. Three MD simulations were carried out, one with a monoprotonated dyad and one with a deprotonated dyad, both of which included a physiological concentration of sodium and chloride ions. The third simulation simulated the deprotonated dyad with just 5 sodium counterions. The simulations revealed no significant difference in the dynamics between the two protonation states, with a "self-blocked" conformation seen in both cases, where the active site is blocked by the flap tips. However, in the simulations involving the deprotonated dyad, the stability of the protease was shown to be sensitive to the number of counterions and a sodium ion is suggested to stabilise the protease at neutral pH. The authors suggest the occurrence of some large conformational changes seen in previous simulations involving a deprotonated dyad may be attributed to the absence of a stabilising counterion to shield the repulsion of the negatively charged aspartates.

In this thesis, the apo-form of the enzyme has been simulated at its most active pH, where the structure is shown to be most stable, and so a monoprotonated catalytic dyad, with the proton lying on the aspartyl group of monomer B (residue Asp125) has been used.

Inhibitor-bound HIV-1 PR

It has been suggested that the protonation states of the catalytic aspartates (residues 25 and 125 of monomer A and B respectively) are dependent on their environment. Therefore, in the inhibitor-bound enzyme, the class of inhibitor will influence their protonation [217] [290], and the protonation of the aspartyl dyad may not be the same as in the apostructure.

The deprotonated state of the aspartates is associated with the binding of positively charged inhibitors [297] and the diprotonated aspartates, with the binding of inhibitors which contain two oxygen atoms in their core (e.g. cyclic ureas) [298] [299] [300]. The protonation state of the aspartates of the saquinavir (SQV) and indinavir (IDV)-bound structures has been investigated by Havlas *et al.* [301] who optimise hydrogen positioning using the Gaussian program at the B3LYP/6-31*G level and calculate the energy of the inhibitor-bound structures. Their results agree with those obtained with semi-empirical methods, where the hydrogen of saquinavir is preferably placed onto the OD2 of monomer A, and on OD2 of monomer B for indinavir, thus a monoprotonated state for the catalytic aspartates. These protonation states are in agreement with other studies which investigate the protonation states of the catalytic aspartates when bound to IDV and/or SQV [302] [303] , and as a result have used in simulations carried out in this thesis.

6.3 HIV-1 Integrase (HIV-1 IN)

6.3.1 Structure

The HIV-1 integrase (HIV-1 IN) enzyme is one of the three enzymes involved in the life-cycle of the HIV virus. This enzyme is responsible for the integration of the viral

DNA into a host chromosome required for viral replication. Prior to this integration process, DNA is cleaved on each strand, which is then attached to the targeted DNA using the integrase enzyme, a process known as strand transfer.

The full length HIV-1 integrase enzyme comprises of 288 residues which, based on partial proteolysis, functional and structural studies, can be divided into three domains, the N- and C-terminal domains and the catalytic domain [304] [305] [306].

The N-terminal domain (residues 1-50) consists of three α-helices and a zinc binding site, where conserved histidine and cysteine residues coordinate a zinc ion, stabilising the interaction between the helices. The binding of zinc is thought to promote the multimerisation of the enzyme which enhances the activity of the enzyme [307] [308] [309]. The third, C-terminal domain (residues 212-288) is a non-specific binding site of DNA [310] [311] [312] [313] [314] and additionally contributes to the oligomerisation, which is essential to the integration process [315]. In isolation, both the C- and N-terminal domains are dimeric in solution, but the C-terminal is monomeric when attached to the catalytic domain [316].

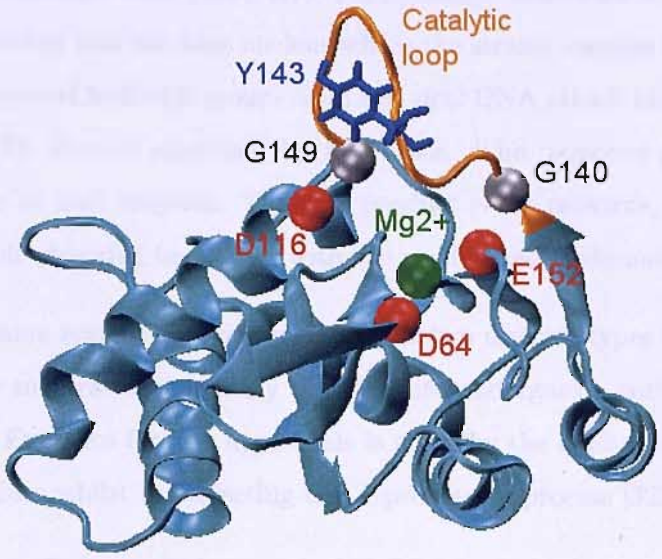


FIGURE 6.5: HIV-1 Integrase enzyme (front view) with the following highlighted residues : α-carbon atoms of active site residues, D64, D116 and E152 (red), catalytically important tyrosine residue, Y143 (blue) and α-carbon atoms hinge residues, G140 and G149 (silver), Mg^{2+} ion (green). The catalytic loop residues (140-149) are shown in orange.

The central catalytic domain (residues 50-212) is composed of a mix of α-helix and β-sheet structures (see figure 6.5). This domain is the most conserved of the three; being responsible for the recognition of the subterminal conserved CA/TG dinucleotide pair near the viral DNA end and contains the active motif, comprised of residues Asp64, Asp116 and Glu152 (DDE).

All three of the domains of the full-length HIV-1 IN are required to carry out the catalytic function of the enzyme, which involves the 3'-processing and strand transfer [305]. However, in isolation, the catalytic domain has the ability to initiate the process of disintegration (the excision of viral DNA) *in vitro* [317] [318]. These two stages of the HIV-1 virus life-cycle occur after the viral-RNA genome is reverse transcribed to produce the double-stranded DNA [319] [320], a process carried out by the HIV-1 reverse transcriptase enzyme. The 3'-processing stage is initiated by the recognition of the two nucleotides, 5'-CA-3', which is followed by the removal of the terminal 5'-GT-3' nucleotides from the viral DNA, causing the hydroxyl group of nucleotide 'A' to become exposed [321]. Following this stage, the HIV-1 IN enzyme, complexed with the pre-integration complex (PIC), which comprises of a number of viral and cellular proteins

including the processed viral DNA, HIV-1 RT, matrix and other auxiliary complex, is actively transported into the host nucleus where the strand transfer takes place. In this reaction, the exposed hydroxyl groups from the viral DNA attack phosphodiester bonds in the host DNA strands separated by five bases. The unpaired gap is then closed, presumed to be by host enzymes. The final product is the provirus, which comprises of the intact double-stranded host DNA with the viral genetic information inserted.

Owing to the same active site being used for the two distinct types of processes, which utilise different substrates, it is likely that the site undergoes a conformational change for each stage. Evidence for this hypothesis is given by the ability to selectively inhibit the strand transfer whilst not affecting the 3'-processing process [322].

The arrangement of the three domains in the full length integrase enzyme still remains unknown due to the low solubility and the tendency of the enzyme to aggregate. Therefore, to gain structural information, the three domains have been determined and studied in isolation by crystallographic [323] [324] [325] [326] and NMR methods [327] [328] [329] [330]. The catalytic core was the first domain to be determined and currently there are approximately 15 structures of this domain available in the Protein Data Bank [8], each containing either the F185K or F185H solubility enhancing mutation. A few more recent studies have published a two-domain structure comprising of the core and C-terminal domains [331] [332] [333]. These structures differ in the arrangement of the C-terminal domain relative to the core domain, indicating the flexibility of the linkage between the two domains. Additionally, as with the determination of the core domain in isolation, not all residues have been resolved, with missing or poorly defined loop residues.

It is thought that independently each of the domains exist as dimers in solution [334], but in the full-length integrase it is proposed to form a higher-order multimer [335], promoted by the zinc-binding site of the N-domain of the enzyme, although, there is some ambiguity concerning this matter. Integration must be consistent with the nucleotide spacing between the phosphates on the two ends of each half of the target DNA. This five nucleotide spacing requires a pair of active sites separated by $\sim$15 Å. However, in the case of a dimeric model of integrase, the spacing between the catalytic sites is $\sim$35 Å. This large distance could be accounted for by the partial unwinding of the DNA during

the reaction and coordinated but sequential cleavage at the target site [336] [337]. An alternative explanation could be the distortion of the domain and/or the DNA substrate occurring in the event of binding. However, this is thought to be unlikely due to the high hydrophobicity of the interface which is also shown to be conserved in the independently determined integrase structures [338]. A more accepted concept is the tetramisation or even higher-order oligomerisation of the dimers [323] in the full length structure, with two pairs of active sites. One pair from each dimer could be positioned within the required 15 Å in order to carry out the catalysis and the other two active sites would not play a catalytic role, but function only as structural components.

It is believed that HIV-1 IN requires at least one divalent metal ion (Mg^{2+} or Mn^{2+}) and an activated water to behave as a nucleophile to be fully functional. The divalent ions are required for both the reactions and for the formation of the HIV-1 IN complex with the viral DNA [315] [339]. However, the first crystallographic studies of the catalytic domain were carried out in the presence of cacodylate (dimethylananic acid) [323], which was critical in the enhancement of the solubility of the enzyme. It is now thought that the presence of this chemical provided a distorted representation of the active conformation, where the active site aspartate residues are not close enough to be able to bind a metal ion. Studies by Goldgur *et al.* [326] carry out the crystallisation of this domain with and without the use of cacodylate and show the structures to be significantly different, attributing this to a chemical reaction occurring between the cysteine residues of the protein and the cacodylate, preventing the binding of the divalent cation and resulting in a poorly defined loop region. Other studies carried out in the absence of the cacodylate show an improved representation of the active site and conclude the presence of one divalent cation placed between the carboxylate groups of Asp64 and Asp116 (E142 is not involved) of the active site (DDE) [340] [339] [340] [339] [341] [342] [343].

The effect of discarding this important metal ion has been additionally studied using MD simulations. Laboulais *et al.* [344] and Lins *et al.* [345] carry out simulations in explicit solvent and find similar conclusions where, in both the presence and absence of an ion, the loop remains flexible. However, significantly heightened conformational change is observed when the ion is absent and some instability in the secondary structure is noted. Additionally, a 2 ns MD study of the full-length HIV-1 IN structure [346] (constructed from two domain structures, more detail included later in this section) with and without

the presence of the Mg^{2+} ion also found increased flexibility of the catalytic domain in the absence of the ion and attribute the increased rigidity in the presence of the ion to be due to interaction between the ion and the catalytic residues.

The binding of a second metal ion in the presence of the DNA substrate has been suggested based on the similarities of the function with DNA polymerase I [347]. In the polymerase mechanism, one of the metal ions activates the water molecule to be utilised as the attacking hydroxide ion and a second metal ion assists in the stabilisation of the transition state at the phosphate to be cleaved. Using the same mechanism for HIV-1 IN, the first metal ion would bind between D64 and D116, and the second metal ion could be ligated by D116 and E152. Lins *et al.* [347] perform an MD study to investigate the effect of the binding of a second Mg^{2+} ion in the presence of a dianonic phosphate group to mimic the presence of a segment of DNA backbone. They demonstrate that the presence of the second ion initiates a change in the conformation of the active site E152 residue, which allows the carboxyl group to position itself in the vicinity of the second Mg^{2+} ion, irrespective of whether the cacodylate chemical has been used in the crystallisation process or not. Their results agree with a previous suggestion [348] that the enzyme mainly exists in an inactive conformation, which converts to its active form upon metal binding. Further to this, Lins *et al.* propose the necessity of the additional ion in the presence of DNA to stabilise the active conformation. However, in the absence of any structural data of integrase with the DNA substrate, there is currently no evidence to prove the existence of the second metal ion.

Studies of the activity of HIV-1 IN have been focused on the catalytic domain since residues 50-190 are sufficient to promote disintegration *in vitro* [317] [318], which has been shown to occur irrespective of whether or not the use of cacodylate has been used in the crystallisation process [349]. However, the structure of this domain has not yet been fully determined unambiguously by experimental techniques, especially the highly mobile loop region, which shows a high degree of disorder, which is vital for the activity of the domain. As a consequence, only a few crystal structures exist with the entire loop resolved [343] [326] [350]. Bujacz *et al.* [350] determine a crystallographic structure said to contain the active loop in an extended conformation, with E152 shown to point away from the other two catalytic carboxylates. According to the assumed roles of these residues in the binding of metal ions, the authors surmise that the conformation of the

loop they observe is not that seen during catalysis. Additionally, this structure does not include the catalytically important divalent metal ion. In the structure resolved by Maignan *et al.* [343], the location of a Mg^{2+} ion has been resolved and the loop is positioned over the active site, in a "closed" and more compact conformation, likely to correspond to an active conformation of the loop.

The loop structure overhangs the active site and although the active conformation adopted by the surface loop during the integration reaction is unknown, correlation has been found between the flexibility of the loop and enzymatic activity [351]. Mutagenesis experiments carried out by Greenwald *et al.* [351] replaced the loop hinge residues G140 and G149 (individually and as the double mutant) with alanine, resulting in reduced flexibility of the loop, demonstrated by the lower B values. The greatest reduction is seen in the double mutant and they attribute the diminished catalytic HIV-1 IN activity observed in their experimental studies as a consequence of this decreased mobility. However, the crystallographic determination of these structures was carried out in the presence of cacodylate which has shown to affect the conformational dynamics of the loop [344].

6.3.2 Role of Tyrosine 143 in Catalysis

The Tyr143 (Y143) residue is situated at the top of the loop of the core domain. Experimental studies have shown the presence of a conserved tyrosine residue in the catalytic domain, near the active site in several retroviral integrase structures [304]. Its role in the mechanism of catalysis is based on the proposed structural arrangement of the active site of E. Coli polymerase, where the residue is suggested to stabilise the activated water molecule [352]. Studies involving the mutation of this residue have resulted in alteration of the preference of the nucleophile during the 3'-processing reaction, from water (3'-processing) to alcohol (strand transfer) [353] [354], thus demonstrating its importance.

The sidechain of the Y143 residue pointing towards the active site is generally assumed to be the active orientation, and simulations carried out by Lee *et al.* [111] and DeLuca *et al.* [355] suggest that loop flexibility is required to position Y143 in close proximity to the substrate DNA when the loop is closed, thus suggesting a correlation between the function of Y143 and loop dynamics. Additionally, it has been speculated that the

mobility of the loop functions to allow Y143 to access the substrate easily and to allow the release of products.

6.3.3 Inhibition of HIV-1 IN

Blocking the HIV-1 IN enzyme is widely considered a viable target for drug development since there is no integrase activity in normal human cells and recombinant HIV virions that lack the integrase gene are not able to infect and destroy cells.

The HIV-1 IN enzyme is responsible for carrying out several distinct enzymatic steps involved in DNA integration, and blocking any one of these steps would be likely to interfere with DNA integration and as a result, prevent HIV infection. Studies indicate that infected cells die at a rapid rate which is matched by (and eventually exceeds) the rate of new infection [356] [357], therefore, reducing the rate of infection could maintain a state of homeostasis that could result in a stabilised immune system.

A large number of HIV-1 IN compounds have been reported [358] [359] [360] [361] [362] [363], with the β-keto acid class of compounds currently looking the most promising [364] but none have yet been FDA approved. The main drawback in drug development is the lack of structural information available for the enzyme.

6.3.4 Theoretical Studies of the Conformational Dynamics of the WT and Mutant HIV-1 IN

Several theoretical studies have been carried out to investigate the dynamics of the WT and mutants of the core domain of HIV-1 IN. The reported mutants have been shown to affect the activity of the enzyme, for which two mechanisms have been suggested by Greenwald *et al.* [351]. One suggestion is the alteration of the equilibrium of the different conformations required at different steps in the catalytic cycle compared with the WT. Alternatively, they suggest that the structure adopted by the constrained loop mutants may represent a non-functional conformation which is more stable.

Lee *et al.* [111] perform a number of MD and locally-enhanced sampling [365] [366] (LES) simulations of the WT and loop hinge mutants, G140A/G149A, using the AMBER2003 forcefield [13] and explicit solvent. LES simulations enable enhanced sampling

by replication of parts of the protein of interest, which in this case is the catalytic loop. The rest of the system sees the mean field of the replicated components, whilst each of the replicas do not interact with each other. This method of sampling leads to a new potential energy surface, where the energy barriers separating the local energy minima are lower, and where the global energy minimum is identical to the original.

Initially, a number of short MD simulations were performed, followed by the selection of the most interesting trajectory where the loop appeared to sample the closed conformation, which was then extended to 40 ns. Their results report the loop to move from its initial open state to a relatively stable closed state at ~8 ns, in which it remains stable for almost 30 ns. After applying LES to this closed state, the reopening of the loop is seen within ~4 ns.

In the same study, MD and LES simulations involving three mutants (G140A, G149A, and the double mutant, G140A/G149A) produced results in agreement with experimental studies of Greenwald *et al.* [351], demonstrating the hinge movement of the loop to be less prominent in the structures containing the single mutants, and completely eliminated in the structure possessing the double mutant and thus reach the same conclusion of Greenwald *et al.* They suggest that the opening/closing ability of the loop is vital for catalytic activity and the mutants studied hinder this loop mobility, thus affecting the activity of the enzyme. Additionally, they suggest a correlation between catalytic loop conformation and the dynamics of the proposed catalytically important Y143 residue. In the mutant enzymes, they find the sidechain to predominantly point towards the active site and speculate the function of the loop to position this residue in the functional form of the enzyme.

The 1QS4 WT crystal structure [367] of HIV-1 IN used in the study by Lee *et al.* contains the Mg^{2+} ion coordinated by the D64 and D116 catalytic residues. The mutant HIV-1 IN structures were created through the mutation of the appropriate residues using the SwissPDB software [368]. This structure has two unresolved loop residues (I141 and P142), which have been modelled in, again using the SwissPDB software [368] from chain B of 1BIS [326] (pdb code) crystal structure, which contains these missing residues. However, inspection of this starting structure, reveals a *cis* amide bond between residues G140 and I141. This is not likely to be a true conformation of this peptide bond since

the 1BIS crystal structure [326], as well as the other crystal structures of the enzyme, do not show this energetically unfavourable state of the prolyl bond between these residues.

Barreca *et al.* [369] carry out MD studies (1.75 ns each) on the WT and a drug-resistant double mutant (T66I/M154I) forms of an inhibitor-bound (5CITEP) HIV-1 IN in explicit solvent. The inhibitor is shown to constrain the loop activity of the WT HIV-1 IN, whereas the simulation of the mutant inhibitor-bound structures show an increased mobility of the catalytic loop compared with the bound WT structure, thus showing the double-mutant to restore the high mobility of the loop, comparable to that seen in the apo form. Differences in the dynamics of the Tyr143 residue is also noted between the WT- and mutant-bound structures. In the WT, where the loop is shown to be more rigid, the Y143 residue is shown to move away from the active site, and is thought to assume an inactive conformation, whereas the distance between the active site and Y143 remains approximately constant in simulations of the bound mutant, with the sidechain pointing towards the active site.

Brigo *et al.* [370] use conventional MD to study the effect of the T66I/M154I mutation on the dynamics of apo HIV-1 IN, compared with the WT HIV-1 IN. This double mutation has been reported to confer resistance to inhibitors with a diketo functionality. Four MD simulations of 2 ns in length, using four different starting structures (obtained from MD trajectory generated in the studies of Barreca *et al.* [369]) were performed on the T66I/M154I HIV-1 IN. They compare the RMSF values from these simulations with previous work carried out on the apo WT HIV-1 IN [347] [371] and inhibitor-bound WT and mutant HIV-1 IN [369]. They find the flexibility of the double mutant to be the same as observed in the apo WT and T66I/M154I inhibitor-bound forms of the HIV-1 enzyme, demonstrating that the mobility of the loop is not influenced by the mutations in the apo form of the HIV-1 IN. This agrees well with the idea that a mechanism of inhibition of this enzyme is to hinder the flexibility of the catalytic loop.

MD studies carried out by Brigo *et al.* [372] also investigated the conformational dynamics of the WT and T66I/M154I double mutant of the integrase enzyme, but bound to a different inhibitor, L-731,988. The mutant is reported to decrease the catalytic activity of the enzyme and to resist inhibition of drugs of this class. A selection of conformations were taken from prior MD performed on the WT [369] and mutant [370]

HIV-1 IN (include Mg^{2+} ion), and docking of the inhibitor was carried out. The HIV-1 IN structures with the most favourable inhibitor conformations were selected and used as the starting structures for 2 ns MD simulations.

Simulations of the WT inhibitor-bound form of HIV-1 IN show the catalytic loop to have decreased mobility compared with the T66I/M154I mutant inhibitor-bound HIV-IN simulations. Additionally, the inhibitor is found to assume different conformations in the mutant and WT forms, with the inhibitor in the mutant HIV-1 IN shown to be buried deeper into the active site compared with the WT.

Comparison with results of a prior study of the apo T66I/M154I HIV-1 IN [370] demonstrate the presence of the inhibitor does not affect the mobility of the loop of the mutant, as both the inhibitor-bound and apo-forms of the mutant HIV-1 IN possess the same mobility. In the simulations of the WT inhibitor-bound HIV-1 IN, direct interaction between the G148 loop residue and the inhibitor is observed, which is not seen in the simulations of the mutant inhibitor-bound HIV-1 IN, as this residue is in a different orientation, making this hydrogen-bond interaction impossible. This is postulated to consequently affect the conformation of the Y143 residue. In simulations of the inhibitor-bound WT HIV-1 IN, the sidechain of Y143 points towards the active site, whereas, in the mutant forms of both the apo and inhibitor-bound HIV-1 IN, the sidechain of Y143 points away from the active site. These results indicate the involvement of the catalytic loop and the Y143 residue to be fundamental to the catalytic mechanism.

De Luca *et al.* [355] used MD simulations to study the full-length HIV-1 IN dimer (1WKN with subunits A and B). This structure is a computational model assembled from the experimentally determined structures of the single domains [373], complexed with viral DNA. The construction of the model started with the core domain (1QS4) [367], which contains the metal ion, but lacks residues of the loop, which were modelled in from the 1BIS crystal structure [326]. The solubility enhancing residues, F185K and W131E were mutated back to their original WT form. A second metal ion has been modelled in using the resolved location of the second ion in the two-metal ion structure of the Avian Sarcoma Virus integrase. Following this, the C-domain (PDB code 1EX4 [374]) was superimposed onto the isolated core domain and linked via residues 209 and 210. After each domain adding stage, minimisation of the structure was carried out.

Finally, the N-terminal was added using the AB dimer of the 1K6Y [375] structure, matching common atoms in the catalytic domain, and missing atoms in the N-terminal domain added in from the 1WJD [376] crystal structure.

Results of 1.5 ns MD simulation of this structure show that only subunit B of the dimer is in direct contact with the viral DNA. A substantial increase in α-helix is observed in regions around the loop structure in subunit B (residues Asn120 - Phe139 and Asn150-Glu170), providing evidence that the opening and closing movement of the loop is a result of the conformational changes in the secondary structure in the regions around the loop, rather than the loop itself. The authors also postulate the importance of the Y143 residue in the catalytic mechanism, observing its change in side chain orientation towards the active site in the presence of the viral DNA substrate.

An MD study by Wijitkosoom *et al.* [346] also studies the dynamics of full-length HIV-1 IN, constructed from the two experimental structures, one of the core domain connected to the N-terminal (1K6Y) and another where the core domain is connected to the C-terminal (1EX4). Any missing residues were modelled in and two cations are included (Mg^{2+} and Zn^{2+}). A comparison is made between this structure and the full-length HIV-1 IN structures of De Luca *et al.* [373] and Podtelezhnikov *et al.* [377]. Although the overall positioning of the three domains is comparable between the three structures, the active site is considerably different compared with the Podtelezhnikov model. Distance measurements between the three active site residues was found to be significantly different and the orientations of the catalytic residue were found to vary. These highlighted differences would impact drastically on the interactions between the ligand/enzyme interactions.

In the same study, 2 ns MD simulations were carried out on the generated full-length structures in explicit solvent in the presence and absence of cations. They find the full-length integrase enzyme possesses increased mobility in the absence of the ions, and the conformation of the active site to differ, with the Mg^{2+} binding to the active site, decreasing the flexibility of the active site. Comparison of distances between the catalytic residues for the full-length HIV-1 IN structures, generated by the different methods shows the structure by Podtelezhnikov *et al.* [377] to be least comparable to their own, which maybe a result of the absent metal ions.

Additionally, the authors demonstrate the size and flexibility of the active site cavity to differ between the two-domain structures and the full-length structure, with an increase in flexibility detected in the two-domain structures. This same observation is made concerning the flexibility of the whole core domain, as the mobility of this domain is likely to be dampened in the full-length structure by the presence of both terminal domains.

Lins *et al.* [347] carried out two 500 ps MD simulations involving the core domain of HIV-1 IN with two magnesium ions and a dianionic phosphate group in the active site to mimic the presence of a segment of DNA substrate. They concluded their results to be in agreement with the hypothesis that the catalytically competent form of the enzyme only occurs in the presence of both metal ions, since conformational change of the active site was only observed in the presence of both ions. The assumed active conformation is thought to be stabilised by the phosphate group, suggesting that the active conformation would only arise in the presence of the substrate. The starting coordinates were obtained from the 1BIS crystal structure [326], which contains one of the ions, the second placed between D64 and D116, according to comparison with the ASV IN structures.

Summary

Some ambiguity concerning the structure of HIV-1 IN domains exists owing to the difficulties in obtaining structural information. The majority of crystal structures of HIV-1 IN have been resolved for the core domain, although a few two-domain crystal structures exist. A few computational studies have constructed the full-length HIV-1 IN, but these structures are modelled, simulations performed on them out are short, (only a few ns in length) and comparison of these structures has revealed major differences between them. Therefore, there is significant uncertainty as to the validity of any conclusions made as a result of the studies.

The vast majority of studies have focused on the core domain, since this domain has the ability to initiate the process of disintegration (the excision of viral DNA) *in vitro* [317] [318] in the absence of the two terminal C- and N-terminal domains. The core domain possesses an catalytic loop, the conformation of which has been demonstrated to play a fundamental role in the catalytic mechanism of the protein. However, resolving the

loop in crystal structures has proved to be difficult owing to its high mobility and as a consequence, the active conformation of the loop is not yet accurately known.

Several mutations of this domain have been demonstrated to affect the dynamics of the catalytic loop [369] [372] [370] [111] [351], thus affecting inhibition or the catalytic activity. From these studies, it has also been suggested that the orientation of the sidechain of Y143 plays a fundamental role in positioning the ligand in the active site.

Two mutant forms have mainly been studied, T66I/M54I and G140A/G149A. The former mutation diminishes the efficacy of the inhibitors by restoring the mobility of the loop, which is lost on binding to the inhibitor in the WT and affects the orientation of the Y143 residue.

The G140A/G149A mutant HIV-1 IN has been shown by an experimental [351] and theoretical [111] study to diminish the catalytic activity of the apo-enzyme through the rigidifying of the loop. This reduction in loop mobility prevents the loop assuming active conformations and prevents the Y143 residue approaching the active site. However, the catalytic loop of the starting structure used by Lee *et al.* in the MD study possesses an unlikely *cis* conformation of a peptide bond involving one of the two missing loop residues modelled-in by the authors. Additionally, the conclusions of the experimental study are based B-values of structures crystallised with the aid of cacodylate, a chemical known to cause deformities of the active site, prevent metal ions binding and affect the dynamics of the loop. Therefore, there is some doubt to the reliability of the results and conclusions reached in these studies.

6.4 Conclusions

An increased understanding of mutation and their effects on dynamics of the HIV-1 PR and HIV-1 IN enzymes would aid understanding of their mechanism, and assist greatly in the development of drugs to inhibit these enzymes. Therefore, in an attempt to gain a better insight into the conformational dynamics of the WT and mutant HIV-1 PR and HIV-1 enzymes, the enhanced sampling technique, RDFMD has been used to enhance the conformational sampling of these enzymes, and the results are shown in chapters 7 and 9.

Previous studies, as described in this chapter have highlighted the deficiencies of using conventional MD, showing simulations to become trapped and sample only a small proportion of conformational phase space. RDFMD is an efficient technique which enhances conformational change by amplifying the low-frequency motions (method described in chapter 4). However, this non-equilibrium method does have some disadvantages. For RDFMD simulations to be carried out, there must be some prior knowledge of the system, as the digital filter is applied to a user-defined specific structural region of the protein. Additionally, some initial system-dependent parameter optimisation must be performed. However, previous studies using RDFMD on protein systems have demonstrated this technique to induce large reversibly sampled biologically relevant conformational changes, not seen in MD through targeting just a few residues [133].

Chapter 7

Human Immunodeficiency Virus-1 Protease (HIV-1 PR): Part 1: Apo

7.1 Aims

The aims of this chapter were to study the conformational dynamics of the WT and mutant HIV-1 PR enzymes using MD and the enhanced sampling technique of RDFMD. The flap structures of the enzymes have been shown to be crucial in the mechanism, but there role is not yet fully understood.

The mutants M46I and the double mutant, V82F/I84V have been chosen for study since experimental studies show them to be commonly occurring mutations which affect the binding of inhibitors. Theoretical studies have predicted that the presence of these mutants cause a deviation from the open/closed equilibrium of the dynamics of the flaps of the WT enzyme, thus affecting the ability of inhibitors to bind in the active site.

7.2 System Preparation and MD Parameters for Simulation of apo HIV-1 PR

The starting structure 1HHP [185], an apo-form of the wild type HIV-1 PR, was taken from the Protein Data Bank [8]. The coordinates of the given monomer were translated (x, y, z) to (x, y, -z) to produce the dimer. The WHATIF [171] program was used to add polar hydrogens and to check the structure. The structure possesses two histidine residues, both of which were doubly protonated, since simulations have been carried out at the pH of the active protease enzyme, pH3-4. The AMBER utility XLEAP [172] was used to add other hydrogen atoms, and to solvate the system, with a minimum distance of 12 Å from the protein, in a box of 9370 TIP3P [44] water molecules. Seven chloride counterions were added to the monoprotonated system to neutralise the overall charge of the system.

All simulations, unless otherwise stated, have been carried out using the NAMD [170] molecular dynamics package and the AMBER99 [27] forcefield.

The minimisation was carried out in stages, starting with the protein only (5,000 steps), followed by solvent (30,000 steps), ions (1,000 steps), solvent and ions (20,000 steps) and finally the entire system (40,000 steps), giving a total of 96,000 steps. Two minimisation algorithms have been used, initially the steepest descent algorithm, followed by the conjugate gradient method.

The minimised system was heated to 300 K in the NVT ensemble. The procedure employed a Langevin thermostat with a 10 ps^{-1} damping parameter and a 2 fs timestep. The heating was carried out gradually in stages at 50 K intervals, each interval being 20,000 steps long.

Equilibration simulations, using a Nosé-Hoover barostat in the NPT ensemble were carried out for 100 ps , with a target pressure of 1 atm. A decay parameter of 100 fs and a piston period of 200 fs were used. A further 2 ps was run, with a decay parameter of 300 fs and a piston period of 500 fs.

The final equilibrated monoprotonated systems had the following box dimensions : apo WT: 61.08, 60.98 and 82.37 Å, apo M46I: 61.6, 61.5, 83.1 Å, apo V82F/I84V: 61.7, 61.6, 83.2 Å.

All production MD simulations were run in the NVT ensemble. The NVT ensemble was chosen in preference to the NPT ensemble for consistency with the RDFMD simulations, which employ this ensemble to avoid system expansion. The MD simulations used a Langevin thermostat with a 5 ps^{-1} damping parameter at a temperature of 300 K and a timestep of 2 fs. Periodic boundary conditions were used, along with a particle mesh Ewald treatment of electrostatic interactions, using a interpolation order of 6, and switching function applied to the Lennard-Jones interactions between 9 Å and the 10.5 Å cutoff. SHAKE [35] was applied to all bonds containing hydrogen, using a tolerance of 10^{-8} Å.

Owing to the absence of apo semi-open crystal structures of the two mutants studied (M46I and V82F/I84V), the equilibrated WT apo structure was taken and the appropriate residues mutated using scap [378] [379]. The appropriate residues have been mutated in both monomers of the enzyme. Following residue mutations, minimisation of these altered residues were carried out for 1,000 steps whilst restraining the rest of the system, followed by the minimisation of the rest of the protein for 5,000 steps, the solvent and counterion for 1000 steps, and lastly, the entire system for 10,000 steps. Heating and equilibration of the system was then carried out in the same manner as for the apo WT structure.

For each system, two MD production simulations have been carried out, one of 20 ns and the other 10 ns in length and each starting from different velocities and coordinates, generated by two sets of equilibration of equal length being carried out, starting from the same minimised structure.

7.2.1 RDFMD Simulation Parameters

The flexibility of the flaps, and their role in substrate entry is the focus of this study. Therefore, flap-tip residues 49-51 and 149-151 have been selected as the target of the filter of RDFMD.

A protocol for parameters to use on regions of interest in proteins has been optimised previously [231], and the parameters are used here. These include a 0-100 cm^{-1} digital filter, with 201 coefficients, an amplification factor of 2, and 4 ps of MD run in between the filter applications. The final trajectory is generated by piecing together the individual 4 ps MD runs. Each RDFMD simulation produces 100 4 ps MD sections, totalling 400 ps.

The temperature cap used in these simulations is different from those used previously [231]. In the prior work on this system and other protein systems studied using the RDFMD technique [7], temperature caps in the range of 900 K - 1500 K were employed and an amplification factor of 2. Owing to the change in the forcefield used in this work, from CHARMM [15] to AMBER [27], the temperature cap has been decreased and the amplification factor also altered. This is due to the AMBER forcefield appearing to be more temperature sensitive to *trans-cis* ω transitions, which were shown to occur at lower temperatures than is seen when studying the same system using the CHARMM forcefield. Therefore, a temperature cap of 800 K has been used, with a range of amplification factors. Both the temperature cap and amplification factors were determined by trial and error.

For each system, 12 RDFMD simulations have been carried out, six of which use a filter delay of 50 steps, and six using a 100 step filter delay. Each simulation employed a different combination of amplification factor and filter delay to increase the variation in the trajectory. For those simulations using a 50 step filter delay, amplification factors of 1.6, 1.8 and 2.0 were used. RDFMD simulations which employed a filter delay of 100 steps used amplification factors of 2.0, 2.2 and 2.4. In addition, two different starting structures were used, which were the same as those generated for the MD simulations.

The MD sections of the RDFMD method were carried out in the NVT ensemble at 300 K, using a Langevin thermostat with a 5 ps^{-1} damping parameter and a 2 fs timestep.

7.3 Results

This study has investigated the conformational dynamics of the HIV-1 PR WT and M46I and V82F/I84V mutant systems. The selected mutants have been speculated to affect the conformations of the enzyme, especially in the region of the flaps. For each apo-system, twelve RDFMD simulations, using a temperature cap of 800 K (table 7.1) and two MD simulations (MD1: 20 ns, MD2: 10 ns) have been carried out, revealing a range of semi-open, open and flap-tip curling conformations of the flaps.

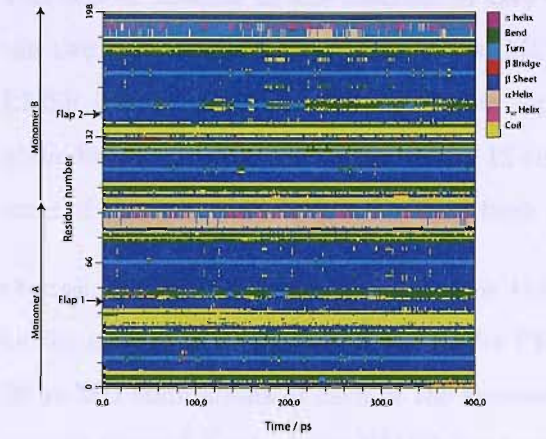
RDFMD Simulations		
Starting Structure No.	Filter Delay	Amplification
1	50	1.6
2	50	1.6
1	50	1.8
2	50	1.8
1	50	2.0
2	50	2.0
1	100	2.0
2	100	2.0
1	100	2.2
2	100	2.2
1	100	2.4
2	100	2.4

TABLE 7.1: RDFMD Simulations: parameters used for simulations of each of the apo-systems (WT, M46I, V82F/I84V)

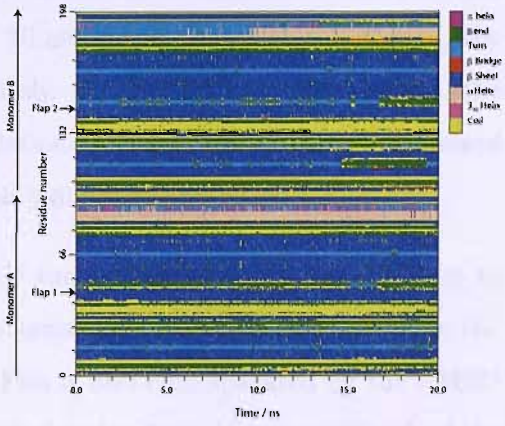
7.3.1 Structural Stability and Flexibility

The secondary structure of the HIV-1 PR WT and mutant enzymes studied have been monitored over the length of the MD and RDFMD simulations (figure 7.1). For each system, the secondary structure plot resulting from the 20 ns MD simulation, and the 400 ps trajectory of a representative RDFMD simulation have been presented. The secondary structure plots for the other RDFMD and MD simulations are comparable to those here and so have not been included.

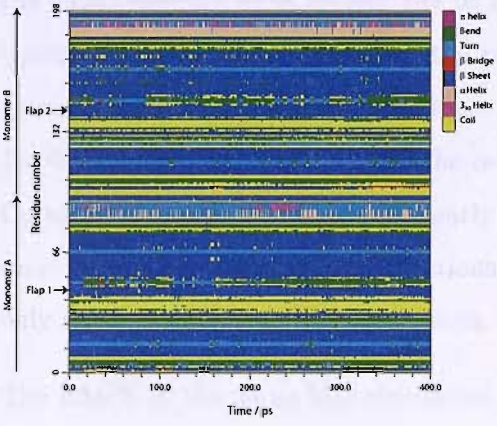
These plots show the HIV-1 PR enzyme to be structurally stable over the duration of the simulations. Comparison of the secondary structure over 20 ns of MD simulation with 400 ps of RDFMD simulation demonstrates that the addition of energy into each of the apo-systems in the filter application stages of the RDFMD method has not caused any substantial change in the secondary structure. In addition to analysing the secondary structure, monitoring for any unfeasible *trans-cis* isomerisation has been carried out throughout the duration of the simulations. None of the simulations reported here observed *trans-cis* isomerisation, which corroborates the work of Hornak *et al.*[90], which refutes the requirement of the isomerism in the flap-tips, which is stated to occur in the simulations of Hamelberg *et al.*[229]. It must be noted however, the temperature cap of 800 K used in the RDFMD simulations presented in this chapter was specifically selected to avoid this event.



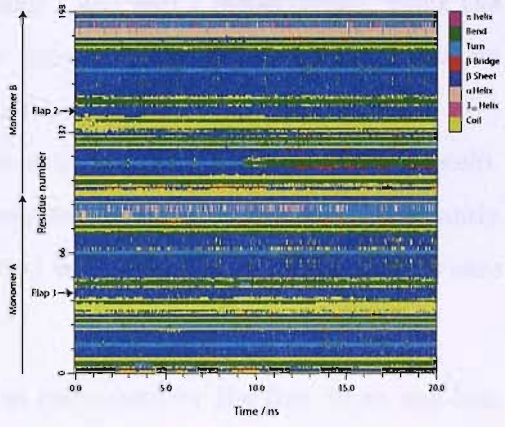
(a) RDFMD: WT



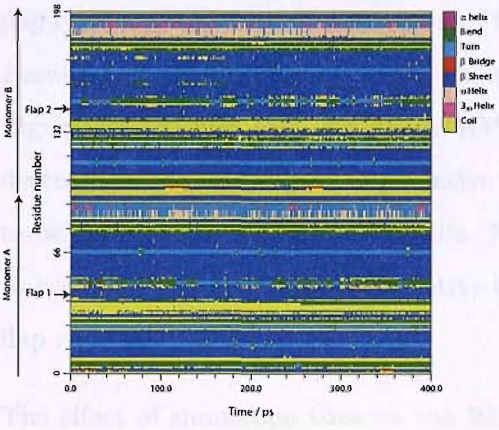
(b) MD:WT



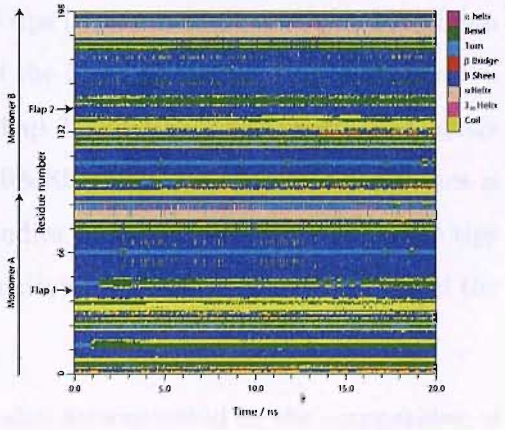
(c) RDFMD: M46I



(d) MD: M46I



(e) RDFMD: V82F/I84V



(f) MD: V82F/I84V

FIGURE 7.1: Representative secondary structure plots for the apo HIV-1 PR structures. Parameters of plots generated by RDFMD: filter delay of 50 steps, temperature cap of 800 K, amplification of 1.6. Plots generated by MD simulation have been produced from 20 ns of simulation.

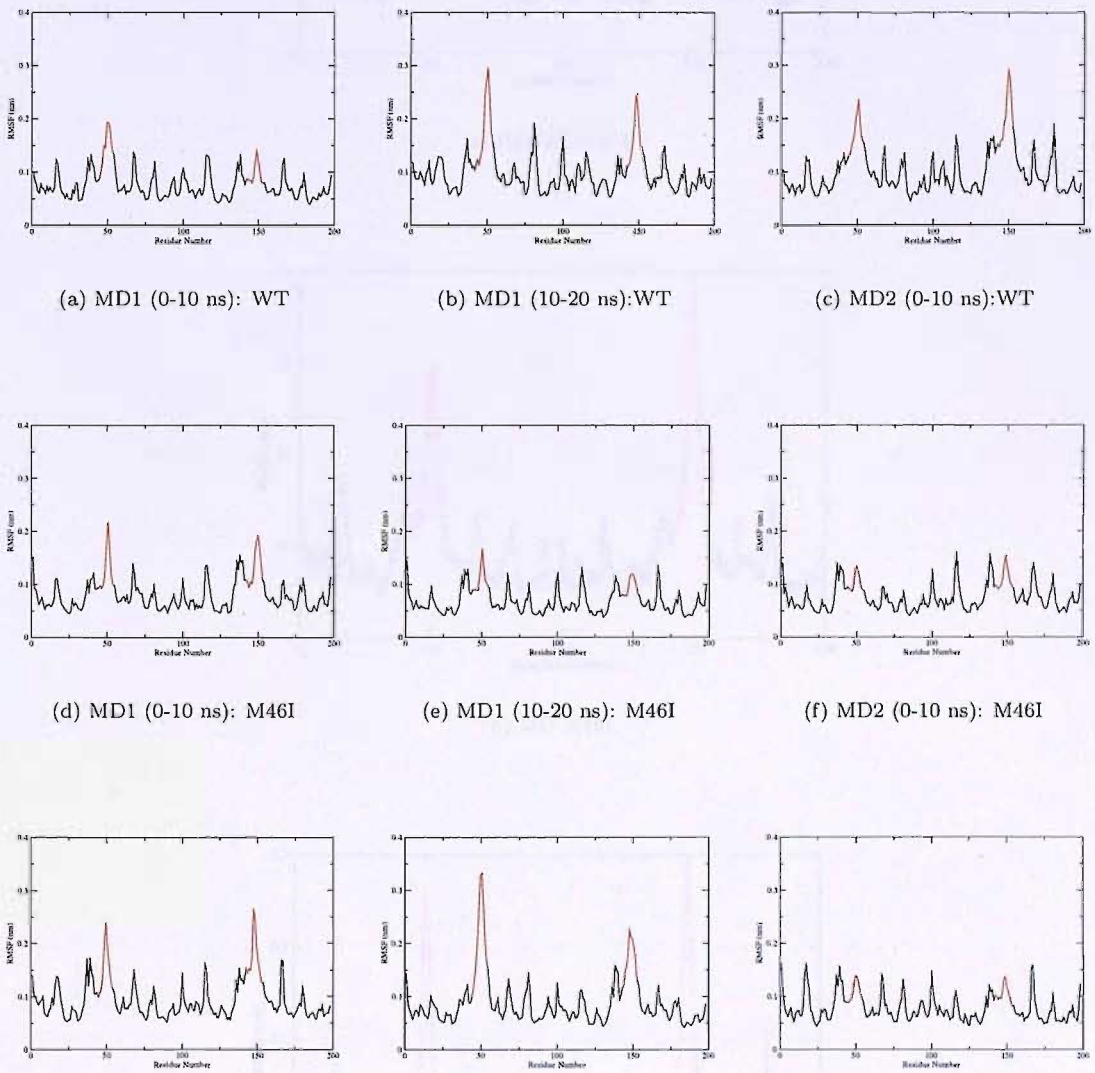
The RMSF of each of the residues of HIV-1 PR is shown in figures 7.2 and 7.3. For the two MD simulations (MD1: 20 ns, MD2: 10 ns) carried out for each system, the RMSF is calculated for each simulation separately. For the RDFMD simulations, the plots display the average RMSF of the 12 simulations for each system and the standard error of the mean has been calculated from each individual simulation.

Overall, the RMSF values generated by the MD simulations show the flap residues to be the most flexible regions of the HIV-1 PR systems, which is especially evident in the 20 ns MD simulations of each of the systems. This is also demonstrated by the RMSD plots in figure 7.4, where the RMSD has been calculated over the duration of each of the MD simulations. Plots 7.4 (a) and (b) display the RMSD of the α-carbons of each of the flaps (residues 44 to 54 and 144 to 154 for flap 1 and flap 2 respectively), using the equilibrated starting structure as the reference structure, and fitting on the α-carbon atoms of the rest of the protein (residues 1 to 43, 55 to 143 and 155 to 198). Figure 7.4 (c) displays the RMSD over time for the α-carbons of the non-flap residues of the protein. Collectively, these RMSD plots clearly demonstrate the flaps to deviate significantly more from their starting configurations compared with the rest of the protein, where only small fluctuations in RMSD occur.

The RMSF of the 20 ns MD simulation has been calculated for the first 10 ns and last 10 ns separately (figure 7.2), where the results are shown to vary. Comparison of the first and last parts of the simulation demonstrate a general increase in the RMSF of the flap residues, especially in the region of the flap-tips (residues 49-52 and 149 -152 of flap 1 and flap 2 respectively), with the exception of the M46I mutant 20 ns MD simulation (figure 7.4 (d) and (e)), where the RMSF of flap 1 and flap 2 are shown to slightly decrease in the last 10 ns. The relative larger RMSF values for the flap-tip residues is to be expected given the NMR results. NMR studies [192] report motion of the flap tips to occur on a shorter timescale relative to the motion involving a greater number of the flap residues.

The effect of simulation time on the RMSF is also demonstrated in the comparison of MD1 and MD2 simulations of each of the apo-systems (figure 7.2), where a reduction in the flexibility of the flap residues for the shorter 10 ns simulation is observed (with the exception of flap 2 of the WT, which is shown to be more flexible in MD2). This is

also evident in the RMSD calculations, where the highest RMSD values for the flaps are generally observed in the second half of the 20 ns simulation (with the exception of flap 1 of the M46I simulation, where the highest RMSD occurs within the first 10 ns). This result indicates that the flaps are highly flexible and longer timescales are required to capture the complete dynamics. Therefore, to simulate the dynamics of the flaps more efficiently, the enhanced sampling technique of RDFMD has been used.

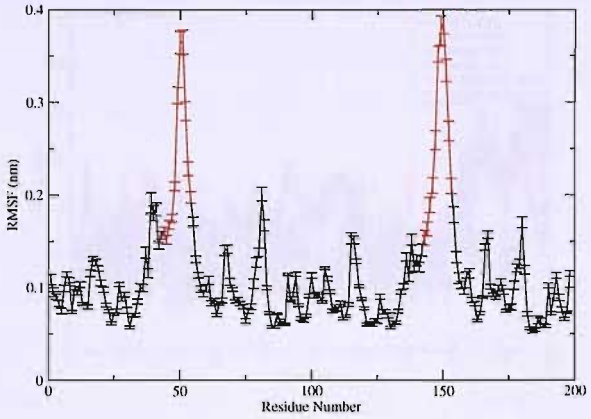


(a) MD1 (0-10 ns): WT (b) MD1 (10-20 ns):WT (c) MD2 (0-10 ns):WT

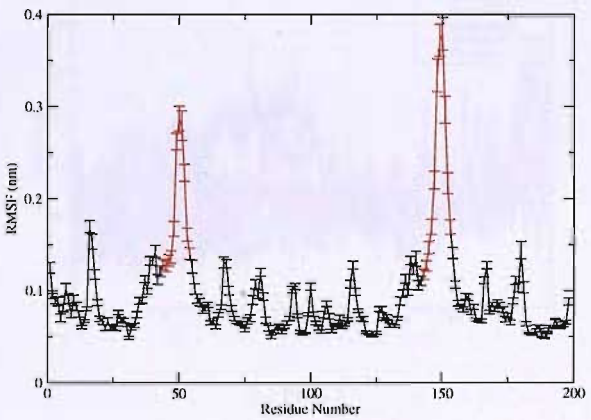
(d) MD1 (0-10 ns): M46I (e) MD1 (10-20 ns): M46I (f) MD2 (0-10 ns): M46I

(g) MD1 (0-10 ns): V82F/I84V (h) MD1 (10-20 ns): V82F/I84V (i) MD2 (0-10 ns): V82F/I84V

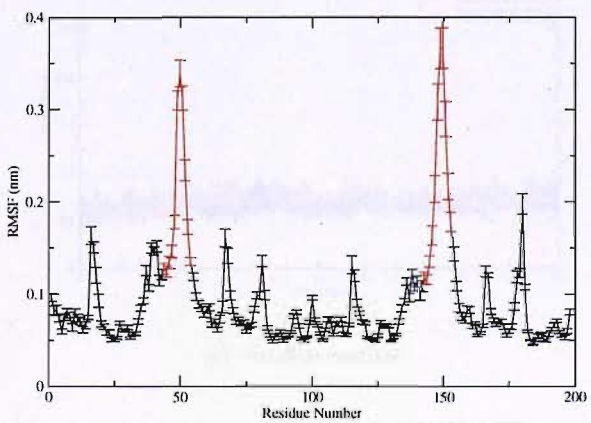
FIGURE 7.2: RMSF of α-carbon atoms of all residues of HIV-1 PR calculated for the two MD simulations of the WT and mutant apo-systems. The flap residues are marked in red.



(a) RDFMD: WT

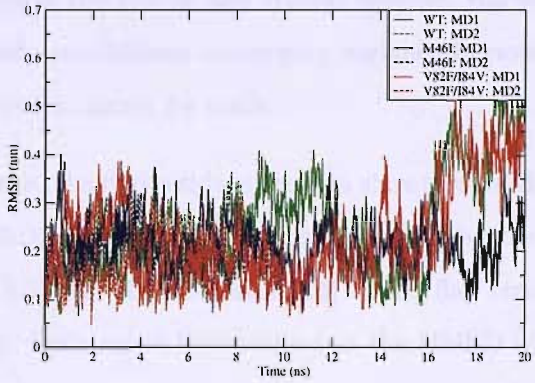


(b) MD:M46I

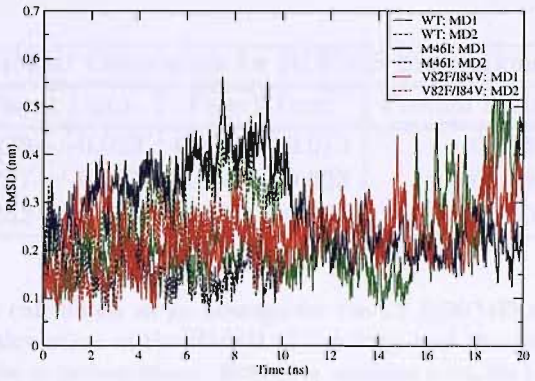


(c) RDFMD: V82F/I84V

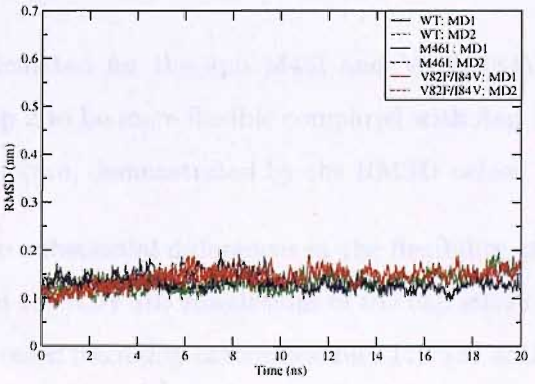
FIGURE 7.3: RMSF of α-carbon atoms of residues of HIV-1 PR calculated for WT and mutant RDFMD simulations, using the equilibrated structure as the reference. RMSF values calculated as an average for all simulations, with the standard error calculated from each individual simulation marked. The flap residues are marked in red.



(a) Flap 1



(b) Flap 2



(c) Non-flap residues

FIGURE 7.4: RMSD of α-carbon atoms of residues of HIV-1 PR calculated for WT and mutant MD simulations (MD1: 20 ns, MD2: 10 ns), using the equilibrated structure as the reference. Figure (a) RMSD of flap 1, fitting on non-flap residues, (b) RMSD of flap 2, fitting on non-flap residues, (c) RMSD of non-flap residues, fitting on non-flap residues. All calculations used the equilibrated starting structure as the reference structure.

Owing to the variability in the RMSF and RMSD between the two MD simulations for the same system, definite conclusions concerning variance between the flexibility of the WT and mutant structures cannot be made.

As in the MD simulations, the flap residues are also shown to be the most flexible regions of the protein in the RDFMD simulations, demonstrated by the relative larger values for the RMSF (figure 7.3) and RMSD (table 7.2) of the flap structures compared with the rest of the protein. Only small fluctuations in the RMSD of the non-flap residues occur, showing that the protein systems remained stable throughout the duration of the RDFMD simulations, supporting the secondary structure information.

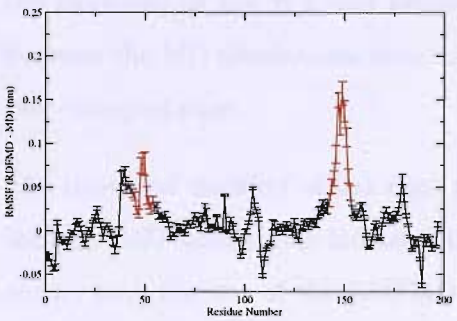
RMSD Calculation for RDFMD Simulations			
System	Flap 1 (nm)	Flap 2 (nm)	Protein excluding flaps (nm)
WT	0.279+/-0.023	0.273+/-0.013	0.128+/-0.004
M46I	0.217+/-0.017	0.256+/-0.023	0.128+/-0.007
V82F/I84V	0.242+/-0.013	0.295+/-0.025	0.130+/-0.002

TABLE 7.2: RMSD calculated as an average for the 12 RDFMD simulations for each apo system. The calculation of the RMSD of the flaps and the protein excluding the flaps was based on the α-carbon atoms, fitting on residues 1-43, 55-143 and 15-198 (non-flap residues,) using the two equilibrated starting structures as the reference structures. The standard error of the mean is calculated from the individual simulations.

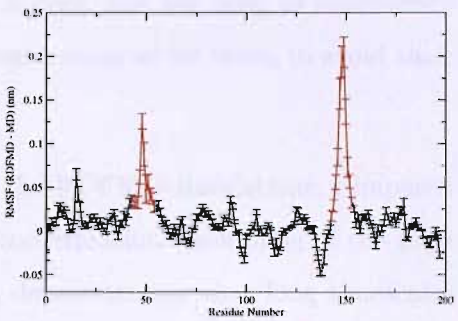
The RMSF values calculated for the apo M46I and V82F/I84V HIV-1 PR RDFMD simulations suggest flap 2 to be more flexible compared with flap 1, and to deviate more from the reference structure, demonstrated by the RMSD values.

As figure 7.3 shows, no substantial differences in the flexibility of the WT and mutant systems are observed in the RDFMD simulations of the flap structures, although there is some indication of increased flexibility of loop residues 179-181 of the V82F/I84V mutant HIV-1 PR compared with the WT and M46I mutant and reduced flexibility of the flap tip residues of flap 1 of the M46I mutant, compared with the WT and V82F/I84V apo simulations.

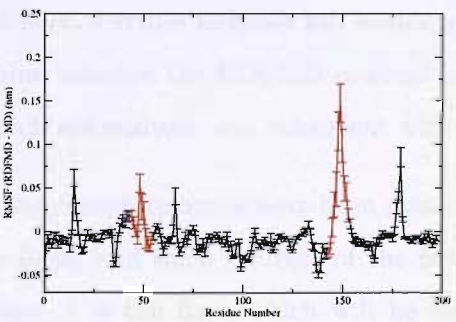
Owing to the number of RDFMD simulations carried out and the complexity of their motion over the duration of the simulation, comparison between the RMSD values obtained for MD and RDFMD have not been made. Comparison of the mobility of the systems using MD and RDFMD methodology has been made using RMSF. Figure 7.5 displays the difference between the average RMSF value calculated over the 12 RDFMD simulations of each system, with the values obtained for the complete 20 ns MD simulation. In each case, there is a positive difference for the flap residues, especially for flap 2, demonstrating that the RDFMD method consistently enhances the conformational dynamics of the flaps.



(a) WT



(b) M46I



(c) V82F/I84V

FIGURE 7.5: The difference ($RMSF_{RDFMD}$ - $RMSF_{MD}$) between the RMSF calculated for 20 ns of MD simulation and the average RMSF value, calculated over the 12 RDFMD simulations for each of the apo-systems. The flap residues are marked in red.

Summary

The analysis carried out in this section demonstrates the HIV-1 PR WT and mutant systems studied to be stable over the duration of the MD and RDFMD simulations, given the stable secondary structure and low RMSD values of the non-flap residues.

RMSD and RMSF analysis of the MD and RDFMD simulations demonstrates the flaps of these protein systems to be highly mobile, especially in the flap-tip region, and to show significantly more motion relative to the rest of the structure. Comparison of the data between the WT and mutants has not resulted in any conclusion regarding the flexibility of the WT and mutant structures. Indeed, the low level of consistency between the MD simulations demonstrates that great care must be taken to avoid their over-interpretation.

The increased mobility of the flaps resulting from the RDFMD simulations, compared with the MD simulations indicates that increased conformational sampling of the flaps occurs with the use of the RDFMD methodology, demonstrating that long timescales may be required to capture the dynamics of the flaps, which is not achieved with the MD simulations here. This is also indicated in the variability between the 10 and 20 ns MD simulations presented here. Further analysis has been carried out and is presented in this chapter, to determine whether the RDFMD method produces conformations of the HIV-1 PR systems which are realistic and consistent with experimental results.

Since in previous studies, the chosen mutants have been postulated to affect the conformational dynamics of the flaps, and since the rest of the protein undergoes relatively little conformational change, it is the flaps which will be the focus of study for this chapter. The majority of analysis has been carried out on flap 2, since the RMSD and RMSF values, in addition to visualisation of the trajectories, do not show flap 1 to undergo conformational change to the extent seen in flap 2. This may be a consequence of the protonation of the catalytic aspartate of monomer B (residue 125) in the active site, whereas the catalytic aspartic acid of monomer A (residue 25) remains in a neutral form.

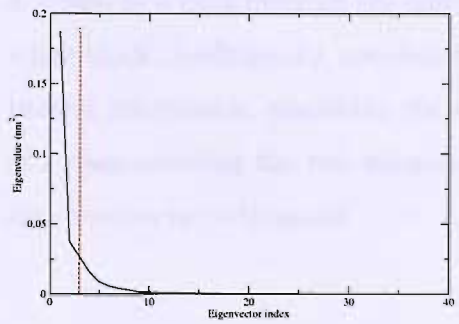
7.3.2 Principal Component Analysis

PCA has been used to study the dynamics of the flaps of each of the monomers of the HIV-1 PR enzyme individually (residues 44-54 - flap 1 and 144-154 - flap 2). The dynamics of the rest of the protein has been disregarded for this analysis to avoid the flap motions of interest being obscured by the motions of the rest of the protein.

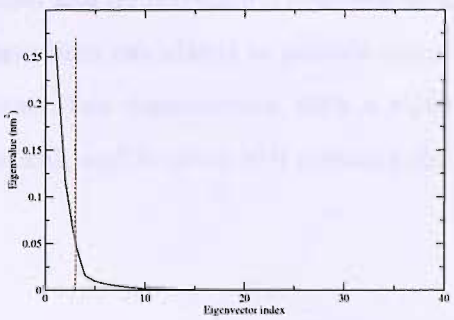
MD Simulations

PCA has been carried out on each of the two trajectories generated by the MD simulations. In each case, the covariance matrix and resulting eigenvectors and eigenvalues have been generated using the α-carbon atom trajectories of flap 2 (residues 144-154), whilst fitted onto the α-carbon atoms of the non-flap residues (residues 1-43, 55-143, 155-198).

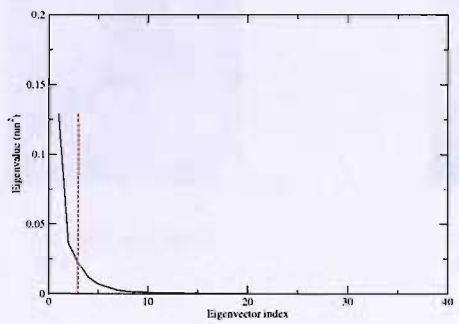
Ordering of the eigenvalues according to their contribution to the total motion of the protein is shown in the scree plot in figure 7.6 for MD1 and MD2 of each of the apo-systems. For proteins in general, literature reports the first few eigenvectors to have large eigenvalues [380], [381] [382], so the overall motion of the protein can be described using just a few eigenvectors. This is also indicated by the scree plots of the HIV-1 PR systems studied here; approximately 80 % of the total motion is represented in the first three eigenvectors for MD1 and MD2 of each of the apo-systems. Thus, the motion represented in these three eigenvectors has been studied in more detail.



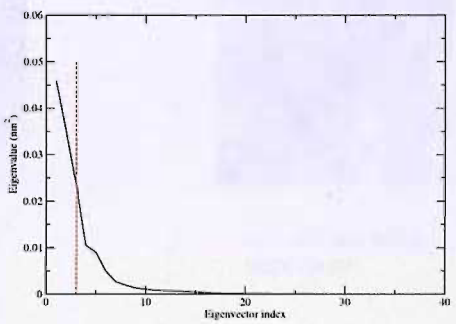
(a) WT: MD1



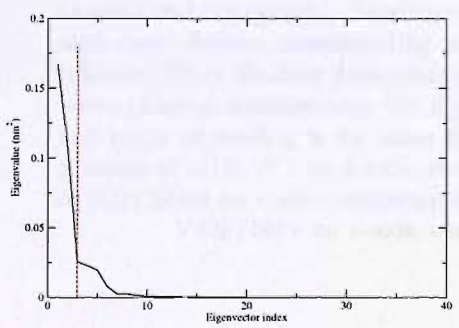
(b) WT: MD2



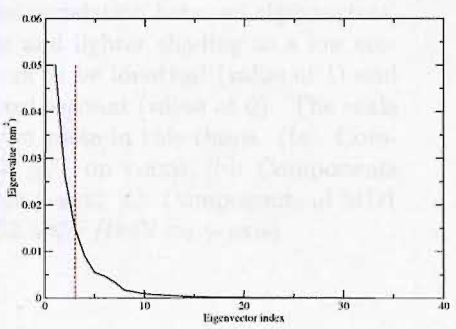
(c) M46I: MD1



(d) M46I: MD2



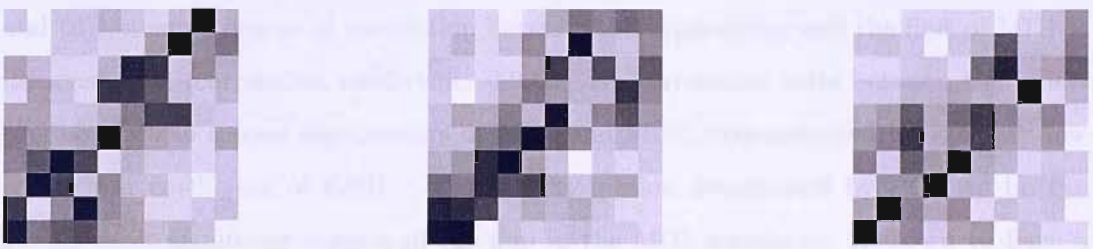
(e) V82F/I84V: MD1



(f) V82F/I84V: MD2

FIGURE 7.6: Scree plot depicting the proportion of motion captured in each of the eigenvectors of flap 2 of MD1 and MD2 of HIV-1 PR WT and mutant apo-systems. The red line depicts the eigenvalues of the first three eigenvectors.

Comparison of the inner-products of the first 10 eigenvectors of flap 2 (figure 7.7) gives the level of correlation between the major motions seen in each of the eigenvectors between the MD simulations of each system. Perfect correlation between two eigenvectors is shown by a dark block on the inner-product matrices and no correlation is shown by a white block. Additionally, correlation coefficients have been calculated to provide quantitative information describing the correlation between two eigenvectors, with a value of 1 demonstrating the two eigenvectors to be identical, and a value of 0 meaning the eigenvectors are orthogonal.



(a) MD1 and MD2: WT

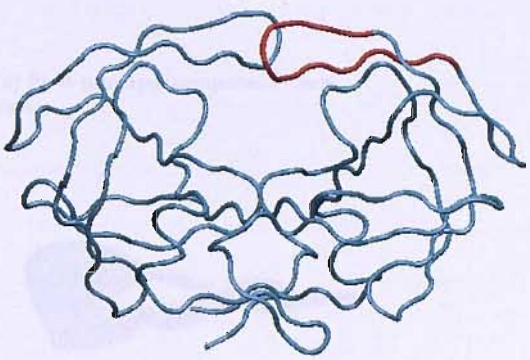
(b) MD1 and MD2: M46I

(c) MD1 and MD2: V82F/I84V

FIGURE 7.7: Comparison of inner products of the first ten eigenvectors of apo WT and mutant MD simulations of flap 2. Eigenvectors are plotted sequentially along the x- and y-axis, with 0 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Shading corresponds to the correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation. Black shading demonstrates the eigenvectors to be identical (value of 1) and white shading demonstrates the eigenvectors to be orthogonal (value of 0). The scale and range of shading is the same for all inner-product plots in this thesis. ([a]: Components of MD1 WT on x-axis, components of MD2 WT on y-axis, [b]: Components of MD1 M46I on x-axis, components of MD2 M46I on y-axis, [c]: Components of MD1 V82F/I84V on x-axis, components of MD2 V82F/I84V on y-axis)

Overall, the MD simulations of the WT show little diagonal correlation between the inner-products of eigenvectors of the first 10 modes of motion. The inner-products of the first eigenvectors of both MD1 and MD2 show a low level of correlation (correlation coefficient of 0.450), which is visualised in the projection of the trajectories of these simulations on to the eigenvectors of the flaps (figures 7.8 (c) and (d) of MD1 and 7.9 (a) and (b) for MD2). In both simulations, flap 2 is shown to move in a vertical direction, towards and away from the active site, with the flap in the MD2 simulation additionally showing a degree horizontal motion.

Since the second eigenvector of MD2 is shown to incorporate an element of vertical motion of the flap, in addition to a twisting movement of the flap-tips (figure 7.9 (c) and (d)), a small degree of correlation between this eigenvector and the first of MD1 is demonstrated (correlation coefficient: 0.292). No correlation exits between the inner-products of the second eigenvectors of MD1 and MD2, demonstrated by the very low correlation coefficient of 0.001. Instead, the motion determined by PCA to be the second most significant motion of the flap in the MD1 simulation is shown to have a high correlation with the third eigenvector of the MD2 simulation (correlation coefficient: 0.617), which involves horizontal and vertical motion of the flap residues (figures 7.8 (e) and (f) and 7.9 (e) and (f) for MD1 and MD2 respectively).



(a) side view

(b) top view

(c) First principal component - side view

(d) First principal component - top view

(e) Second principal component - side view

(f) Second principal component - top view

FIGURE 7.8: The first two principal components of flap 2 created from the MD 1 simulation of the WT HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling. The region coloured in red in (a) and (b) are the flap 2 residues, shown in figures (c)-(f). A complete set of figures showing the representative motion of the first three eigenvectors is given in the Appendix B.1.

(a) First principal component - side view

(b) First principal component - top view

(c) Third principal component - side view

(d) Third principal component - top view

FIGURE 7.9: The first and third principal components of flap 2 created from the MD 2 simulation of the WT HIV-1 PR system WT. Colour grading corresponds to values depicting the extent of sampling. A complete set of figures showing the representative motion of the first three eigenvectors is given in Appendix B.2.

Analysis of the inner-products of the eigenvectors generated by MD1 and MD2 simulations of the M46I mutant (figure 7.7) displays a relatively high correlation along the diagonal of the first six eigenvectors (correlation coefficients of eigenvectors 1: 0.462, 2: 0.329, 3: 0.321, 4: 0.548, 5: 0.494, 6: 0.455).

The first two eigenvectors additionally show off-diagonal correlation. For MD1, the first two eigenvectors incorporate a mixture of the vertical and horizontal motion (figures 7.10 (a) and (b) for MD1 and 7.11 (a) and (b) for MD2) also represented in the first two eigenvectors of MD2. This is demonstrated by the similar correlation coefficients between eigenvectors 1 and 2 of MD1 and MD2 (correlation coefficients: first eigenvector of MD1 and MD2: 0.462, second eigenvectors of MD1 and MD2: 0.329, second of MD1 and first eigenvector of MD2: 0.501, and the first eigenvector of MD1 and second eigenvector of MD2: 0.377).

(a) First principal component - side view

(b) First principal component - top view

FIGURE 7.10: The first principal component of flap 2 created from the MD1 simulation of the apo M46I mutant HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling. A complete set of figures showing the representative motion of the first three eigenvectors is given in the Appendix B.3.

(a) First principal component - side view

(b) First principal component - top view

FIGURE 7.11: The first principal component of flap 2 created from the MD2 simulation of the apo M46I mutant HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling. A complete set of figures showing the representative motion of the first three eigenvectors is given in B.4

The simulations of the V82F/I84V mutant show high correlation along the diagonal for the first two principal modes (correlation coefficients of 0.782 for eigenvector 1 and 0.701 for eigenvector 2), which are shown to capture the horizontal and vertical motion of the flap (figures 7.12 (a), (b) and (c), (d) for MD1 and 7.13 (a), (b) and (c), (d) for MD2). Additionally, off-diagonal correlation is observed for some of the less significant eigenvectors.

(a) First principal component - side view

(b) First principal component - top view

(c) Second principal component - side view

(d) Second principal component - top view

FIGURE 7.12: The first two principal components of flap 2 created from the MD1 simulation of the apo V82F/I84V mutant HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling. A complete set of figures showing the representative motion of the first three eigenvectors is given in Appendix B.5

(a) First principal component - side view

(b) First principal component - top view

(c) Second principal component - side view

(d) Second principal component - top view

FIGURE 7.13: The first two principal component of flap 2 created from the MD2 simulation of the apo V82F/I84V mutant HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling. A complete set of figures showing the representative motion of the first three eigenvectors is given in Appendix B.6.

Overall, the inner-product comparisons for MD1 and MD2 for each of the apo-systems shows low diagonal correlation for the principal modes of motion between the two MD simulations carried out for each system, which generally decreases for the less significant eigenvectors. As shown by the RMSD and RMSF results, flap 2 in shorter 10 ns simulations is indicated to be less mobile, so motions are of a lower amplitude and less well-defined. This has resulted in the dissimilarity observed between the major motions of the two MD simulations. Owing to the variability between the motions of the two MD simulations for the same system, comparisons cannot be made between the WT and mutant MD trajectories, and alternative simulation procedures must therefore be sought.

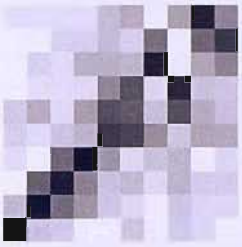
RDFMD Simulations

PCA analysis has been carried out in the same manner for the RDFMD simulations as for the MD simulations.

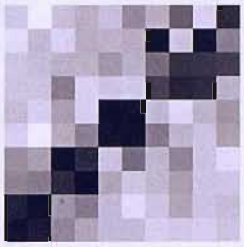
Trajectories for analysis have been created from the concatenation of the twelve individual RDFMD simulation trajectories, thus producing three trajectories, one for each apoenzyme (WT, M46I and V82F/I84V). The α-carbon atom coordinates of each of the trajectories for one of the flaps constitutes the data-set for the calculation of the covariance matrix, created whilst fitting on the non-flap residues of the structure. The reference structure used for PCA must be the same if comparison between trajectories is required. Therefore, the reference structure used in each case was the α-carbon coordinates of the equilibrated WT apo-structure.

The process described above was then repeated for the other flap. To compare directly the sampling of the eigenvectors generated by the WT and mutant simulations, the eigenvectors must be similar. If the eigenvectors being compared were identical, the inner-product would be unity.

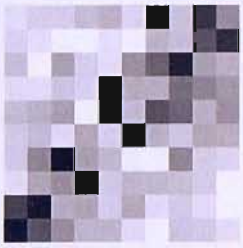
The inner-products, comparing the eigenvectors of the WT and mutant trajectories for both flap 1 and flap 2 are shown in figures 7.14 and 7.15 respectively. For both flaps, correlation along the diagonal is evident, as shown by the relatively high correlation values for eigenvectors 1-3 observed between the WT and mutant simulations (see table 7.3) which is observed in most cases. For the remainder of the 10 eigenvectors shown in the inner-product matrices, flap 1 shows less diagonal correlation, and additionally displays a larger degree of off-diagonal correlation compared with flap 2. This larger amount of off-diagonal correlation for flap 1 is due to the smaller amplitude of motions of this flap, as indicated by the RMSF and RMSD values calculated earlier.



(a) M46I and V82F/I84V

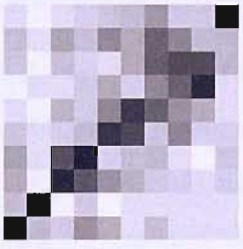


(b) WT and M46I

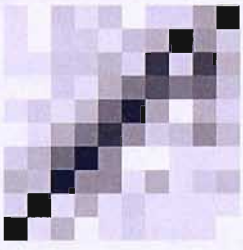


(c) WT and V82F/I84V

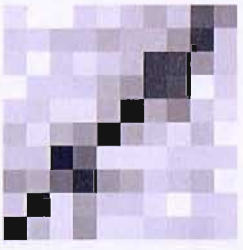
FIGURE 7.14: Comparison of inner products of the first ten eigenvectors of apo WT and mutant RDFMD simulations of flap 1. Eigenvectors are plotted sequentially along the x- and y-axis, with 1 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Shading corresponds to correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation ([a]: Components of M46I on x-axis, components of V82F/I84V on y-axis, [b]: Components of WT on x-axis, components of M46I on y-axis, [c]: Components of WT on x-axis, components of V82F/I84V on y-axis).



(a) M46I and V82F/I84V



(b) WT and M46I



(c) WT and V82F/I84V

FIGURE 7.15: Comparison of inner products of the first ten eigenvectors of apo WT and mutant RDFMD simulations of flap 2. Eigenvectors are plotted sequentially along the x- and y-axis, with 1 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Shading corresponds to correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation. ([a]: Components of M46I on x-axis, components of V82F/I84V on y-axis, [b]: Components of WT on x-axis, components of M46I on y-axis, [c]: Components of WT on x-axis, components of V82F/I84V on y-axis).

Correlation Coefficients along the Diagonal		
Flap	Simulations compared	Correlation Coefficient (Eigenvector)
1	M46I and V82F/I84V	0.976 (1), 0.720 (2), 0.592(3)
1	WT and M46I	0.535 (1), 0.511 (2), 0.285 (3)
1	WT and V82F/I84V	0.626 (1), 0.565 (2), 0.072 (3)
2	M46I and V82F/I84V	0.838 (1), 0.970 (2), 0.545 (3)
2	WT and M46I	0.962 (1), 0.946 (2), 0.696 (3)
2	WT and V82F/I84V	0.876 (1), 0.912 (2), 0.180 (3)

TABLE 7.3: Correlation coefficients along the diagonal for the first three eigenvectors of WT and mutant RDFMD simulations. Matrix representation demonstrating the levels of correlation between the first 10 eigenvectors are shown in figures 7.14 and 7.15.

For flap 2, the important major modes of motion between the trajectories of the WT, M46I and V82F/I84V are shown to be similar, demonstrated by the high correlation of the inner-products of the first 2-3 eigenvectors of the WT and mutant trajectories of flap 2 (see table 7.3). The α-carbon trajectories of all apo RDFMD simulations (totalling 36) have therefore been concatenated together, creating a single trajectory. In this way, projections of each trajectory onto these composite eigenvectors allows the differences in phase space sampled to be more clearly resolved. Again, the covariance matrix was calculated using the trajectory coordinates of the α-carbon atoms of each of the flaps individually, using the same equilibrated WT reference structure as described previously, and the results are discussed below.

The decision of an appropriate number of the resultant eigenvectors to study has been determined through the use of scree plots (figures 7.16 and 7.17) and visualisation of the trajectories projected onto the eigenvectors. The first three principal components of each flap, accounting for 77.2 % and 83.2 % of the total motion respectively have been chosen for analysis (see tables 7.4 and 7.5). The remaining eigenvectors correspond to smaller, less significant motions of the flaps and have not been analysed further.

First ten principal components of flap1			
Eigenvector	Eigenvalue (nm^2)	% total variance	Cumulative %
1	0.226	40.2	40.2
2	0.150	26.6	66.8
3	0.0592	10.5	77.3
4	0.0467	8.30	85.6
5	0.0179	3.19	88.8
6	0.0155	2.75	91.5
7	0.0102	1.80	93.3
8	0.00963	1.71	95.0
9	0.00580	1.03	96.1
10	0.00462	0.821	96.9

TABLE 7.4: Motion captured in the first ten principal components of flap 1

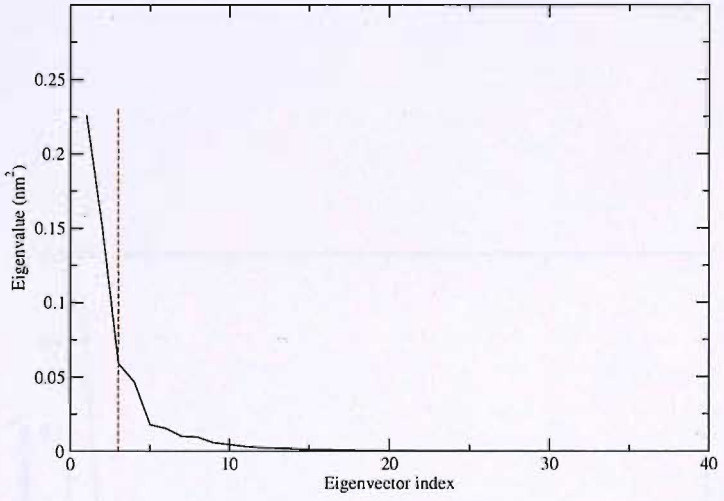


FIGURE 7.16: Scree plot depicting the proportion of motion captured in each of the eigenvectors of flap 1. The red line depicts the eigenvalues for the first three eigenvectors.

First ten principal components of flap 2			
Eigenvector	Eigenvalue (nm^2)	% total variance	Cumulative %
1	0.428	55.3	55.3
2	0.151	19.6	74.9
3	0.0635	8.22	83.2
4	0.0539	6.98	90.12
5	0.0191	2.48	92.6
6	0.0134	1.74	94.3
7	0.00909	1.18	95.5
8	0.00803	1.04	96.6
9	0.00641	0.829	97.4
10	0.00376	0.487	97.9

TABLE 7.5: Motion captured in the first ten principal components of flap 2

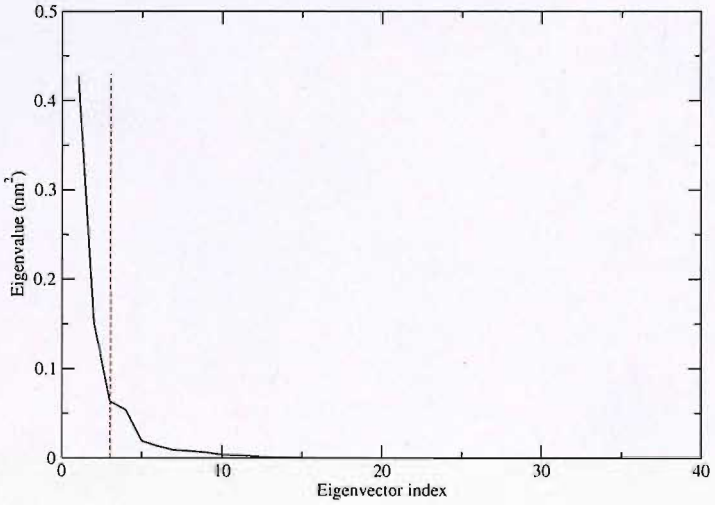
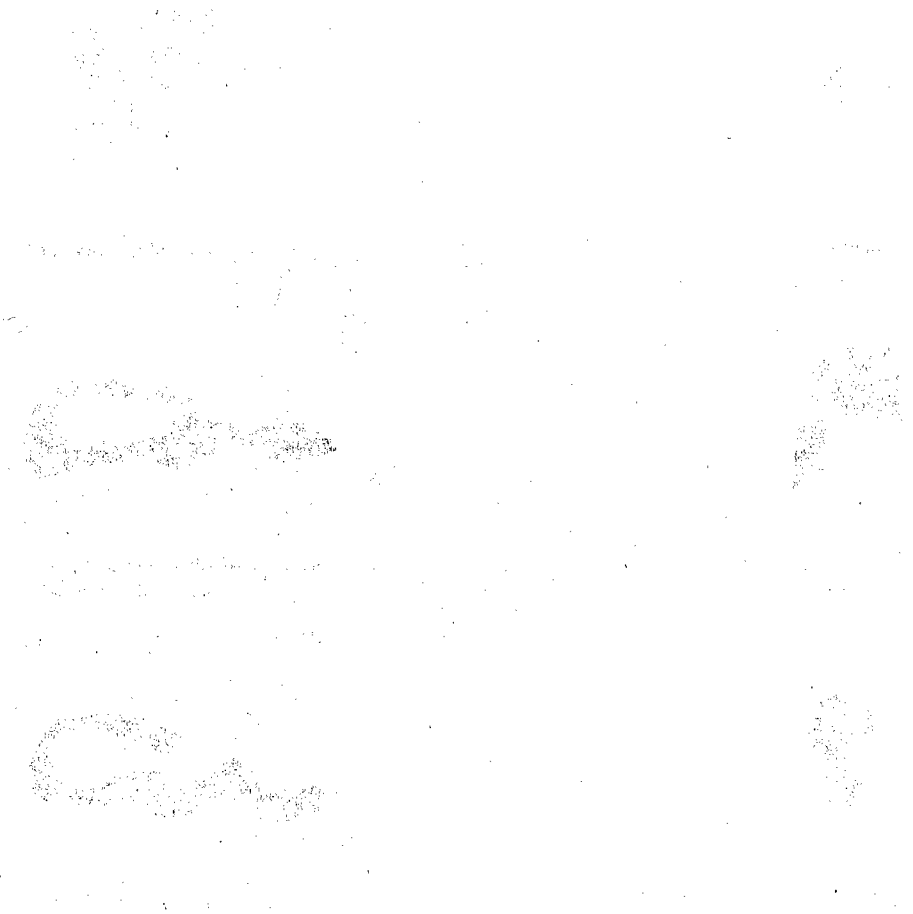
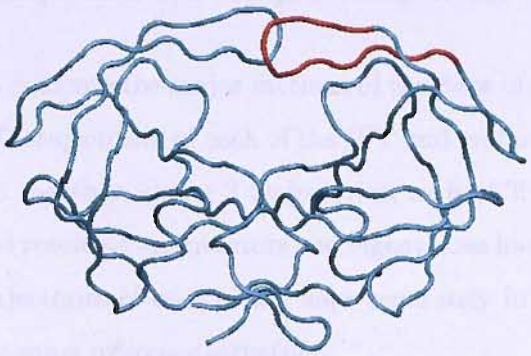


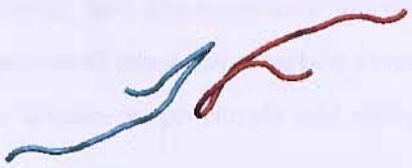
FIGURE 7.17: Scree plot depicting the proportion of motion captured in each of the eigenvectors of flap 2. The red line depicts the eigenvalues for the first three eigenvectors.

Projection of the RDFMD trajectories onto the eigenvectors generated by flap 2 displays the motion represented by the first three eigenvectors (figure 7.18). Figures 7.18 (c) and (d) show the first principal component of flap 2 to capture the vertical movement of the flap, towards and away from the active site, and figures 7.18 (e) and (f) shows the second principal component to demonstrate a horizontal gating motion with the flap moving perpendicular to the vertical motion of the flap. Together, these first two principal components provide evidence of conformations of the flaps which allow/disallow access to the active site cavity. The third principal component shows the flap-tip curling motion, which will be discussed later in this chapter, in section 7.3.2.2.





(a) side view



(b) top view

(c) First principal component - side view

(d) First principal component - top view

(e) Second principal component - side view

(f) Second principal component - top view

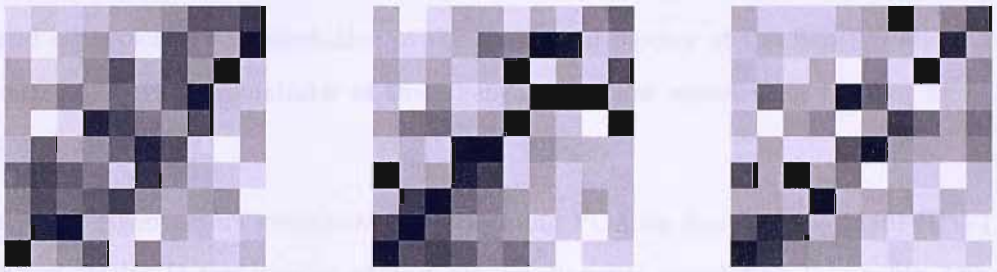
(g) Third principal component - side view

(h) Third principal component - top view

FIGURE 7.18: The first three principal component of flap 2 created from a concatenated trajectory of all WT and mutant RDFMD simulations. Colour grading corresponds to values depicting the extent of sampling. The region marked in red on figures (a) and (b) are the flap 2 residues shown in figures (c)-(h).

Comparison of Principal Components of RDFMD and MD Simulations

To compare the major motions of the flaps of the MD and RDFMD simulations, the two MD trajectories of each of the WT and mutant HIV-1 PR systems have been concatenated together, giving 3 trajectories, each of 30 ns. Following this, the covariance matrix and resulting eigenvectors and eigenvalues have been generated using the α-carbon atom trajectories of each of the flaps separately in the same manner as previously and using the same reference structure.



(a) (b) (c)

FIGURE 7.19: Comparison of inner products of the first ten eigenvectors of apo WT MD and RDFMD simulations of flap 1. Shading corresponds to correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation. ([a]: components of WT MD on x-axis, components of RDFMD on y-axis, [b]: components of M46I MD on x-axis, components of RDFMD on y-axis, [c]: components of V82F/I84V MD on x-axis, components of RDFMD on y-axis).

Figures 7.19 and 7.20 compare the first ten eigenvectors generated by the RDFMD and MD simulations through inner-product analysis. The analysis shows the motion captured in the first three eigenvectors of the RDFMD simulations, corresponding to the vertical motion, horizontal gating motion and the flap-tip curling are also observed in the first few eigenvectors of the MD simulations, although a large proportion of correlation between the eigenvectors is off-diagonal figures 7.19 and 7.20.

Comparison of the inner-products of the MD and RDFMD simulations for flap 1 of the WT show correlation across the diagonal for the first two eigenvectors (0.651 for eigenvector 1, and 0.367 for eigenvector 2, see table 7.6), demonstrating the MD simulations of this system to incorporate vertical and horizontal motion as with the RDFMD simulations. The third eigenvector of the MD simulations only has a low correlation along

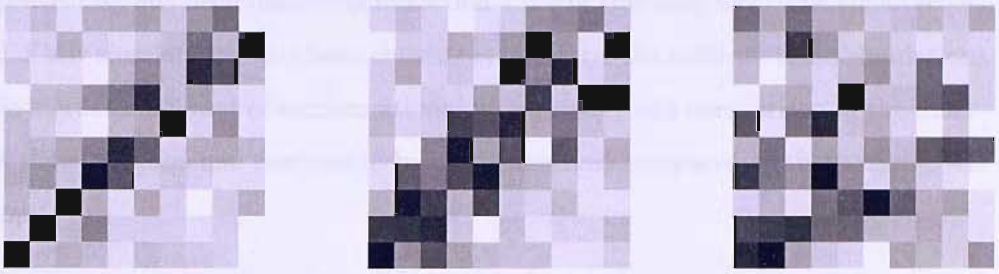
Correlation Coefficients along the Diagonal		
Flap	Simulations compared	Correlation Coefficient (Eigenvector)
1	WT MD and RDFMD	0.651 (1), 0.367 (2), 0.106(3)
1	M46I MD and RDFMD	0.367 (1), 0.528 (2), 0.485 (3)
1	V82F/I84V MD and RDFMD	0.483 (1), 0.391 (2), 0.011 (3)
2	WT MD and RDFMD	0.907 (1), 0.855 (2), 0.848 (3)
2	M46I MD and RDFMD	0.656 (1), 0.306 (2), 0.255 (3)
2	V82F/I84V MD and RDFMD	0.262 (1), 0.195 (2), 0.00 (3)

TABLE 7.6: Correlation coefficients along the diagonal for the first three eigenvectors of WT and mutant RDFMD and MD simulations. Matrix representation demonstrating levels of correlation between the first 10 eigenvectors are shown in figures 7.14 and 7.15.

the diagonal (correlation coefficient: 0.106), but a significantly stronger correlation with the second eigenvector, representative of the horizontal motion of the flap (correlation coefficient: 0.527). The remainder of the eigenvectors show correlation, but not along the diagonal axis.

The first three eigenvectors generated by performing PCA on flap 1 of the M46I HIV-1 PR MD and RDFMD trajectories show a strong diagonal correlation (see table 7.6). Good off-diagonal correlation between the first eigenvector of the MD simulations and the fourth of the RDFMD simulations (correlation coefficient: 0.381) is also displayed. As observed in the comparison of the eigenvectors generated by WT HIV-1 PR MD and RDFMD simulations, the significant motions depicted by the first three eigenvectors of the RDFMD simulations are also captured as the principal motions of the M46I MD simulations, but not necessarily in the same order, and in many cases, incorporate off-diagonal correlation.

Comparison of the eigenvectors generated by the MD and RDFMD simulations of flap 1 of the V82F/I84V HIV-1 PR systems demonstrates diagonal (see table 7.6 for diagonal correlation values) and off-diagonal (correlation coefficients: between eigenvector 1 of MD simulations and eigenvector 2 of RDFMD simulations: 0.329, between eigenvector 2 of MD simulations and eigenvector 1 of RDFMD simulations: 0.287) correlation between the first two eigenvectors. The flap-tip curling motion, represented as the third eigenvector of the RDFMD simulations is shown to have strongest correlation with the fourth eigenvector of the MD simulations (correlation coefficient: 0.739).



(a) WT (b) M46I (c) V82F

FIGURE 7.20: Comparison of inner products of the first ten eigenvectors of apo WT MD and RDFMD simulations of flap 2. Shading corresponds to correlation between eigenvectors, with dark shading corresponding to high correlation and light shading to a low correlation. ([a]: Components of WT MD on x-axis, components of RDFMD on y-axis, [b]: Components of M46I MD on x-axis, components of RDFMD on y-axis, [c]: Components of V82F/I84V MD on x-axis, components of RDFMD on y-axis).

Comparison of the inner-products for flap 2 (figure 7.20) shows the RDFMD and MD simulations of the WT apo-system to have strong diagonal correlation for the first five eigenvectors (values for the first three are given in table 7.6), demonstrating similar ordering of the motions which contribute most to the total motion in the simulations. Comparing the first three eigenvectors of MD and RDFMD simulations of the mutant HIV-1 PR systems does not show as strong diagonal correlation as the WT, but does demonstrate that the important motions represented in the first three eigenvectors of the RDFMD simulations are also observed in the few eigenvectors of the MD simulations, but are correlated in an off-diagonal fashion.

Overall, the comparison of the important eigenvectors of the WT and mutant HIV-1 PR MD simulations with the eigenvectors of the RDFMD simulations reveal the major modes of motion to be the similar, thus validating the RDFMD method to sample the same types of motion as MD simulations.

7.3.2.1 Analysis of Eigenvectors 1 and 2

To study the conformations shown by the flaps, and to investigate whether the presence of the mutations studied here have an effect on their dynamics, the MD and RDFMD trajectories of each of the apo-systems have been projected onto the eigenvectors created from the covariance matrix of the concatenated RDFMD simulations (the generation of

these eigenvectors has been described in section 7.3.2 of this chapter). The eigenvectors of the RDFMD simulations have been chosen, as initial results indicate these simulations to sample a greater amount of conformational phase space, and comparison of the inner-products show the principal motions to be better resolved compared to those of the MD simulations.

As discussed earlier, analysis has been focused on flap 2 since earlier analysis indicates this flap to display more conformational change compared to flap 1. The flaps move in an asymmetric fashion and it is expected that the flaps adopt different conformations at any given point, and although it is perhaps expected that both flaps have the same average properties, it is shown in this study, and in other theoretical studies also, that a difference in the motions undergone by the flaps of the two monomers exists [221].

Two-dimensional PCA plots shown in this chapter demonstrate the correlation between the sampling of two selected eigenvectors. As mentioned earlier, the first three eigenvectors have been considered to represent the major motions of the HIV-1 PR systems studied here, and therefore, the sampling of these eigenvectors are the focus of study.

The x-axis of the two-dimensional plots represents the sampling along the first eigenvector (vertical motion), with positive values representing the flap to move towards the active site, and negative values where the flap is moving away from the active site in a vertical direction. The y-axis demonstrates the sampling of the gating motion of flap 2, with negative values corresponding to conformations where the flap is over the active site, and positive values correspond to conformations where the flap is away from the active site.

The vertical movement and the horizontal gating motion, perpendicular to the vertical movement of flap 2, are motions captured by eigenvectors 1 and 2 of the RDFMD simulations respectively. Collectively, these eigenvectors describe the overall conformation of the flap as being open (access to the active site), closed (no access to the active site, and flaps are swapped) or semi-open (intermediate conformations between open and closed) (note: the different conformations of HIV-1 PR are discussed in more detail in section 6.2.1 of chapter 6). Eigenvector 3 corresponds to the flap-tip curling motion, which is described later in this chapter.

Distance measurements have also been used to confirm conformations identified by PCA in this chapter. Measuring the distance between the flap-tip residues (α-carbon of residues Gly50 and Gly150 of flap 1 and flap 2 respectively) over the length of the simulations gives an indication of accessibility to the active site, and measurement between the α-carbon atom of Gly150 flap-tip residue and the α-carbon atom of the catalytic Asp125 residue indicates the vertical mobility of flap 2.

WT HIV-1 PR

Figure 7.21 displays the sampling of the WT MD and RDFMD simulations when projected onto the first two eigenvectors, and representative conformations of flap 2 sampled in the different regions of the plot are shown in figures 7.22 and 7.23.

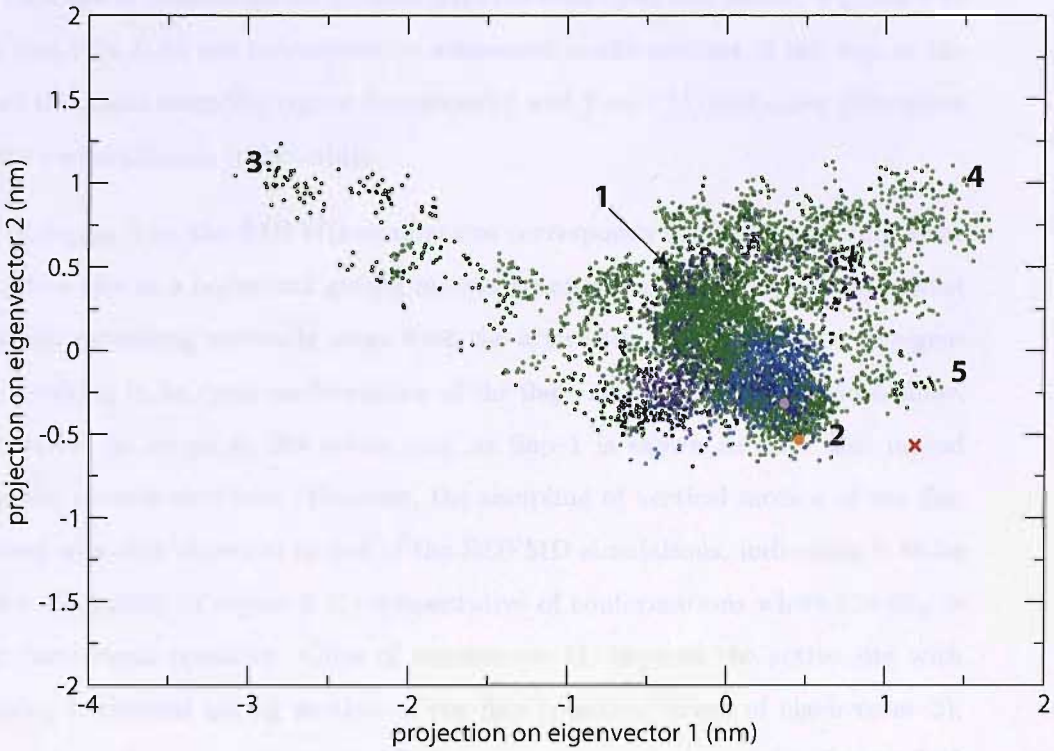
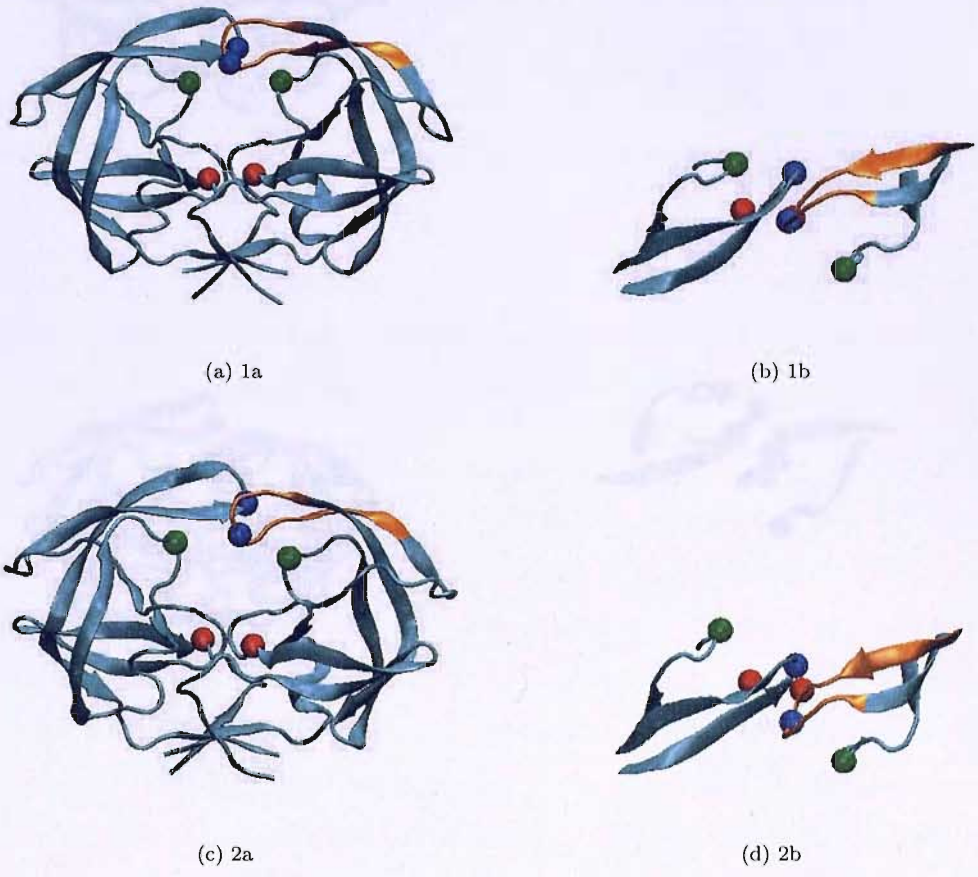


FIGURE 7.21: Sampling of the first two eigenvectors, generated from the projection of the apo WT RDFMD (green) and the two MD (blue and maroon) simulations onto the eigenvectors of the second flap created from the concatenation of all RDFMD trajectories. (the two starting structures are represented by an orange and brown circle). The numbers correspond to conformations of flap 2 represented in figures 7.22 and 7.23. The red cross depicts the projection of the closed conformation of flap 2 of the SQV-bound structure (1HXB) onto the eigenvectors.

The projection of the RDFMD simulations onto the first two eigenvectors show the main area of sampling to lie within a defined region, with an approximate boundary of projection values -0.5 to 0.5 nm on the x-axis, and -0.5 to 0.5 nm on the y-axis. The two different starting semi-open structure coordinates are shown to be positioned near the perimeter of this area. Sampling in this described region is representative of a variety of semi-open conformations, with no large conformational changes of flap 2. There is no rigorous definition of the semi-open conformation of the flaps, it is loosely described as when the flaps are in conformations intermediate between open and closed. Figures 7.22 1 (a & b) and 2 (a & b) are representative semi-open conformations of the flap at the extremes of this main sampling region (numbered 1 and 2 on 7.21) and show differences between the conformations to be subtle.

Sampling of region 3 by the RDFMD simulations corresponds to the flap moving away from the active site in a horizontal gating motion (positive values along y-axis), whilst simultaneously extending vertically away from the active site (negative values of eigenvector 1), resulting in an open conformation of the flap. As figures 7.23 3a and b show, this has resulted in access to the active site, as flap 1 is shown to have also moved away from the protein structure. However, the sampling of vertical motion of the flap to this extent was only observed in one of the RDFMD simulations, indicating it to be a rare event. Sampling in region 4 is representative of conformations where the flap is positioned downwards (positive values of eigenvector 1), towards the active site with accompanying horizontal gating motion of the flap (positive values of eigenvector 2), moving away from the active site cavity. This type of motion is depicted in figures 7.23 4a and 4b and could also result in access to the active site if flap 1 is in a suitable conformation.

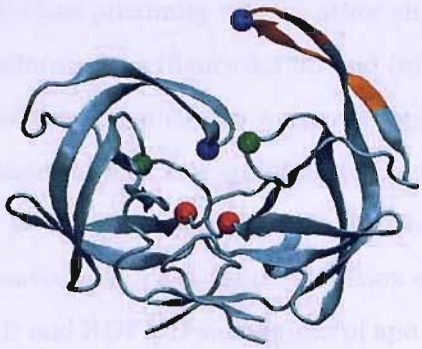
Conformations of flap 2 corresponding to sampling in region 5 of the plot, like region 4, show the flap to point downwards, towards the active site (positive values of eigenvector 1), but unlike conformations associated with sampling in region 4, projection onto eigenvector 2 shown in this region corresponds to simultaneous sampling of the flap lying above the active site (negative values of eigenvector 2), similar to its position in the starting conformation. The flap in this conformation does not provide access to the active site (figures 7.23 5a and 5b).



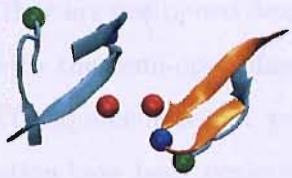
(a) 1a (b) 1b

(c) 2a (d) 2b

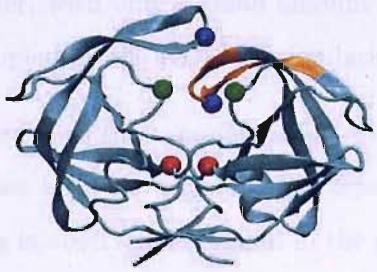
FIGURE 7.22: Conformations of flap 2 (in orange) representative of the main area of sampling of apo WT HIV-1 RDFMD simulations projected onto eigenvectors 1 and 2. Numbering corresponds to the region sampled in figure 7.21, with a: side view, b: top view. Flap 2 structure: orange, atoms highlighted in blue represent the α-carbon atoms of residue 50 and 150, red: α-carbon atoms of residues 25 and 125 of the active site and green: α-carbon atoms of residues 81 and 180 of the 80's loops



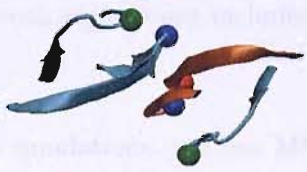
(a) 3a



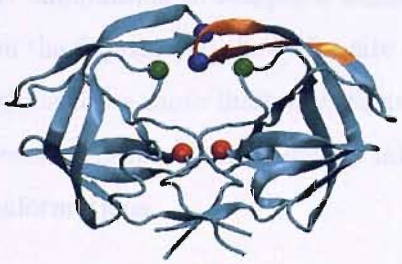
(b) 3b



(c) 4a



(d) 4b



(e) 5a

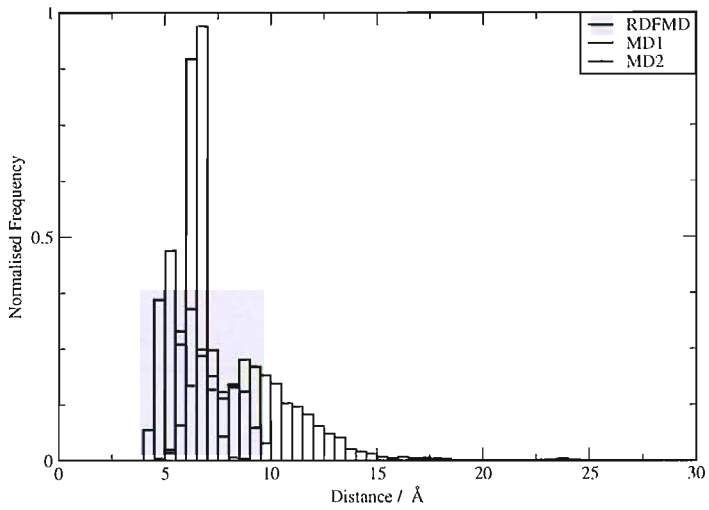
(f) 5b

FIGURE 7.23: Conformations of flap 2 of apo WT HIV-1 from the projection of RDFMD simulations onto eigenvectors 1 and 2. Numbering corresponds to the region sampled in figure 7.21, with a: side view, b: top view. Flap 2 structure: orange, atoms highlighted in blue represent the α-carbon atoms of residue 50 and 150, red: α-carbon atoms of residues 25 and 125 of the active site and green: α-carbon atoms of residues 81 and 180 of the 80's loops

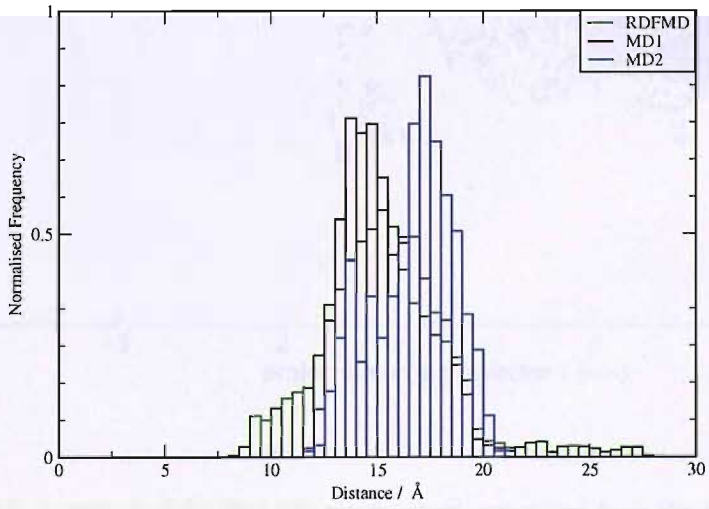
The closed flap conformation of HIV-1 PR (figure 6.3 (b) and (d)) is most commonly seen in studies where an inhibitor is bound. In this conformation, the two flaps are shown to lie in close proximity to each other and are swapped in compared with the semi-open flap conformations (figure 6.3 (a) and (c)). Additionally, they are positioned downwards towards the active site to a greater degree compared with the semi-open flaps. The coordinates of flap 2 of an inhibitor-bound structure (WT saquinavir-bound, pdb code 1HXB [383]), where the flaps are in this closed conformation have been projected onto the eigenvectors. The resulting location of the projections are represented by a red cross. The MD and RDFMD simulations of apo WT HIV-1 PR have not sampled the closed flap conformations in the simulations presented in this chapter, with conformations sampled in regions 2 and 5 being the most similar.

The sampling of the two MD simulations is shown by the blue and maroon circles for MD1 and MD2 respectively. These simulations are shown to sample different regions to each other, with only a small amount of overlap, but both regions are included in the area sampled by the RDFMD simulations.

Compared with the sampling achieved by the RDFMD simulations, the two MD simulations are both shown to sample fewer conformations of these two eigenvectors. The sampling is observed to remain in the region representative of the semi-open conformation of the flap, and does not sample the vertical motion of the flap to such an extent as seen in the RDFMD simulations (region 3), although, the sampling in region 3 is the result of a rare vertical opening event, only seen in one of the RDFMD simulations to such an extent. This is also demonstrated in distance measurements. Figure 7.24 shows the MD simulations to sample a smaller range of distances between the flap-tips and between the flap-tip and the active site aspartic acid residue. This demonstrates the MD simulations to be more limited in their sampling and indicates that the simulations to have become trapped in local energy minima, preventing the sampling of all the possible flap conformations.



(a) residues 50-150



(b) residues 150-125

FIGURE 7.24: Distance measured between the α-carbon atoms of residues 50 and 150 (flap-tips) and between the α-carbon atoms of residues 50 and 25 (flap-tip - catalytic aspartic acid residues) over the length of the 2 MD and 12 RDFMD simulations. For each of the MD simulations and the total 12 RDFMD simulations, the total number of distances measured has been normalised to a sum of 1.

M46I HIV-1 PR

Sampling of the first two eigenvectors by the M46I mutant by RDFMD and MD simulations is presented in figure 7.25, and conformations of flap 2 which are representative of the sampled regions of the 2-dimensional plot are shown in figures 7.26 and 7.27.

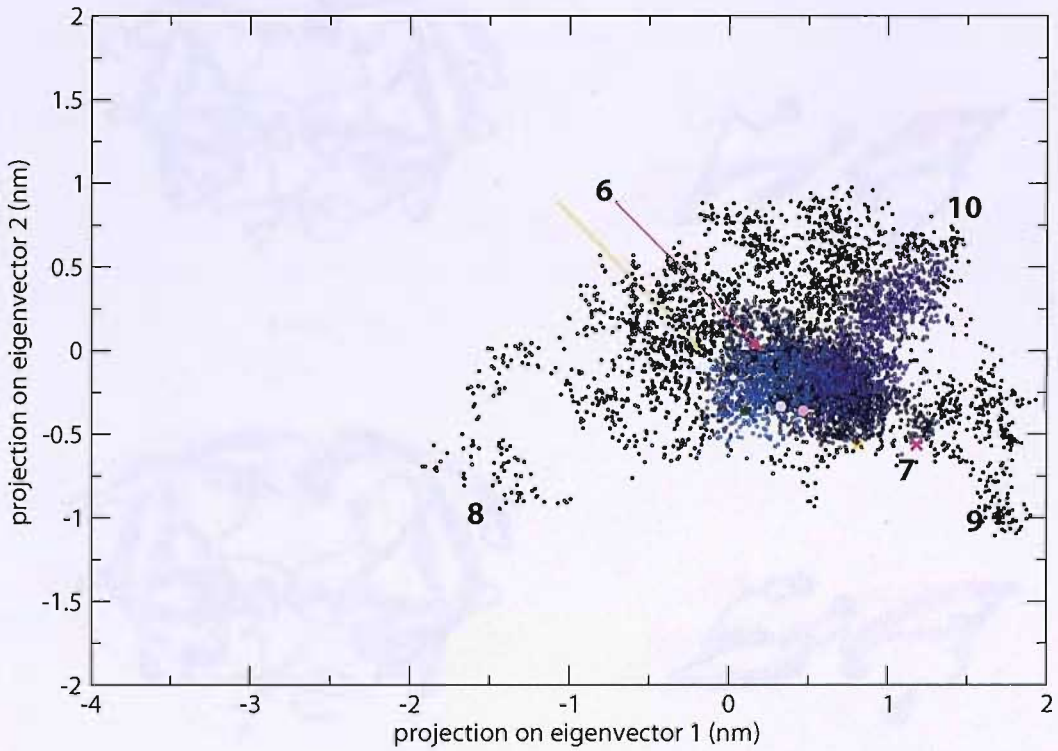
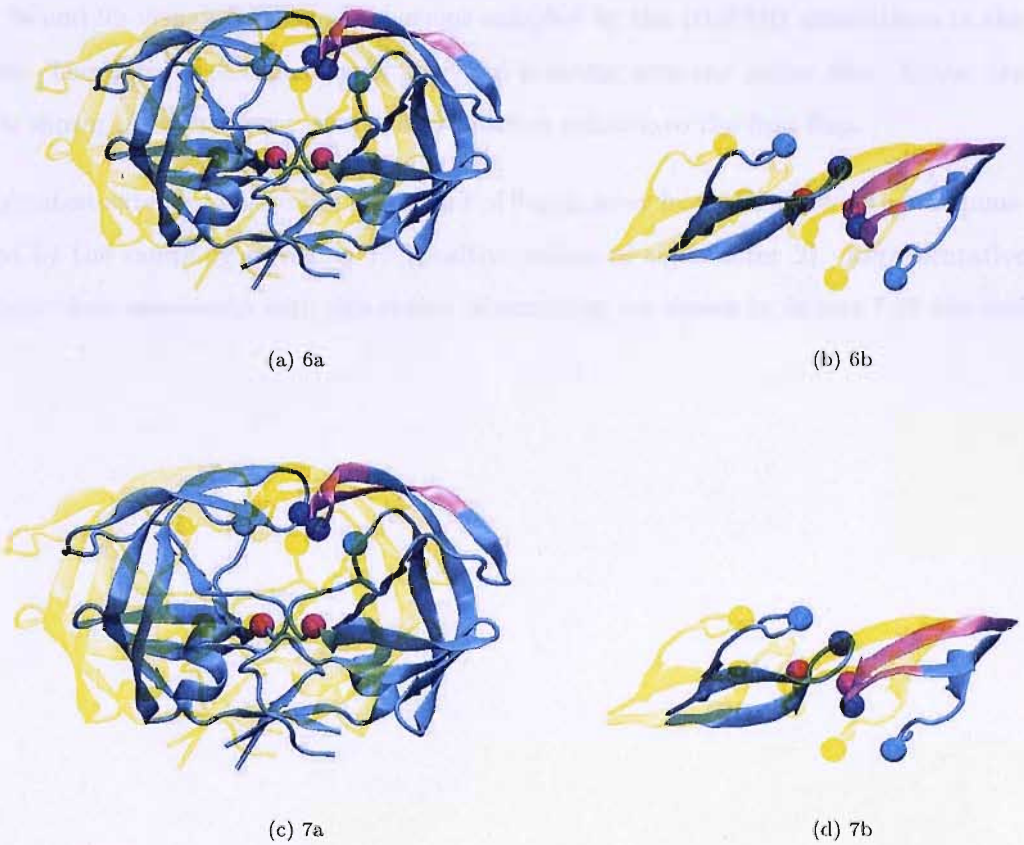


FIGURE 7.25: Sampling of the first two eigenvectors, generated from the projection of the apo M46I mutant RDFMD (black) and the two MD (blue and maroon) simulations onto the eigenvectors of the second flap created from the concatenation of all RDFMD trajectories. (The two starting structures are represented by orange and brown circles). The numbers correspond to conformations of flap 2 represented in figures 7.26 and 7.27. The red cross depicts the projection of the closed conformation of flap 2 of the SQV-bound structure (1HXB) onto the eigenvectors

Projection of the apo M46I HIV-1 PR RDFMD simulations onto the first two eigenvectors shows the main area of the sampling in the region marked by the values 0 - 1 nm on the x-axis, and -0.5 - 0 nm on the y-axis. Representative conformations of the flap at the extremes of this region (labelled as 6 and 7 on figure 7.25) are shown in figures 7.26 6(a & b) and 7(a & b)). They show the flap to be in similar semi-open conformations, therefore demonstrating that no major conformational change occurs in this region.



(a) 6a (b) 6b

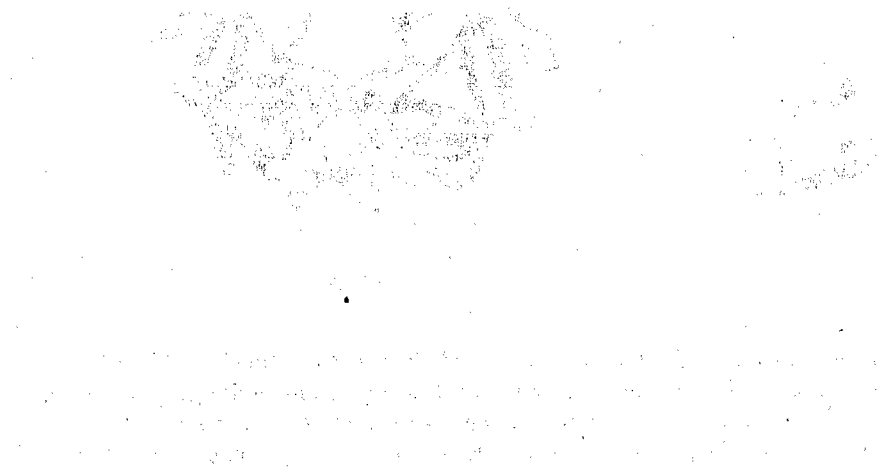
(c) 7a (d) 7b

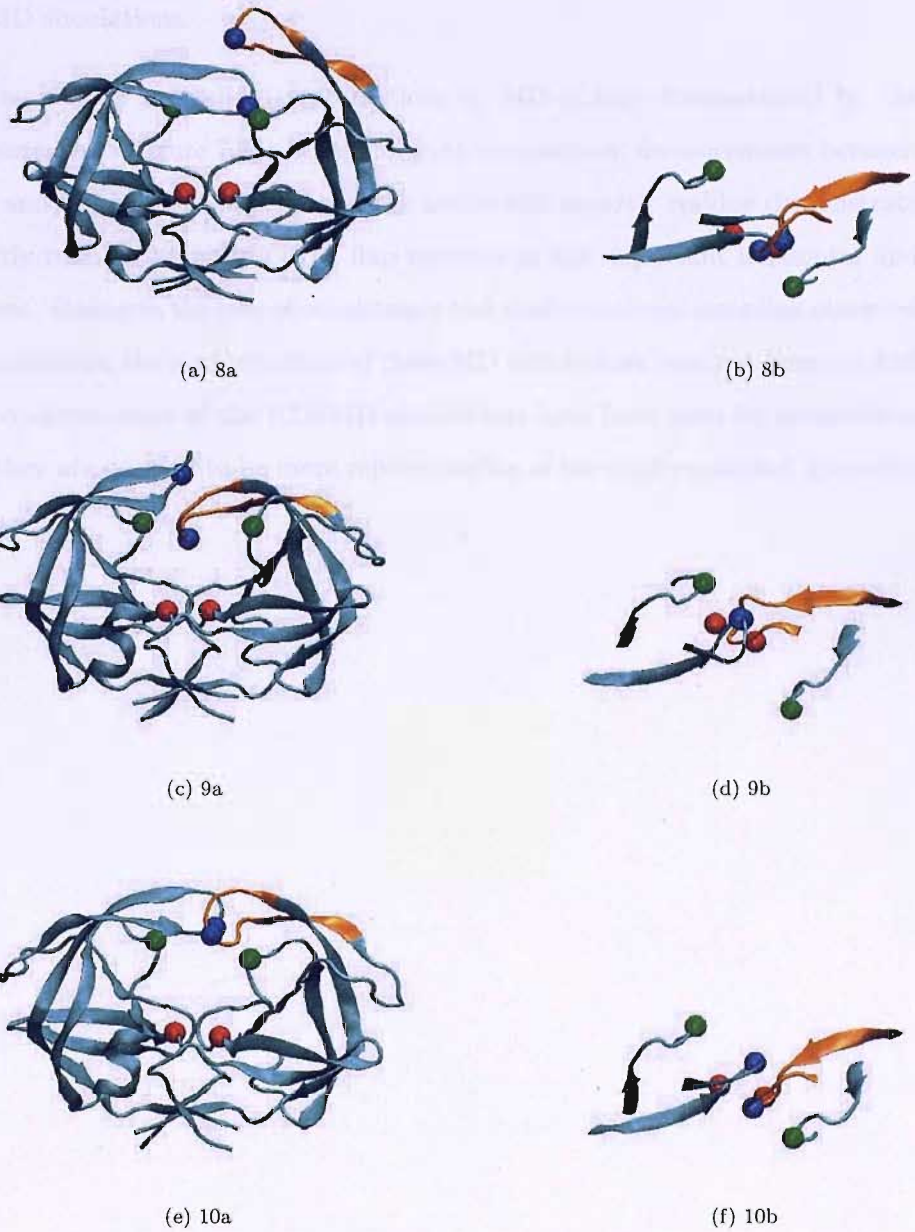
FIGURE 7.26: Conformations of flap 2 representative of the bulk area of sampling of apo M46I HIV-1 RDFMD simulations projected onto eigenvectors 1 and 2. Numbering corresponds to the region sampled in figure 7.25, with a: side view, b: top view. Flap 2 structure: orange, atoms highlighted in blue represent the α-carbon atoms of residue 50 and 150, red: α-carbon atoms of residues 25 and 125 of the active site and green: α-carbon atoms of residues 81 and 180 of the 80's loops

A small amount of sampling occurs in region 8 of the 2-dimensional plot, with the associated conformations shown in figures 7.27 8a and 8b. Flap 2 is shown to vertically extend away from the non-flap residues (negative values of eigenvector 1), whilst lying over the active site (negative values of eigenvector 2), almost swapping in orientation with respect to flap 1.

Sampling of region 9 overlaps the projection of the coordinates of the closed conformation of flap 2 of the inhibitor-bound HIV-1 PR enzyme, demonstrating that RDFMD simulations of this mutant have sampled the closed-type conformation of flap 2. Figures 7.27 9a and 9b visualise the conformations sampled by the RDFMD simulations in this region. The flap is positioned lying over and pointing into the active site. Again, the flap is shown to have almost swapped in position relative to the first flap.

The greatest extent of horizontal movement of flap 2, away from the active site is demonstrated by the sampling in region 10 (positive values of eigenvector 2). Representative conformations associated with this region of sampling are shown in figures 7.27 10a and 10b.





(a) 8a (b) 8b

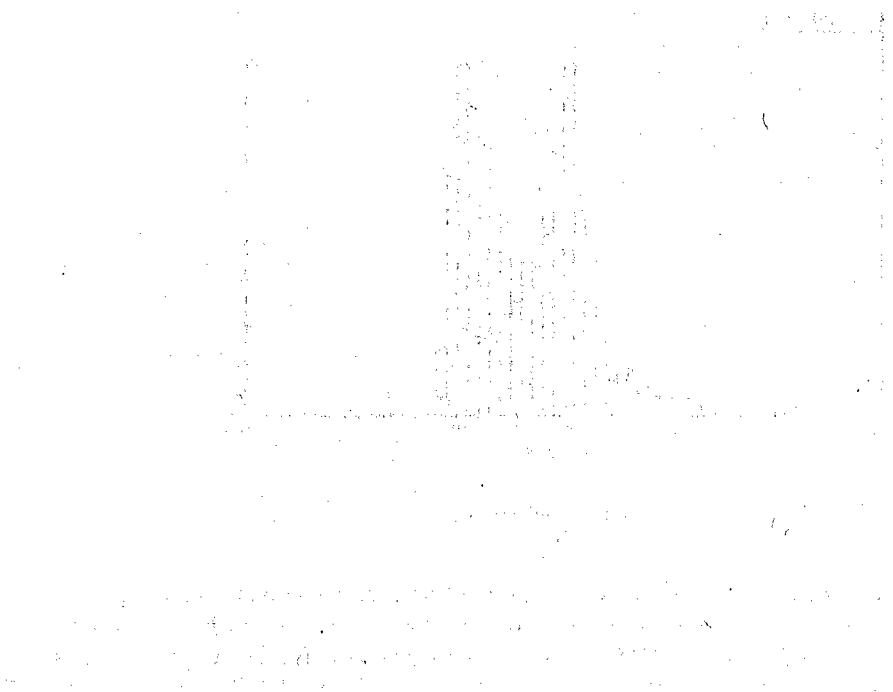
(c) 9a (d) 9b

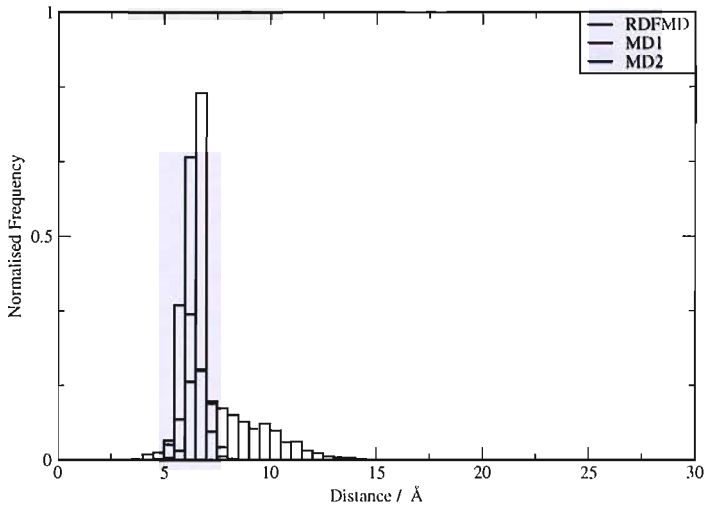
(e) 10a (f) 10b

FIGURE 7.27: Conformations of flap 2 of apo M46I HIV-1 from the projection of RDFMD simulations onto eigenvectors 1 and 2. Numbering corresponds to the region sampled in figure 7.25, with a: side view, b: top view. Flap 2 structure: orange, atoms highlighted in blue represent the α-carbon atoms of residue 50 and 150, red: α-carbon atoms of residues 25 and 125 of the active site and green: α-carbon atoms of residues 81 and 180 of the 80's loops.

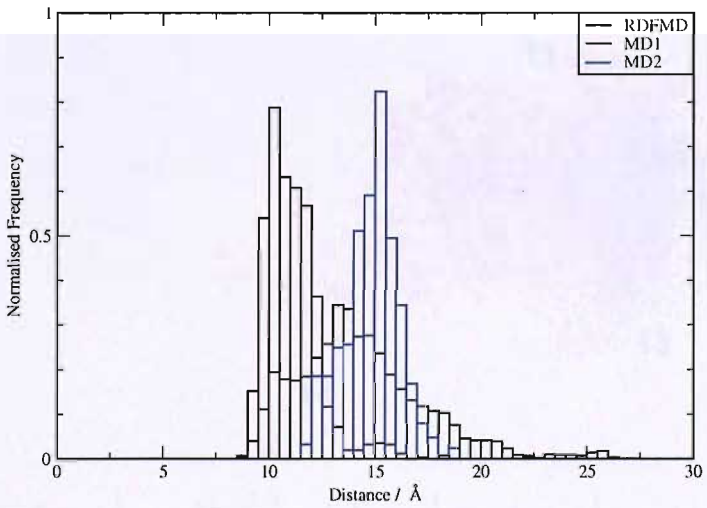
The projection of the two MD simulations onto the first two eigenvectors show sampling of different regions, with some overlap. All the regions sampled correspond to are semi-open type conformations of flap 2 and only covers a small proportion of that sampled by the RDFMD simulations.

The smaller number of accessible conformations to MD is also demonstrated by the distance measurements (figure 7.28), where for both simulations, measurements between the flap-tips, and between the flap-tip and the active site aspartic residue demonstrate the significantly reduced mobility of the flap residues in the important horizontal and vertical motions. Owing to the lack of consistency and conformational sampling observed in the MD simulations, the conformations of these MD simulations have not been studied further. The conformations of the RDFMD simulations have been used for comparison purposes, as they are deemed to be more representative of the conformational dynamics of this system.





(a) residues 50-150



(b) residues 50-150

FIGURE 7.28: Distance measured between the α-carbon atoms of residues 50 and 150 (flap-tips) and between the α-carbon atoms of residues 50 and 25 (flap-tip - catalytic aspartic acid residues) over the length of the 2 MD and 12 RDFMD simulations of the apo-M46I HIV-1 PR. For each of the MD simulations and the total 12 RDFMD simulations, the total number of distances measured has been normalised to a sum of 1.

V82F/I84V HIV-1 PR

Projection of the RDFMD simulations onto the first two eigenvectors is shown in figure 7.29 and representative conformations of flap 2 in each of the numbered regions are shown in figure 7.30.

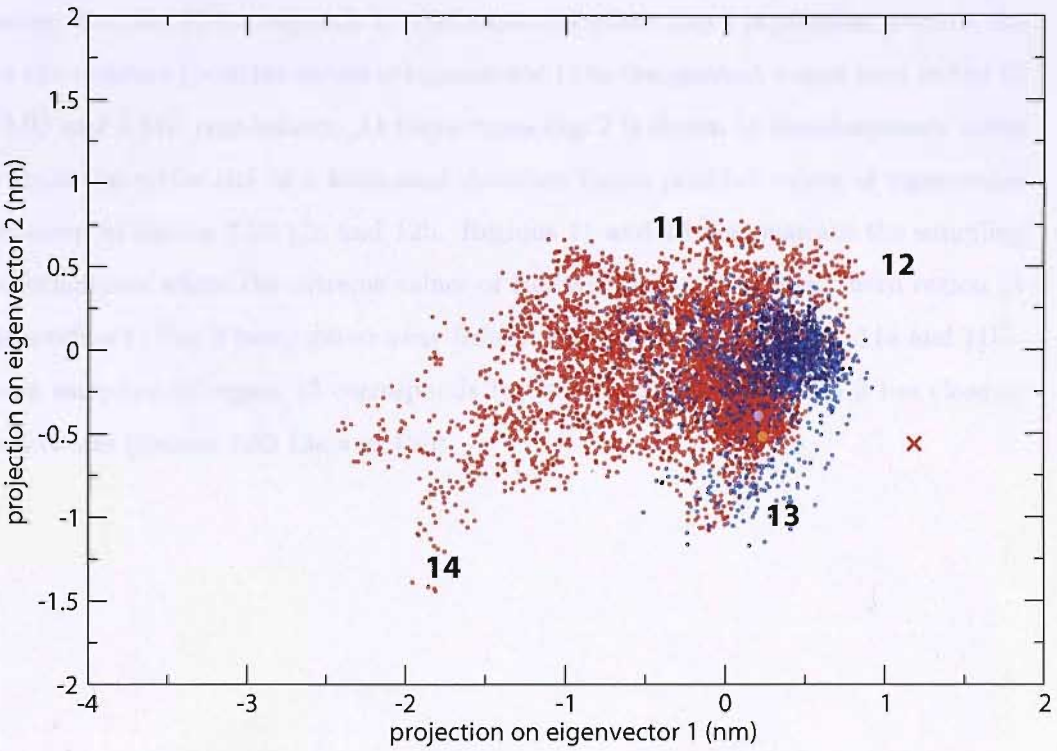
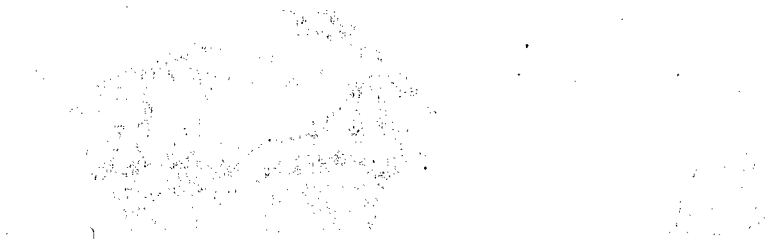
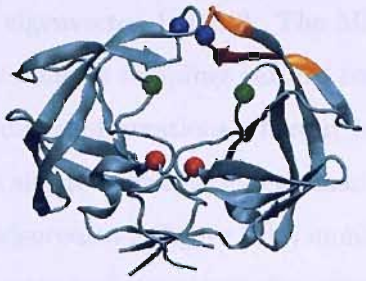


FIGURE 7.29: Sampling of the first two eigenvectors, generated from the projection of the apo V82F/I84V mutant RDFMD (red) and the two MD (blue and maroon) simulations onto the eigenvectors of the second flap created from the concatenation of all RDFMD trajectories. (The two starting structures are represented by orange and brown circles). The numbers correspond to conformations of flap 2 represented in figure 7.30. The red cross depicts the projection of the closed conformation of flap 2 of the SQV-bound structure (1HXB) onto the eigenvectors.

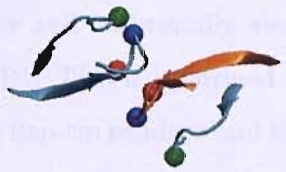
The sampling of the RDFMD simulations is shown to be diffuse, with a less defined region characterising the main area of the sampling, as seen with the sampling generated by the RDFMD simulations WT and M46I apo-systems (figures 7.21 and 7.25). Flap 2 of the RDFMD simulations of this mutant does not appear to sample any conformation which is similar to that of the closed. Instead, the RDFMD simulations are shown to largely sample regions corresponding to conformations where the flap is vertically extended away from the active site and starting conformation. Region 14 of figure 7.29 and the corresponding flap 2 conformations shown in figures 7.30 14a and b show the flap at its most vertically extended position. As these figures (figures 7.30 14a and 14b) show, vertical mobility does not necessarily allow access to the active site, the horizontal gating motion away from the active site and the conformation of flap one at that time is also important.

Sampling of region 12 corresponds to conformations where flap 2 is pointing towards the active site residues (positive values of eigenvector 1) to the greatest extent seen in the 12 RDFMD and 2 MD simulations. At these times flap 2 is shown to simultaneously move away from the active site in a horizontal direction (more positive values of eigenvector 2) as shown in figures 7.30 12a and 12b. Regions 11 and 13 demonstrate the sampling of conformations where the extreme values of eigenvector 2 are sampled, with region 11 corresponding to flap 2 being gated away from the active site (figures 7.30 11a and 11b), whereas sampling of region 13 corresponds to conformations where flap 2 lies close to the active site (figures 7.30 13a and 13b).

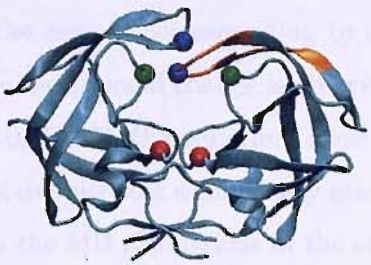




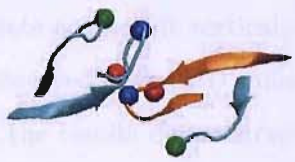
(a) 11a



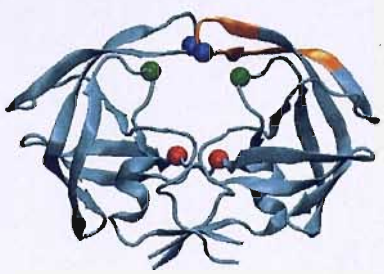
(b) 11b



(c) 12a

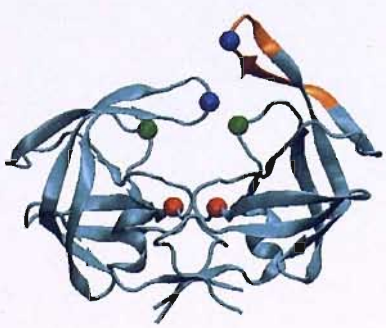


(d) 12b

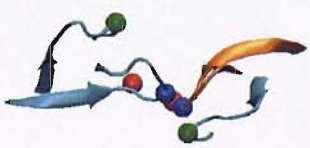


(e) 13a

(f) 13b



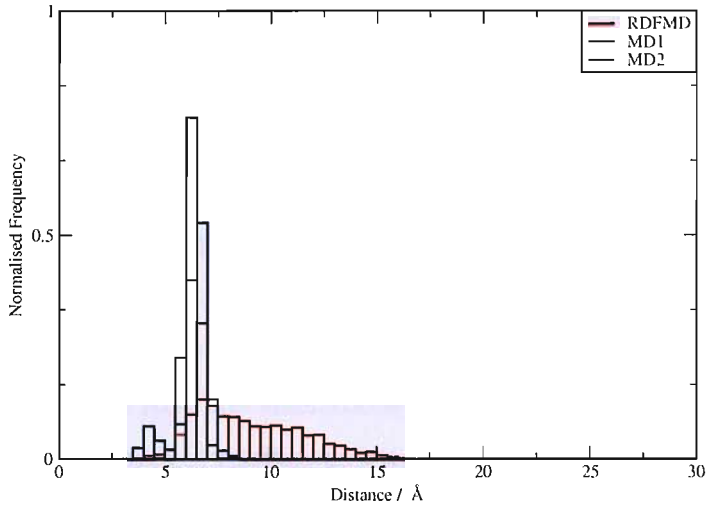
(g) 14a



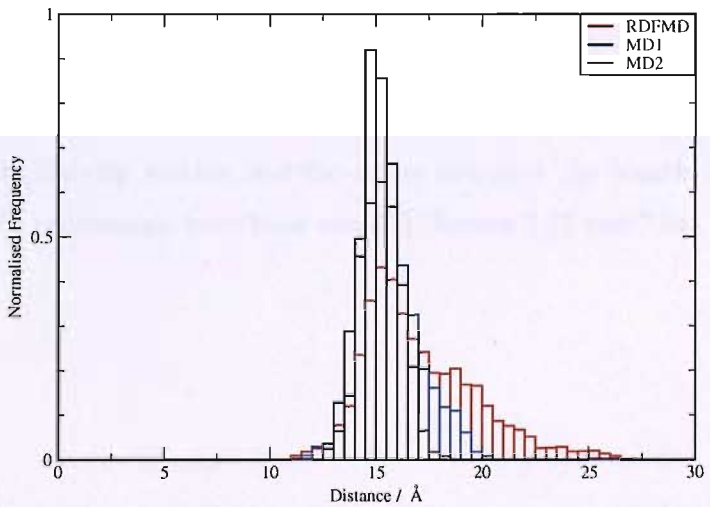
(h) 14b

FIGURE 7.30: Conformations of flap 2 of apo V82F/I84V HIV-1 from the projection of RDFMD simulations onto eigenvectors 1 and 2. Numbering corresponds to the region sampled in figure 7.29, with a: side view, b: top view. Flap 2 structure: orange, atoms highlighted in blue represent the α-carbon atoms of residue 50 and 150, red: α-carbon atoms of residues 25 and 125 of the active site and green: α-carbon atoms of residues 81 and 180 of the 80's loops

The two MD simulations are shown to overlay in their sampling of regions of conformational space of eigenvector 1 and 2. The MD1 simulation is shown to sample a larger area with a more diffuse sampling pattern compared with the MD2 simulation, showing the flap to assume conformations where it has moved vertically and horizontally away from the active site to a greater extent than is observed in MD1. This is confirmed in the distance measurements (figure 7.31) monitored between the flap-tip residues and the between the flap-tip and the active site, where MD1 is shown to sample a larger range of values compared with MD2. However, the MD simulations fail to sample the extent of the conformational space that the RDFMD simulations do, with the MD simulations sampling only the regions corresponding to a variety of semi-open conformations and no significant conformational change is observed. Together with the PCA, the distances measured over the RDFMD simulations show the flaps to separate and to lift vertically away from the active site to a significantly greater extent than observed in the MD simulations. As with the MD simulations of the other apo-systems, the results demonstrate the requirement for longer timescales to be simulated using MD to capture the events observed using RDFMD.



(a) residues 50-150



(b) residues 50-150

FIGURE 7.31: Distance measured between the α-carbon atoms of residues 50 and 150 over the 12 RDFMD simulations, MD1 and MD2 of V82F/I84V HIV-1 PR. For each of the MD simulations and the total 12 RDFMD simulations, the total number of distances measured has been normalised to a sum of 1.

Comparison of WT and Mutant Projections onto Eigenvectors 1 and 2

MD Simulations

Owing to the significantly reduced conformational sampling of flap 2 by the MD simulations, comparisons between the WT and mutants have not been made. This has been demonstrated by the reduced amount and the varying regions of conformational space sampled when MD1 and MD2 have been projected onto eigenvectors 1 and 2, compared with the sampling of the RDFMD simulations. This is also evident from the reduced range of distances measured between the flap-tip residues and the flap-tip and the active site catalytic aspartic residue.

RDFMD Simulations

Figure 8.22 overlays the sampling of the WT and mutant RDFMD simulations onto the first two eigenvectors to highlight any differences seen between the WT and mutants. Additionally, the histograms displaying the distances measured between the flap-tips and between the flap-tip residue and the active site over the length of the WT and mutant RDFMD simulations have been overlaid (figures 7.33 and 7.34).

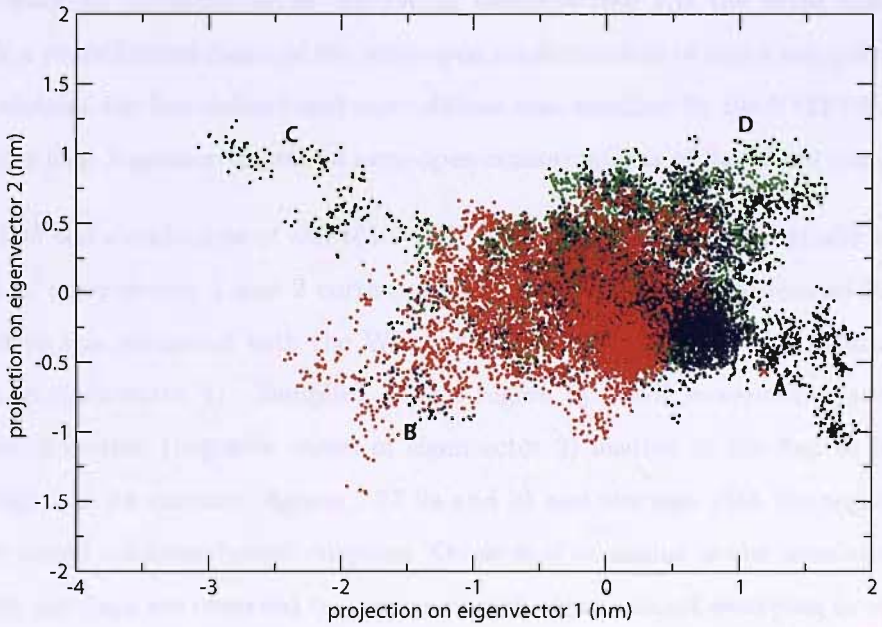


FIGURE 7.32: Sampling of the first two eigenvectors, generated from the projection of the apo WT (green), M46I (black) and V82F/I84V (red) RDFMD onto the eigenvectors 1 and 2 generated using the concatenated RDFMD trajectories. The red cross shows the location of the projection of the coordinates of flap 2 in the closed conformation of the HIV-1 SQV bound WT structure (pdb code 1HXB). Regions labelled A to D mark the extreme areas of sampling

The 2-dimensional plot clearly demonstrates a significant difference in the conformational sampling of flap 2 for the WT and mutant simulations.

A comparison of the main areas of the sampling for the second flap of each of the apoenzymes demonstrates the RDFMD simulations of the M46I mutant to have the most restrained sampling distribution (figure 7.25) compared with the WT system (figure 7.21), whereas the V82F/I84V mutant demonstrates a less well defined main region, the sampling being more diffuse (figure 7.29). For each of the WT and mutant systems, these described main regions of sampling observe only semi-open conformations. However, the distribution of the sampling in this region indicates flap 2 of the M46I apoenzyme to sample a more limited range of the semi-open conformations of flap 2 compared with the WT, whereas the less defined and more diffuse area sampled by the V82F/I84V mutant indicates that a greater variety of semi-open conformations of flap 2 are sampled.

The RDFMD simulations of the M46I HIV-1 PR apo-system preferentially sample the region of eigenvectors 1 and 2 corresponding to the flap being positioned further into the active site compared with the WT and V82F/I84V RDFMD simulations (positive values of eigenvector 1). Sampling of this region by M46I incorporates simultaneous horizontal motion (negative values of eigenvector 2) leading to the flap to be overlaid with flap 1 at its extreme (figures 7.27 9a and b) and overlaps with the region sampled by the closed inhibitor-bound enzymes. On several occasions in the simulations of this mutant, the flaps are observed to overlay on each other, almost swapping in orientation, without lifting notably away from the active site (figure 7.27). The flaps of the inhibitor-bound enzymes persist mainly in the conformation as shown in figure 6.3 ([b] and [d]), where they are positioned slightly downwards towards the active site and are swapped in orientation to each other in comparison to the apo WT HIV-1 PR. Although the RDFMD simulations do not sample this conformation to the same extent, probably due to the absence of a ligand, the most comparable conformations in the apostructures studied here are observed with the M46I mutant. This region is significantly less sampled by the simulations of the WT, and not at all by those of the V82F/I84V mutant. Where the WT samples flap conformations where the flap is positioned downwards, into the active site (region 4 on figure 7.21 and figures 7.23 4a and 4b), simultaneous sampling of the second eigenvector (positive values of eigenvector 2) results in conformations where the flap moves horizontally away from the active site, in the opposite direction needed

to achieve closed-type conformations. This type of conformation can lead to access to the active site cavity, providing the other flap is in an appropriate conformation also.

In addition, relatively reduced sampling of regions B and C in figure 8.22, corresponding to the vertical motion of the flap away from the active is observed in the simulations of the apo M46I HIV-1 PR, compared to the WT and V82F/I84V simulations. This is also demonstrated in the relatively greater sampling of the smaller flap-tip to catalytic residue distances (α-carbon of the flap tip Gly150 - α-carbon of the active site Asp125) compared with the WT and V82F/I84V simulations (figure 7.33). Sampling of eigenvectors 1 and 2 also demonstrate the simulations of this mutant to not observe horizontal motion of the flap to the extent seen in the simulations of the other apo-systems. Although this mutant samples an overall similar range of horizontal motion (eigenvector 2), the bulk of the sampling of this eigenvector lies in a narrow range, relative to the other apo-systems, which corresponds to the flap to preferentially lie over/near the active site. This is also evident in the distances between the flap-tips. Figure 7.34 shows the flaps in the apo M46I HIV-1 PR simulations to not move as far apart as seen in the WT and V82F/I84V simulations, and to largely sample the smaller distances, indicating that the flaps do not move away from each other, with this mutant having reduced access to the active site. This data reinforces the conclusion that the flap of this mutant preferentially positions itself close to the active site, with a relatively smaller population of conformations which would result in possible access to the active site, as observed in the analysis of the PCA.

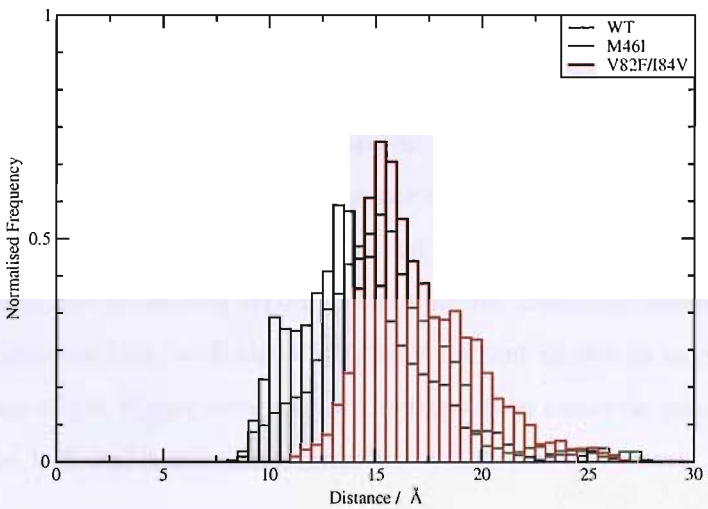


FIGURE 7.33: Distance measured between the α-carbon residues of 150 and 125 (flap 2) over the 12 RDFMD simulations

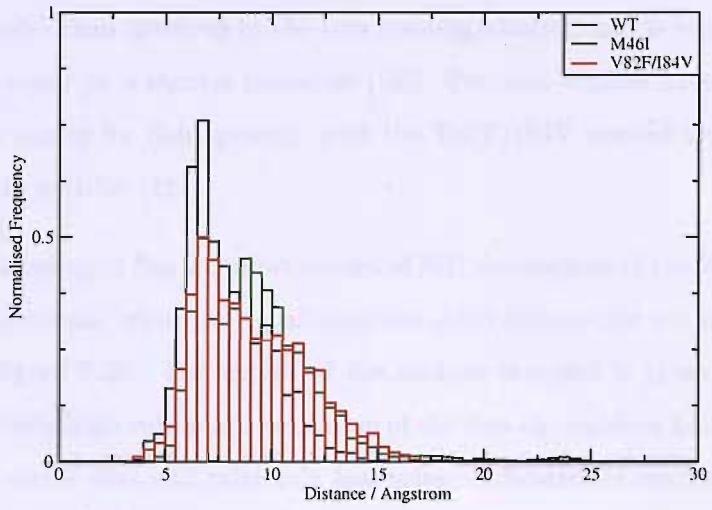


FIGURE 7.34: Distance measured between the α-carbon atoms of residues 50 and 150 over the 12 RDFMD simulations.

PCA shows the V82F/I84V mutant to sample a greater number of conformations where the flap vertically extends away from non-flap residues compared with the WT and M46I simulations. Although, the WT simulations are shown to sample the lowest values of eigenvector 1, corresponding to the flap lifting the furthest away, this is a rare event, and is only observed in one of the WT RDFMD simulations. Distance measurements (figure 7.33) also demonstrate this, with the V82F/I84V mutant shown to sample a relatively larger proportion of the bigger distances of the flap-tip to catalytic aspartate distances compared to the WT and M46I simulations.

In summary, the analysis of the M46I RDFMD simulations indicate preferential sampling of closed-type conformations, with reduced mobility of the flap. The simulations of the V82F/I84V mutant do not sample closed-type conformations, with the sampling of the RDFMD simulations corresponding to semi-open and open conformations of flap 2 only and indicating this mutant to possess increased mobility compared with the WT.

7.3.2.2 Analysis of Eigenvector 3

Extent of Flap-Tip Motion

The flap-tip curling motion is captured by the third principal component, as visualised in figures 7.18 (g) and (h). As mentioned earlier in the discussion of the literature (chapter 6.2), this motion involves fewer flap residues (residues 49-53 and 149-153 of flap 1 and

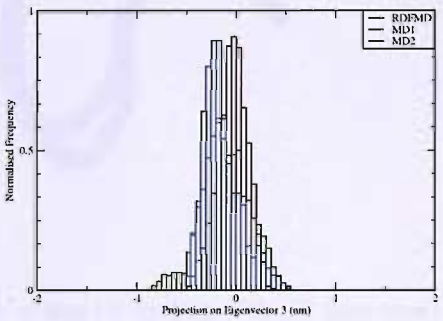
flap 2 respectively) than involved in the flap opening/closing, and is reported by NMR experiments to occur on a shorter timescale [192]. Previous studies have suggested this motion to be a trigger for flap opening, with the V82F/I84V mutant reported to show increased flap-tip activity [221].

The extent of sampling of flap 2 conformations of MD simulations of the WT and mutant HIV-1 PR apo-systems, when projected onto the third eigenvector are displayed in the histograms in figure 7.35. The extent of the motion sampled is given on the x-axis, where the relatively high values are indicative of the flap-tip residues being curled-back, away from the active site, and relatively low values indicative of conformations where the flap-tip residues are curled inwards.

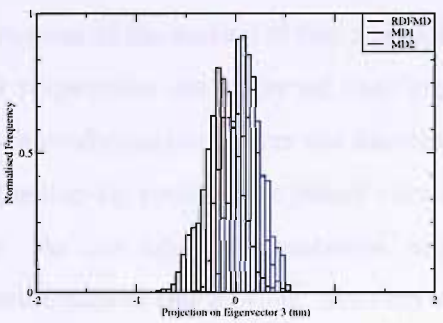
Comparison of MD1 and MD2 simulations for each of the apo-systems show in most cases sampling of different regions of eigenvector 3. As concluded from the projection of the MD simulations onto the first two eigenvectors, the MD simulations are indicated to have become trapped in energy minima, suggesting that the conformations they sample not to be representative of the full range of accessible conformations. For this reason, the results of the MD simulations have not been used to make judgements on the extent of the sampling of the flap-tip curling motion between the WT and mutants.

(a) uncurled: negative values

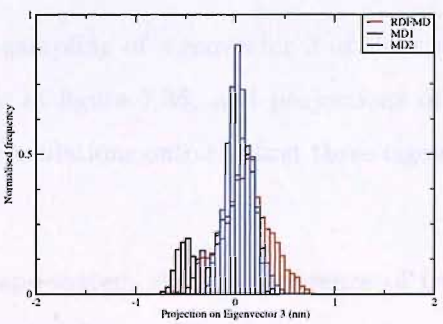
(b) curled-back: positive values



(c) apo WT



(d) apo M46I



(e) apo V82F/I84V

FIGURE 7.35: Sampling of the third principal component by RDFMD simulations of the WT, M46I and V82F/I84V apo-systems. Eigenvectors generated from a covariance matrix from the concatenation of the all the apo-HIV-1 PR simulations. Figures (a)-(b) show conformations corresponding to positive and negative projection values of eigenvector 3.

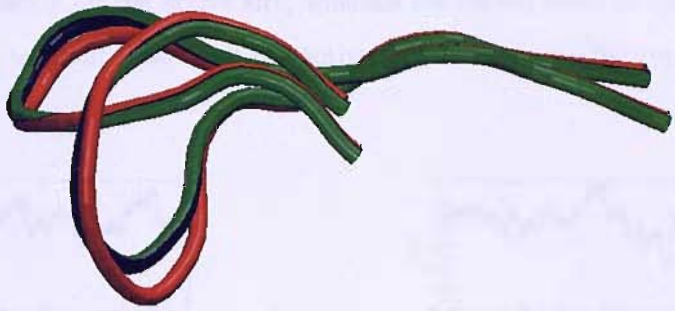


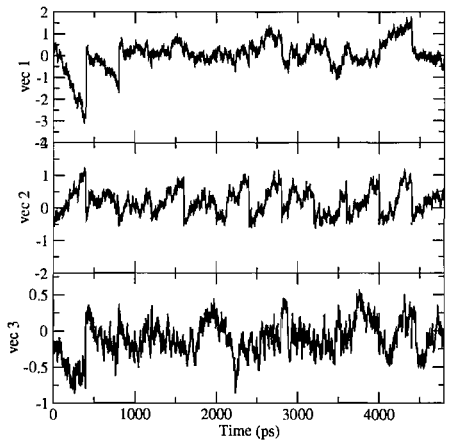
FIGURE 7.36: Extremes of flap tip curling of flap 2 (top view) in the WT and mutant apo systems. Blue: WT, green : M46I, red: V82F/I84V

Figure 7.36 displays the extremes of the motion of flap 2 sampled when the concatenated WT and mutant RDFMD trajectories are projected onto eigenvector 3. Projection of the simulations demonstrate conformations where the flap-tips are in an uncurled state at one extreme, to where the flap-tip residues are curled outwards, away from the active site at the other extreme. As this figure demonstrates, only subtle differences exist between the extreme conformations of this motion. The only slight difference is that the flap-tip residues in the simulations of the V82F/I84V mutant are seen to curl back to a slightly greater extent than what is observed in the RDFMD simulations of the WT and M46I mutant.

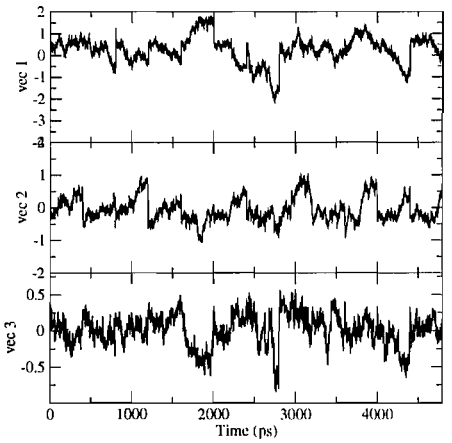
Histograms depicting the sampling of eigenvector 3 of the apo WT and mutant HIV-1 PR simulations are shown in figure 7.35, and projections of the trajectories over the length of the 12 RDFMD simulations onto the first three eigenvectors are given in figure 7.37.

Simulations of the M46I apo-system show a preference of flap 2 in an uncurled state (negative values of eigenvector 3), positioned close to the active site, compared with the WT. Representative conformations of this type are shown in figures 7.38 (a)-(d) (where conformation depicted in (a)-(b) occur at 2757 ps and (c)-(d) at 2959 ps in

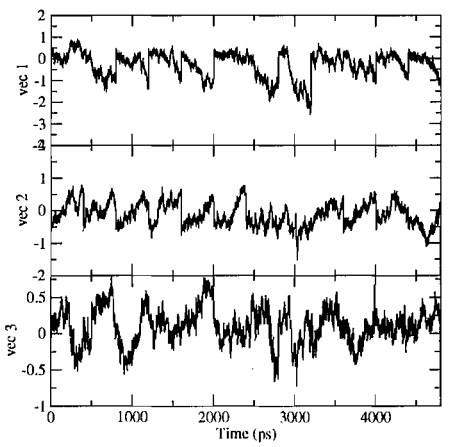
figure 7.37(b)), where (a)-(b) shows the most uncurled, and (c)-(d), the most curled conformation attained during the RDFMD simulations. In both cases, the flaps are positioned to lie closely over the active site, with the flap-tip residues of figure (a)-(b) shown to overlay directly on the active site, whereas the curled state in figures (c) -(d), the flap-tip is shown to point away from the active site, towards the hydrophobic cluster



(a) apo WT



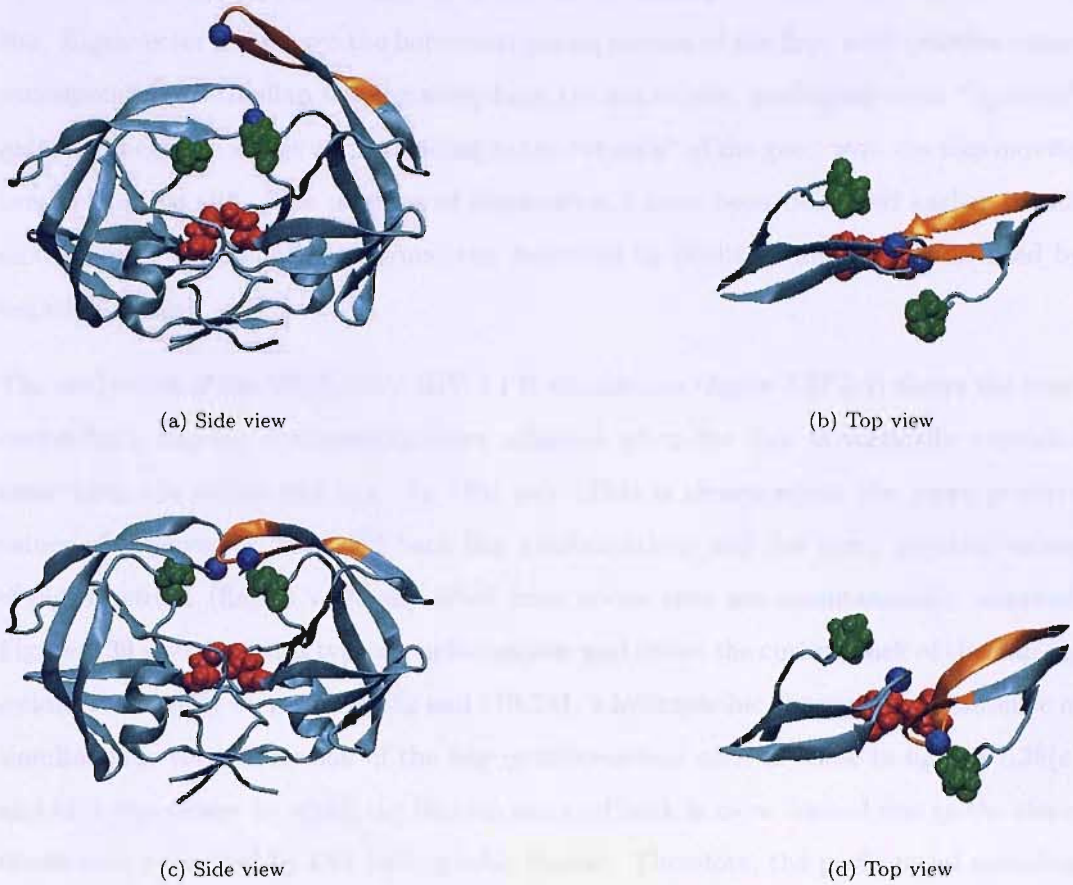
(b) apo M46I



(c) apo V82F/I84V

FIGURE 7.37: Projection of WT and mutant RDFMD simulations onto the first three eigenvectors. Eigenvectors generated from a covariance matrix of the trajectories of all the RDFMD simulations of the HIV-1 PR apo-systems.

The RDFMD simulations of the V82F/I84V mutant appear to sample conformations



(a) Side view

(b) Top view

(c) Side view

(d) Top view

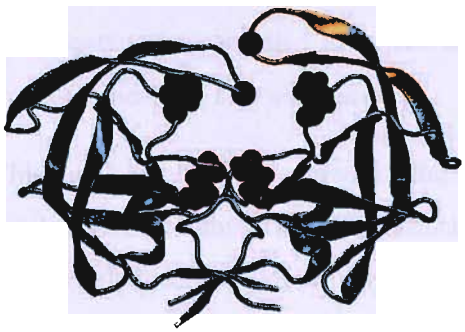
FIGURE 7.38: Conformations of flap 2 when concatenated RDFMD simulations of the M46I apo-system are projected onto eigenvector 3 ((a) and (b): flap-tips in uncurled conformation. (c) and (d): curled conformation of flap-tips). Residues in blue: Flap tip (residue 50 and 150 in flaps 1 and 2 respectively), residues in red: active site catalytic aspartic acid (25 and 125 in flaps 1 and 2 respectively), residues in green: proline of hydrophobic cluster (81 and 181 in flaps 1 and 2 respectively).

where flap 2 is slightly more curled back than those seen in the WT simulations (figure 7.36) , and also display a preferential sampling of this type of conformation, where the flap-tip residues curl back away from the active site (figure 7.35 (e)). Consequently, there is lesser sampling of the uncurled flap-tip conformations compared with the WT and M46I simulations.

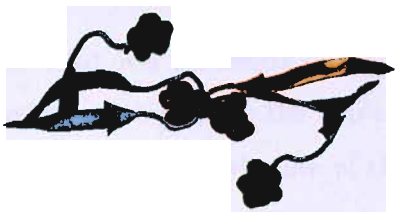
Figure 7.37 shows the projection of the concatenated RDFMD simulations of each of the WT and mutant systems onto the first three eigenvectors. The conformations achieved by flap 2, described by the first two eigenvectors has been carried out using 2-dimensional plots (figures 7.21 7.25 7.29). As described earlier, eigenvector 1 captures the vertical movement of the flap, with positive values corresponding to the flap in a position pointing

towards the active site, and negative values, where the flap is lifted away from the active site. Eigenvector 2 captures the horizontal gating motion of the flap, with positive values corresponding to the flap moving away from the active site, analogous to an "opening" gate, and negative values corresponding to the "closing" of the gate, with the flap moving over the active site. The motions of eigenvector 3 have been described earlier in this section, with curled-back conformations described by positive values, and uncurled by negative values.

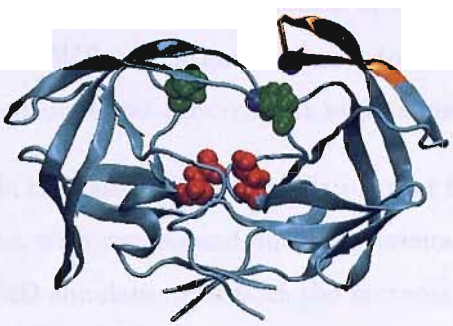
The projection of the V82F/I84V HIV-1 PR simulations (figure 7.37 (c)) shows the most curled back flap-tip conformations are achieved when the flap is vertically extended away from the active site (e.g. at 1951 ps). This is shown where the more positive values of eigenvector 3 (curled back flap conformation) and the more negative values of eigenvector 1 (flap is vertically lifted from active site) are simultaneously sampled. Figure 7.39 visualises this type of conformation, and shows the curling back of the flap-tip residues to overlay on residues 132 and 179-181, a hydrophobic cluster. In the absence of simultaneous vertical motion of the flap (conformations such as those in figures 7.38(c) and (d)), the extent to which the flap-tip can curl back is more limited due to the steric hinderance presented by this hydrophobic cluster. Therefore, the preferential sampling of the curled-back conformations of the V82F/I84V simulations compared with the WT and M46I simulations may be a consequence of the increased mobility of the flap of this mutant. As analysis of eigenvector 1 (representing vertical motion of flap residues) shows, along with the flap-tip to active site distance measurements, flap 2 of the RDFMD simulations of the V82F/I84V mutant samples a greater population of conformations where the flap is lifted vertically above the active site compared to simulations of the WT and M46I mutant. In contrast, projection of the RDFMD simulations of the M46I apo-system onto the first two eigenvectors, in addition to distance measurements, shows this mutant preferentially samples conformations where flap 2 remains close to the active site. This would result in this mutant being less able to sample the same extent of the curled-back conformations as seen in the V82F/I84V simulations, as the nearby hydrophobic cluster would prevent this.



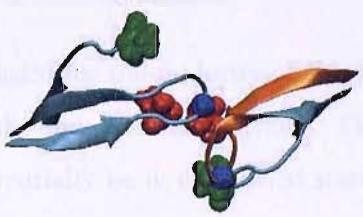
(a) Side view



(b) Top view



(c) Side view



(d) Top view

FIGURE 7.39: Conformations of flap 2 of apo V82F/I84V representing the different conformations sampled by the third eigenvectors. ((a) and (b): flap tip in uncurled conformation, (c) and (d): curled-back conformation of flap-tip. Residues in blue: Flap tip (residue 50 and 150 in flaps 1 and 2 respectively), residues in red: active site catalytic aspartic acid (25 and 125 in flaps 1 and 2 respectively), residues in green: proline of hydrophobic cluster (81 and 181 in flaps 1 and 2 respectively).

7.3.2.3 Frequency of Flap-Tip Curling

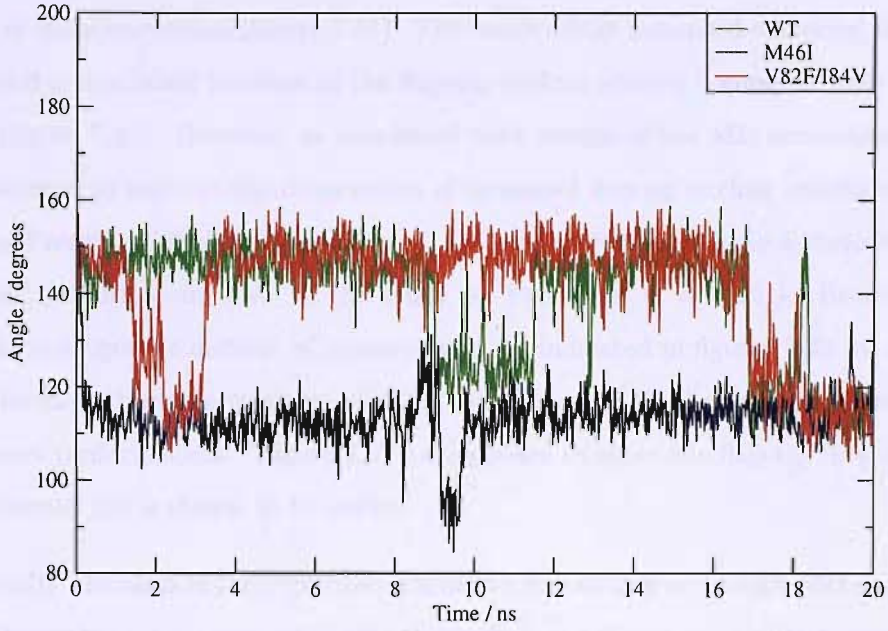
A previous theoretical study by Perryman *et al.* [221], as mentioned in the background section 6.2 showed the V82F/I84V HIV-1 PR mutant to display increased flap-tip curling activity compared with the WT. The frequency of flap-tip curling was monitored through measurement of the angle between the α-carbon atoms of the G48-G49-I50 (flap 1) and G148-G149-I150 (flap 2) flap tip residues. For the simulations of each of the apostructures studied here, the same angle has been monitored over the length of the MD and RDFMD simulations.

This angle has been measured for flap 1 and flap 2 (figure 7.40) over the length of MD1, the longer MD simulation. The relatively lower angle values correspond to the flap-tip being curled, and the relatively higher values correspond to the uncurled state of the flap.

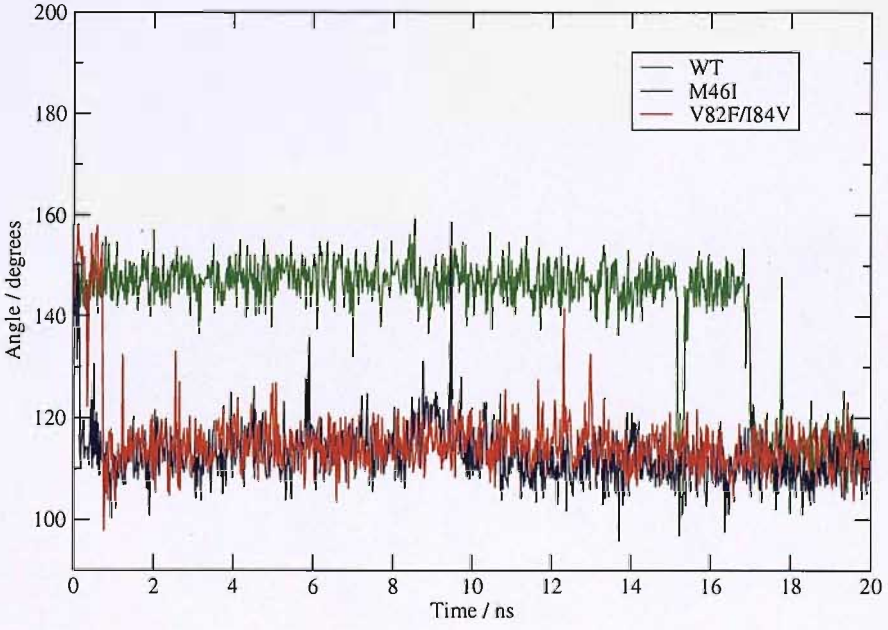
For flap 1, the MD simulation of the M46I HIV-1 PR is shown to remain in a mostly curled state for the duration of the whole simulation. The same flap in the V82F/I84V and WT simulations is shown to remain mostly in the uncurled conformation, with an curling event occurring at approximately 9 ns and 17 ns in each simulation.

In the case of flap 2, the flap-tips of the WT remain uncurled for the majority of the 20 ns, with curling and uncurling events occurring towards the end of the simulations. The MD simulations of both the mutants are shown to preferentially be in the curled state, with infrequent uncurling events.

There is no conclusive evidence attained from the MD simulations carried out here, concerning variations in the frequency of flap-tip curling between the WT and mutants, although the simulation is of similar length (22 ns in Perryman *et al.* study) as that of the theoretical study which identified this.



(a) flap 1

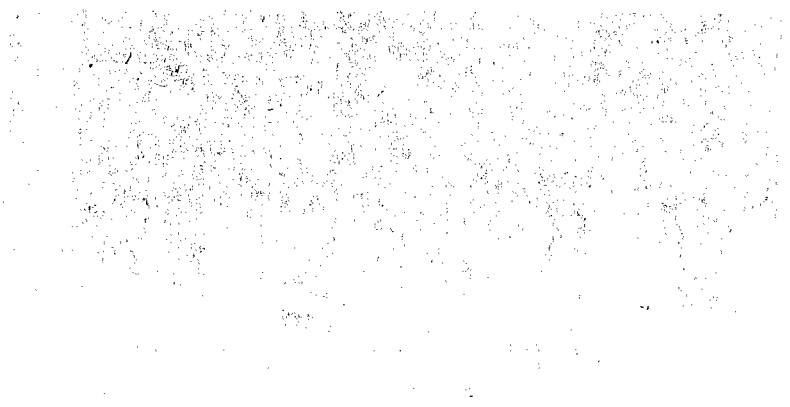


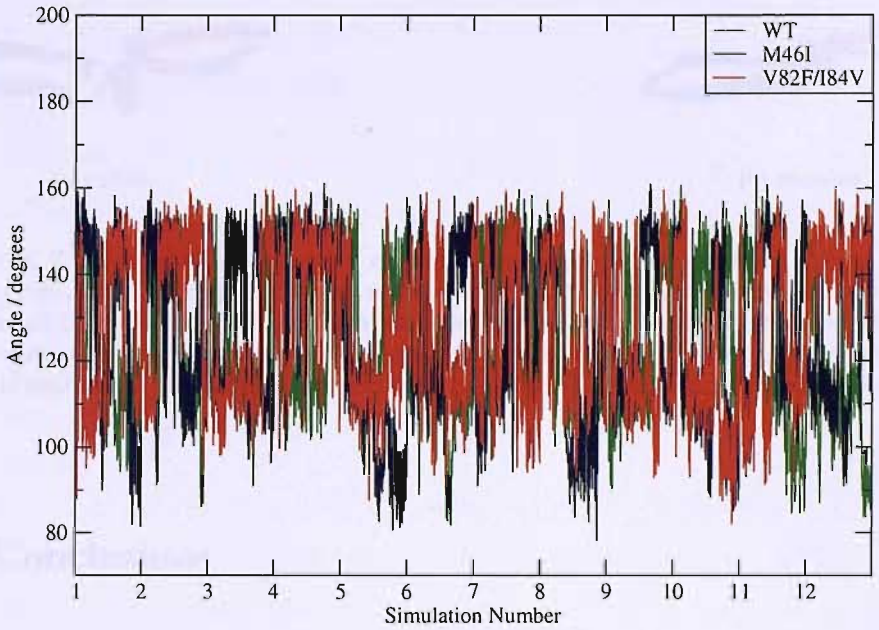
(b) flap 2

FIGURE 7.40: Measurement of angle between the α-carbon atoms of [a] G48-G49-I50 (flap 1) and [b] G148-G149-I150 (flap 2) flap-tip residues over the length of MD1 for each of the apo HIV-1 PR systems

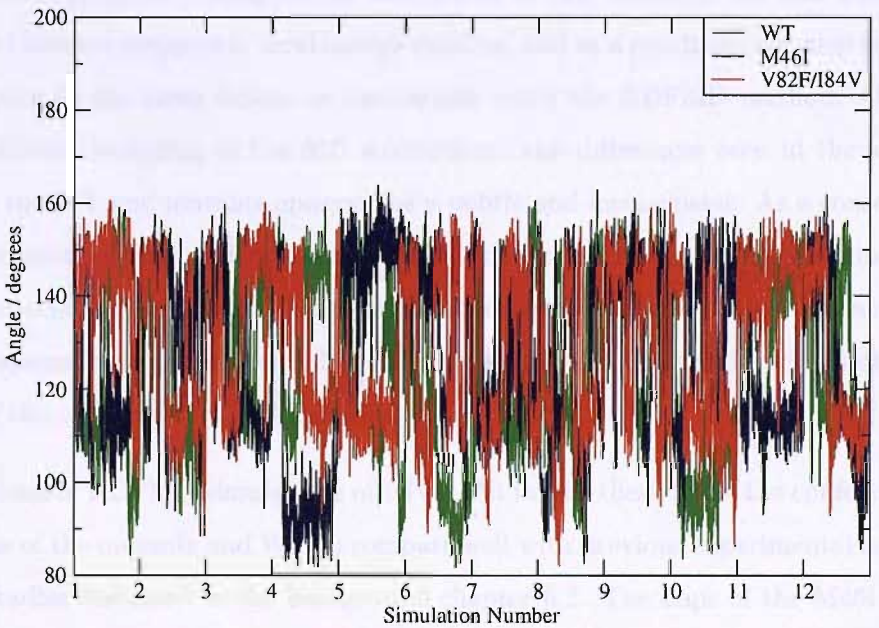
The same angle was measured over the length of the concatenated RDFMD simulations of each of the apo-systems (figure 7.41). The result of the enhanced sampling technique has caused a significant increase in the flap-tip curling activity compared with the MD results (figure 7.40). However, as concluded with results of the MD simulations, there is no evidence to support the observation of increased flap-tip curling concluded in the study by Perryman *et al.* [221]. Angles of 115° are stated to indicate a curled-in state, and 145°, a curled-out state in the study by Perryman *et al.* [221]. However, this may not be a rigorous method of measurement, as indicated in figures 7.42 (a) and (b), where the angle between residues 148-149-150 are both 140°, but the flaps are clearly in different conformations. Figure 7.42 (a) appears to show the flap-tip in a uncurled state, whereas (b) is shown to be curled.

Additionally, the basis of their conclusion is shown for one flap on a single MD simulation. The MD simulations carried out in this thesis show significant variability between MD1 and MD2, so their results may be a consequence of an artefact of the MD simulation, rather than being a true result.





(a) flap 1



(b) flap 2

FIGURE 7.41: Measurement of angle between the α-carbon atoms of [a] G48-G49-I50 (flap1) and [b] G148-G149-I150 (flap 2) flap-tip residues over the 12 RDFMD simulations of the apo-systems.

(a) curled (b) uncurled

FIGURE 7.42: Curled and uncurled conformations of flap 2 (coloured in orange, flap 1 in cyan), both with approximate values of 140° for the angle between the α-carbon atoms of G148-G149-I150. α-carbon atoms of the catalytic aspartate residues in the active site (residues 25 and 125) are represented by red-coloured balls. α-carbon atoms of the residues of the flap-tips (residues 50 and 150) are depicted by orange-coloured balls.

7.4 Conclusions

The results of the MD simulations presented here show the flaps of the WT, M46I and V82F/I84V HIV-1 apo protease simulations to undergo conformational change. PCA demonstrates the significant motions of the flaps to be observed in the RDFMD simulations. However, owing to the limitations of MD method, the MD simulations appear to become trapped in local energy minima, and as a result are not able to sample phase space to the same extent as simulations using the RDFMD method. Owing to the insufficient sampling in the MD simulations, any differences seen in the sampling between the WT and mutants aposystems is subtle and inconclusive. As a consequence, the conformations sampled by the RDFMD simulations have been studied in greater detail, especially flap 2, which has been demonstrated to undergo a greater amount of conformational change relative to flap 1. This is possibly due to the differing protonation states of the catalytic aspartic acid residues.

The analysis of RDFMD simulations of HIV-1 PR in this thesis show the conformational dynamics of the mutants and WT to compare well with previous experimental and theoretical studies discussed in the background chapter 6.2. The flaps of the M46I mutant appear to be less mobile than the WT and V82F/I84V mutant apo HIV-1 PR enzymes, and sample an increased number of conformations where the flaps are in a similar orientation to those observed in the closed crystal structure. The limited flexibility of this mutant results in the reduced population of conformations sampled which confer access

to the active site. Experimental results postulate the reduced flexibility to be due to the increased size of the isoleucine residue, compared with the methionine residue, which results in the restricted movement of the flaps due to steric hinderance [218].

The V82F/I84V mutant is shown to be more flexible, sampling a greater range of semi-open conformations where the flap is vertically lifted away from the active site to a greater extent than seen in the simulations of the WT and M46I mutant, and is less likely to sample the closed-type conformations. The population of conformations which have the potential to provide access to the active site appear to be of a similar number to that of the WT, judged by the flap-tip to flap-tip distances measured.

The differences in the conformational dynamics of the WT and mutant enzymes are of biological relevance. The M46I mutant is a commonly occurring secondary mutation, which is predicted to recover catalytic activity lost by the presence of a primary mutation (e.g. V82F/I84V). The mechanism by which it does this is predicted to be through stabilisation of the flaps in the closed conformation, thus reducing the enthalpic cost in closing the flaps in order for substrate to be processed by the enzyme. The V82F/I84V mutant is a primary mutant, reducing the efficacy of all current protease inhibitors (PIs) and also reducing the catalytic efficiency of the enzyme for substrates. The flaps of this enzyme are reported in previous studies to be highly mobile, with increased flap tip curling activity [221]. The increased frequency of flap-tip curling is not confirmed in the results here of the V82F/I84V simulations, but this mutant is shown to preferentially sample conformations where the flap-tips are curled back, away from the active site to a greater extent than observed in the WT and M46I simulations. To the author's knowledge, this result has not been reported in previous literature. The ability of this mutant to be able to curl back further is postulated to be due to the increased sampling of the conformations where the flap lifts vertically above the non-flap residues of the protein, away from a hydrophobic cluster of residues, which would normally hinder the event if curling back occurs with no vertical motion.

The diminished catalytic activity and resistance to PIs is thought to be a consequence of a higher enthalpic cost by this enzyme to close the flaps for its function. For both of these mutants, there is a shift in the equilibrium, with the V82F/I84V favouring semi-open conformations, with reduced sampling of the closed conformation, and the M46I

mutant, whose reduced mobility contributes to preferential sampling of more closed-type conformations.

RDFMD has assisted in this study as the increase in conformational sampling it has achieved over MD has been vital in reaching these conclusions. MD alone is not sufficient to sample enough of phase space in order to be able to compare the dynamics between the WT and the mutant enzymes.

Chapter 8

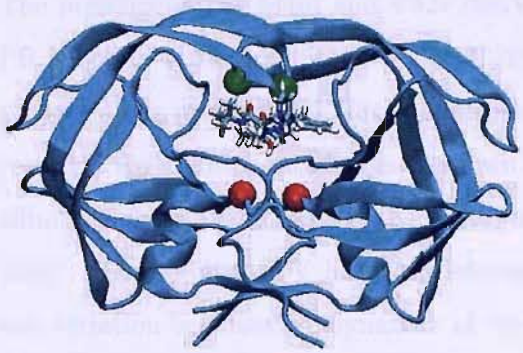
Human Immunodeficiency Virus-1 Protease (HIV-1 PR): Part 2: Inhibitor-bound

8.1 Aims

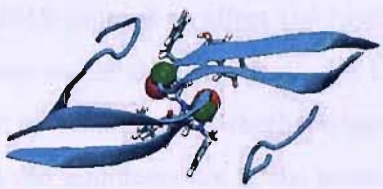
The aims of this chapter are to study the conformational dynamics of the WT and mutant HIV-1 PR enzymes when bound to inhibitors, using conventional MD and the enhanced sampling technique of RDFMD.

The mutations chosen for study are M46I and V82F/I84V, in addition to the WT system, when the HIV-1 PR enzyme is bound to the inhibitors indinavir (IDV) and saquinavir (SQV) (figure 8.1). The mutants studied in this chapter are the same as those studied in the apo HIV-1 PR systems in this thesis (chapter 7).

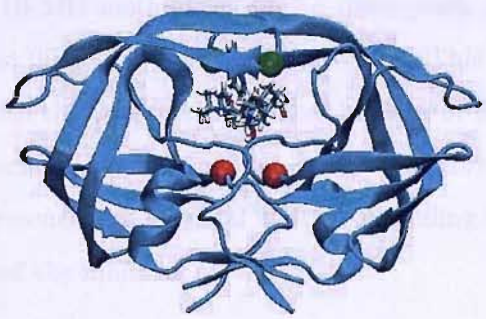
As previously discussed in chapters 7 and 6, these mutations are thought to affect the dynamics of the flaps, altering the equilibrium between the open and closed conformations, thus altering the dissociation rates of drugs. Clinical studies show the V82F/I84V mutant to be a primary mutant, having a detrimental effect on the binding affinity of all the protease inhibitors (PIs) currently in clinical use, including saquinavir and indinavir [241] [242]. The M46I mutant exhibits a compensatory role. This mutant does not affect



(a) IDV-bound



(b) IDV-bound



(c) SQV-bound

(d) SQV-bound

FIGURE 8.1: Crystal structures of IDV- and SQV-bound HIV-1 PR. [a]: side view of HIV-1 PR bound with IDV (1HSG); [b]: top view of HIV-1 PR bound with IDV (1HSG) [384]; [c]: side view of HIV-1 PR bound with SQV (1HXB) [383]; [d]: top view of HIV-1 PR bound with SQV (1HXB). Top view representations show flap and P1 residues only (residues 44-59, 76-83, 144-159 and 176-183). Atoms coloured green represent the α-carbon atoms of residues 50 and 150 in monomers A and B respectively. Atoms coloured red represent the α-carbon atoms of residues 25 and 125 of monomers A and B respectively. In HIV-1 PR IDV-bound crystal structure, the distances measure between α-carbon atoms 50 and 150: 5.74 Å, 50 and 25: 13.78 Å and 150 and 125: 14.12 Å. In the HIV-1 PR SQV-bound crystal structure, distances measured between α-carbon atoms 50 and 150: 6.19 Å, 50 and 25: 14.48 Å and 150 and 125: 14.39 Å.

the binding affinity of inhibitors, but is thought to partially restore the activity of the enzyme from the effects of primary mutation, allowing the protein to possess a sufficient level of functionality to carry out the processing of substrates, while not being inhibited. [243] [244] [245] [246] [247] [248].

The presence of the M46I and V82F/I84V mutations have been reported when HIV-1 PR is bound to the inhibitor SQV [277] [278] [279], whereas in the case of IDV binding, the existence of the V82F/I84V double mutant, but not the M46I mutant has been recorded. In addition, studies have shown the V82F/I84V mutant to affect the binding affinity of the SQV-bound HIV-1 PR enzyme to a greater extent than the binding of IDV [269]. The inhibitor IDV has been selected for study to demonstrate whether there is any variation between the dynamics of the flaps when the inhibitors are in the presence of effective and ineffective mutations.

Additionally, the study of these inhibitor-bound structures will evaluate whether the RDFMD method is able to distinguish differences in the conformational dynamics of the flaps of the WT and mutants. This is predicted to be harder to assess compared with the apo-systems, since the flaps of the inhibitor-bound structures are known to be significantly less mobile, being held more rigidly in the closed conformation due to the presence of additional hydrogen bonding between the flap tips and between the flap tips and the inhibitor.

8.2 System Preparation and MD Parameters for Simulation of Inhibitor-bound HIV-1 PR

A total of six systems were setup, with the WT, M46I and V82F/I84V mutant HIV-1 PR bound to saquinavir and indinavir (Systems - IDV-bound: WT, M46I, V82F/I84V, SQV-bound: WT, M46I, V82F/I84V).

The starting coordinates for the SQV-bound structures were taken from the WT 1HXB crystal structure [383] (resolution 2.3 Å), and the coordinates of the WT IDV-bound structure [384] taken from the 1HSG crystal structure (resolution 2.0 Å), both obtained from the Protein Data Bank [8].

The parameters for the inhibitors were obtained from Lepsik *et al.* [301], who carried out optimisation of the partial charges using quantum mechanics calculations and used the Antechamber utility of the AMBER7.0 [385] package to obtain the topologies of the inhibitors.

As with the apo-structures, the WHATIF [171] program was used to add polar hydrogens and to check the structure. The AMBER utility XLEAP [172] was used to add other hydrogen atoms and to solvate the system, with a minimum distance of 12 Å from the protein to the box boundary, in a box of 11370 and 10818 TIP3P [44] water molecules for the SQV-bound and IDV-bound systems, respectively. The crystal waters were included in the system set-up. Seven chloride counterions were added to each of the monoprotonated systems to neutralise the overall charge of the system.

All simulations, unless otherwise stated, have been carried out using the NAMD [170] molecular dynamics package and the AMBER99 [27] forcefield, using the SHAKE [35] algorithm to constrain all bonds involving hydrogen.

The minimisation was carried out in stages, starting with the inhibitor (1,000 steps) protein only (5,000 steps), followed by solvent (30,000 steps), ions (1,000 steps), solvent and ions (20,000 steps) and finally the entire system (40,000 steps), giving a total of 97,000 steps. Two minimisation algorithms have been used, initially the steepest descent algorithm, followed by the conjugate gradient method.

The minimised system was heated to 300 K in the NVT ensemble. The procedure employed a Langevin thermostat with a 10 ps^{-1} damping parameter and a 2 fs timestep. The heating was carried out gradually in stages at 50 K intervals, each interval being 20,000 steps long.

Equilibration simulations, using a Nosé-Hoover barostat in the NPT ensemble were carried out for 50,000 steps, with a target pressure of 1 atm. A decay parameter of 100 fs and a piston period of 200 fs were used. A further 100,000 steps were run, with a decay parameter of 300 fs and a piston period of 500 fs. These different stages of equilibration allowed the system to gradually adjust to the new conditions.

To generate the M46I and V82F/I84V mutated structures for both the IDV- and SQV-bound structures, the coordinates of the equilibrated WT system were taken, and the

appropriate residues mutated using scap [378] [379]. As with the mutated apo-structures, following residue mutation, minimisation of the altered residues were carried out for 1,000 steps whilst restraining the rest of the system, followed by the minimisation of the rest of the protein for 5,000 steps, the solvent and counterion for 1000 steps, and lastly, the entire system for 10,000 steps. Heating and equilibration of the system was then carried out in the same manner as for the inhibitor-bound WT structure.

Final equilibrated box-sizes were as follows : SQV-bound: WT 66.6, 71.1, 80.3 Å, M46I 66.5, 71.0, 80.2 Å, V82F/I84V 66.7, 71.2, 80.4 Å. IDV-bound: WT 61.9, 75.8, 77.1 Å, M46I 61.7, 75.5, 77.8 Å, V82F/I84V 61.8, 75.7, 78.0 Å.

All production MD simulations were run in the NVT ensemble. The NVT ensemble was chosen in preference to the NPT ensemble for consistency with the RDFMD simulations, which employ this ensemble to avoid system expansion. The MD simulations used Langevin thermostat with a 5 ps^{-1} damping parameter at a temperature of 300 K and a timestep of 2 fs. Periodic boundary conditions were used, along with a particle mesh Ewald treatment of electrostatic interactions, using a interpolation order of 6, and switching function applied to the Lennard-Jones interactions between 9 Å and the 10.5 Å cutoff. SHAKE [35] was applied to all bonds containing hydrogen, using a tolerance of 10^{-8} Å.

For each system, two MD production simulations have been carried out, one of 20 ns and the other 10 ns in length. RDFMD and MD simulations started from two different equilibrated structures possessing different velocities and coordinates, generated by two sets of equilibration being carried out, starting from the minimised structure at the start of the heating stage each time.

8.3 RDFMD Simulation Parameters

As with the RDFMD simulations of the apo HIV-1 PR structures, flap residues 49-51 and 149-151 have been selected as the target of the filter of RDFMD.

The same protocol of RDFMD parameters used for the apo HIV-1 PR structures has been used for the inhibitor bound simulations. The RDFMD parameters used here were

based on previous work which investigated the optimisation of a protocol of suitable parameters for the use of the RDFMD methodology to study protein systems. [231]. These included a 0-100 cm^{-1} digital filter, with 201 coefficients, a filter delay of 50 or 100 steps, a temperature cap of 800 K, amplification factors of 1.6, 1.8, 2.0, 2.2 and 2.4, and 4 ps of MD run in between the filter applications. The final trajectory was generated by piecing together the individual 4 ps MD runs. Each RDFMD simulation produced 100 4 ps MD sections, totalling 400 ps.

For each system, 12 RDFMD simulations have been carried out, six of which used a filter delay of 50 steps, and six with a 100 step filter delay. For those simulations using a 50 step filter delay, amplification factors of 1.6, 1.8 and 2.0 were used. RDFMD simulations which employed a filter delay of 100 steps used amplification factors of 2.0, 2.2 and 2.4. Each simulation employed a different combination of amplification factor and filter delay to increase the variation in the trajectory (see table 8.1). In addition, RDFMD simulations started from two different starting structures with different random number seeds, generated by two sets of equilibration being carried out for each system, starting from the minimised structure each time.

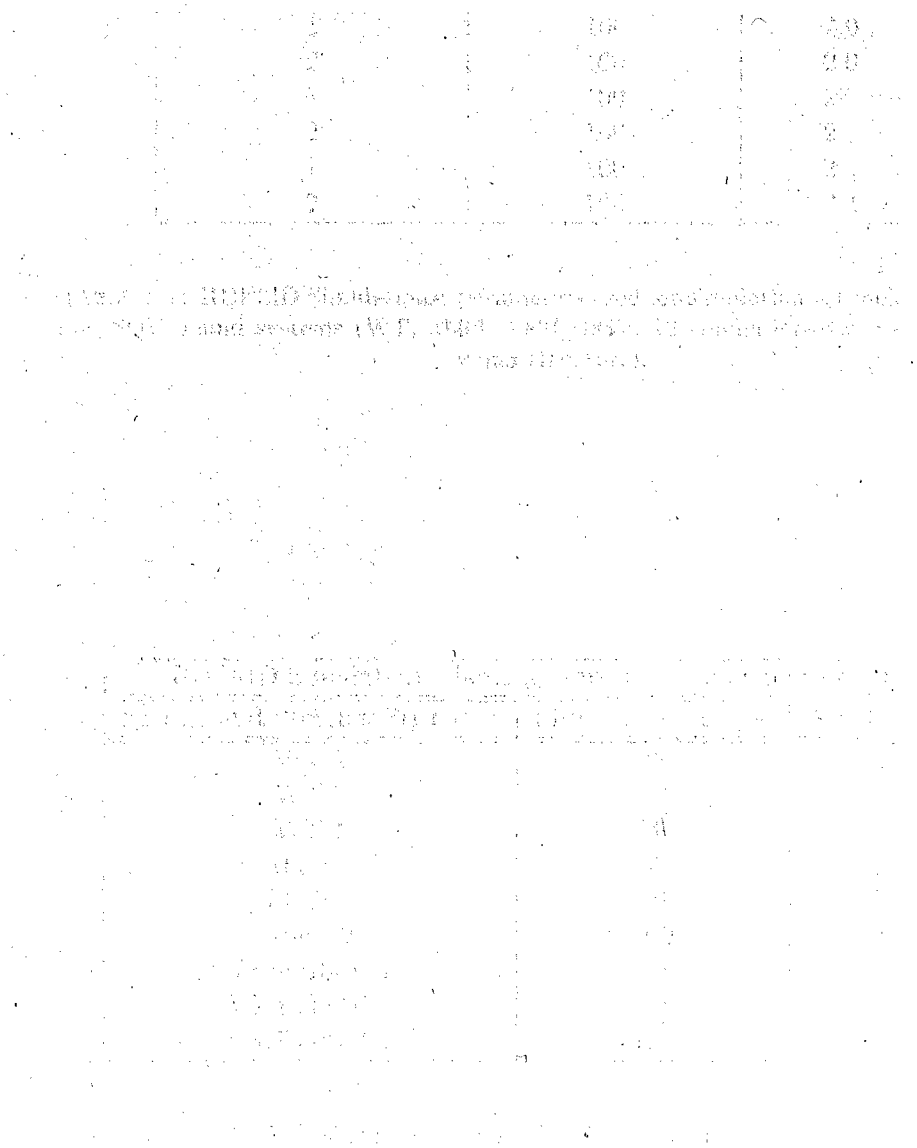
The MD sections of the RDFMD method were carried out in the NVT ensemble at 300 K, using a Langevin thermostat with a 5 ps^{-1} damping parameter and a 2 fs timestep.

8.4 Results

This study has investigated the conformational dynamics of the HIV-1 PR WT and M46I and V82F/I84V mutant systems, when bound to IDV and SQV. The selected mutants have been shown to affect the conformational dynamics of the apo-enzyme, especially in the region of the flaps in chapter 7 of this thesis. The aim of this chapter is to investigate the effects of the same mutants on the dynamics of the flaps when in the presence of the inhibitors.

RDFMD simulations of the IDV and SQV inhibitor-bound HIV-1 PR systems have been carried out using the parameters displayed in table 8.1. In total, 12 RDFMD simulations and 2 MD (MD1: 20 ns, MD2: 10 ns) simulations exist for each system. For the SQV-bound RDFMD simulations, a few of the trajectories showed the presence of unrealistic

trans to *cis* ω isomerism events of the peptide bond in a few of the residues targeted by the digital filter. The RDFMD simulations in which this occurred are shown in table 8.2. The results of these simulations have been disregarded and have not been included in the analysis presented in this chapter.



RDFMD Simulations		
Starting Structure No.	Filter Delay (steps)	Amplification
1	50	1.6
2	50	1.6
1	50	1.8
2	50	1.8
1	50	2.0
2	50	2.0
1	100	2.0
2	100	2.0
1	100	2.2
2	100	2.2
1	100	2.4
2	100	2.4

TABLE 8.1: RDFMD Simulations: parameters used for simulations of each of the IDV- and SQV-bound systems (WT, M46I, V82F/I84V, 12 simulations for each inhibitor-bound structure).

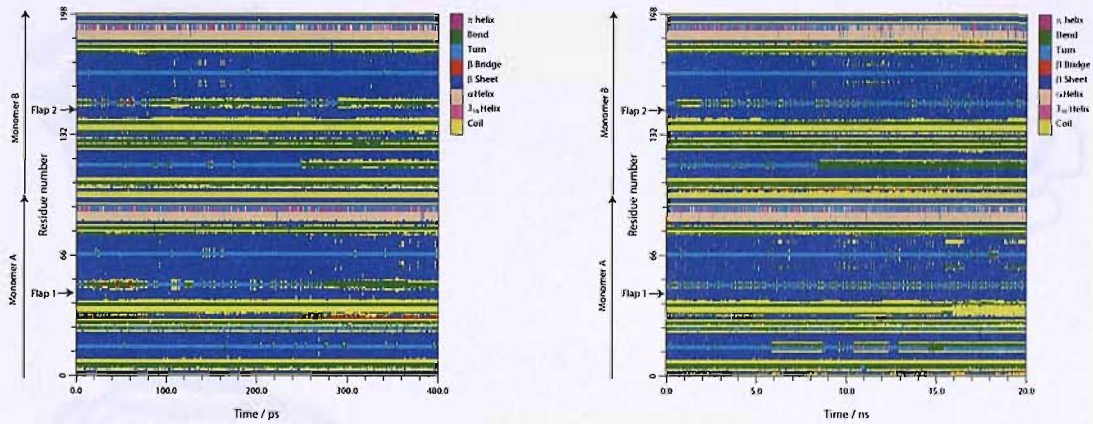
RDFMD Simulations showing trans to cis ω angle isomerism		
System and Starting Structure	Filter Delay (steps)	Amplification
WT 1	50	1.6
WT 2	50	1.6
WT 1	50	1.8
M46I 2	50	1.8
M46I 1	50	2.0
M46I 2	100	2.2
V82F/I84V 1	50	1.6
V82F/I84V 1	50	1.8
V82F/I84V 1	100	2.0

TABLE 8.2: RDFMD simulations of HIV-1 SQV-bound systems which displayed ω *trans* to *cis* transitions of peptide bonds in the residues targeted by RDFMD.

8.4.1 Structural Stability and Flexibility

The secondary structure of the inhibitor-bound HIV-1 PR WT and mutant enzymes studied here have been monitored over the length of the MD and RDFMD simulations. Figure 8.2 displays representative plots resulting from 20 ns of MD and 400 ps of RDFMD simulation of the IDV-bound WT HIV-1 PR. The secondary structure plots of the MD and RDFMD simulations of the other IDV- and SQV inhibitor-bound WT and mutant HIV-1 PR structures are similar and therefore have not been included.

The plots confirm the secondary structure of the inhibitor-bound HIV-1 PR systems to be stable throughout the simulations carried out using both MD and RDFMD methodologies. Comparison of the secondary structure plots generated by the MD simulations with those of the RDFMD simulations demonstrates that the addition of energy into the systems during the RDFMD filter application stages has not affected the secondary structure.

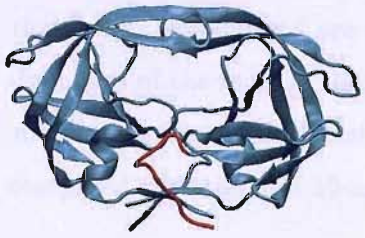


(a) RDFMD: WT

(b) MD: WT

FIGURE 8.2: Representative secondary structure plots for the IDV-bound HIV-1 PR structures. Parameters of plot generated by RDFMD: filter delay of 50 steps, temperature cap of 800 K, amplification of 2.0. Plots generated by MD simulation have been produced from 20 ns of simulation.

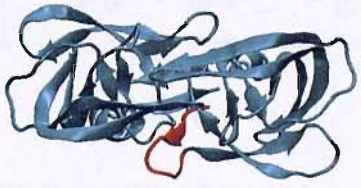
The MD1 simulation of the WT SQV-bound HIV-1 PR shows some unusual movement of the residues of the N-terminal of monomer B (residues 195 to 199). Figure 8.3 shows the conformation which this region of residues assumes at the start (figures 8.3 (a) and (b)) and the end (8.3 (c) and (d)) of the 20 ns MD simulation. The secondary structure is shown to remain intact for these residues over the duration of the MD simulation, but this region is shown to deviate from its usual conformation. In all other simulations, the overall configuration of this region is not shown to noticeably differ from that observed in the crystal structure. However, in the last 10 ns of the simulation of the WT SQV-bound HIV-1 PR system, this region is shown to move away from the rest of the protein structure and to become more extended in its shape, deforming away from the more compact shape of its usual conformation. Although this motion of N-terminal is not seen in the other simulations carried out, it has not caused deformation of the active site or other residues of the enzyme, and thus the results of the simulation have been treated as normal.



(a) 0 ns: Side view

(b) 0 ns: Top view

(c) 20 ns: Side view



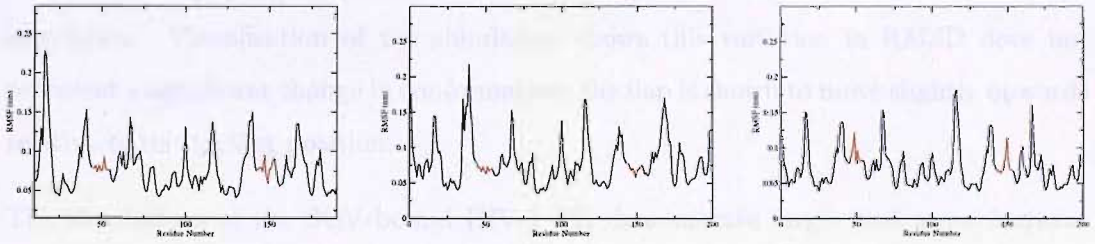
(d) 20 ns: Top view

FIGURE 8.3: Conformations of WT SQV-bound HIV-1 PR at start and end of MD1 simulation. The residues of the N-terminal are in red (residues 100 to 108).

The root mean squared fluctuations (RMSF) of each of the residues of the HIV-1 PR structures have been calculated over the duration of each of the MD simulations. For the RDFMD simulations, the plots display the average RMSF of the simulations for each of the systems, and the standard error has been calculated from the individual simulations.

The RMSF of the IDV-bound WT HIV-1 PR MD simulations are shown in figure 8.4. Since the RMSF of the MD simulations of the IDV- and SQV-bound HIV-1 PR systems are comparable, with no notable differences between them, only this plot is presented here (RMSF plots of the mutant IDV-bound and WT and mutant-bound HIV-1 PR MD simulations can be found in Appendix 8.4 and Appendix C.1).

The RMSF of MD1 simulations have been calculated for the first 10 ns and the last 10 ns separately. In the analysis of the apo HIV-1 PR systems, the RMSF of the flap residues were shown to be consistently larger in the 20 ns MD simulations compared with the 10 ns MD simulations (figure 7.2 of chapter 7). Analysis of the first and last 10 ns revealed that the larger RMSF values were obtained in the last 10 ns, showing that longer simulations are required to capture the full extent of the conformational dynamics of the flaps of the HIV-1 PR systems. However, this is not the case in these simulations, as no consistent increase in the RMSF value is observed in the last 10 ns compared with the first 10 ns of the 20 ns MD simulation.



(a) MD1 (0-10 ns): WT (b) MD1 (10-20 ns):WT (c) MD2 (0-10 ns):WT

FIGURE 8.4: RMSF of α-carbon atoms of all residues of IDV-bound HIV-1 PR calculated for the two MD simulations of the WT and mutant IDV-bound HIV-1 PR systems. Fitting carried out using the equilibrated WT starting structure (starting structure 1) The flap residues are marked in red.

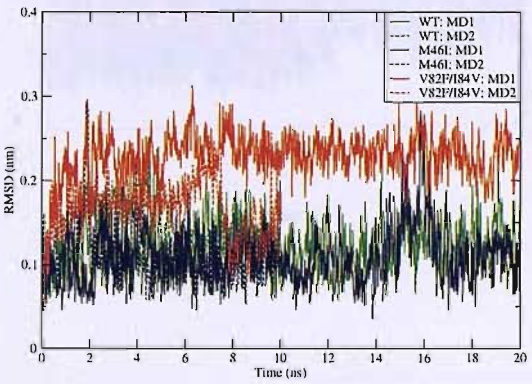
Overall, the RMSF generated by the MD simulations shows a significant reduction in flap flexibility for all the IDV and SQV-bound HIV-1 PR systems compared with the MD simulations of the HIV-1 PR apo-systems. This is also demonstrated by the RMSD plots

in figures 8.5 and 8.6. In the generation of these plots, the RMSD has been calculated over the duration of each of the MD simulations. Figures 8.5 (a) and (b) and 8.6 (a) and (b) display the RMSD of the α-carbons of each of the flaps (residues 44-54 and 144-154 for flap 1 and flap 2 respectively), using the equilibrated starting structures as the reference structures in each case, with fitting carried out on the non-flap α-carbon atoms (residues 1 to 43, 55 to 143 and 155- 198). Figures 8.5 (c) and 8.6 (c) display the RMSD over time for the α-carbon atoms of the non-flap residues of the protein.

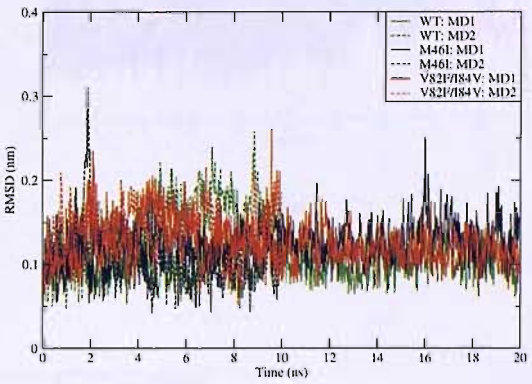
The RMSD values measured over the duration of the MD simulations for the non-flap residues do not fluctuate to any significant extent away from the equilibrated reference structure. This signifies the non-flap residues to be stabile and relatively immobile. This same observation was seen in the simulations of the apo WT and mutant HIV-1 PR systems (chapter 7, figures 7.4).

Generally, the RMSD values for the flap residues are shown to deviate away from the reference structure to a greater extent compared with the non-flap residues 8.5(a), (b) and 8.6(a), (b)). In the simulations of the IDV-bound HIV-1 PR systems, the average values appear to be largely similar to those observed for the non-flap residues, with the exception of flap 1 of the V82F/I84V mutant HIV-1 PR. The RMSD of this flap is shown to deviate from the RMSD of the equilibrated structure early on in the simulation, to a value of approximately 0.23 nm, where it then remains for the remainder of the simulation. Visualisation of the simulation shows this variation in RMSD does not represent a significant change in conformation; the flap is shown to move slightly upwards relative to its starting position.

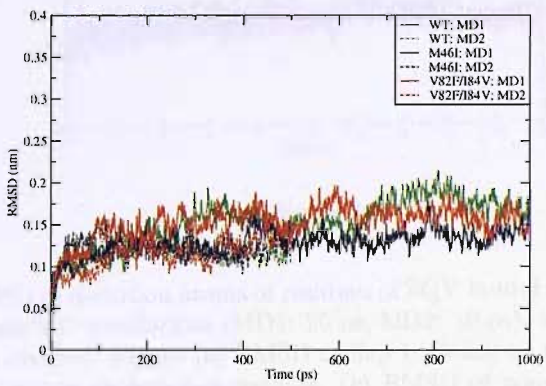
The simulations of the SQV-bound HIV-1 PR demonstrate larger and more frequent fluctuations of the flap residues compared with the IDV-bound simulations. The fluctuations in the RMSD correspond to small, but frequent movement of the flaps. These movements do not lead to any significant conformational change, and the flaps remain close to their starting conformations, but this difference may be an indication of increased flap movement of SQV-bound HIV-1 PR structures compared to those bound to IDV. As with the IDV-bound simulations, no evidence exists which confers differences in mobility between the WT and the mutant simulations of this system.



(a) Flap 1

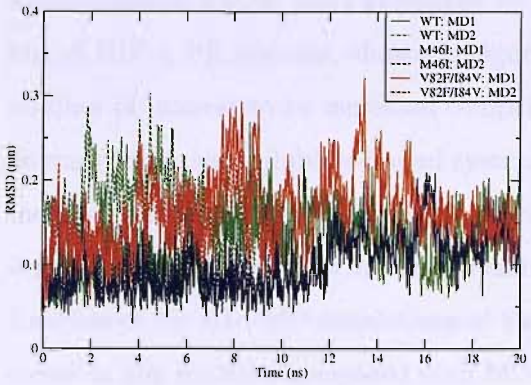


(b) Flap 2

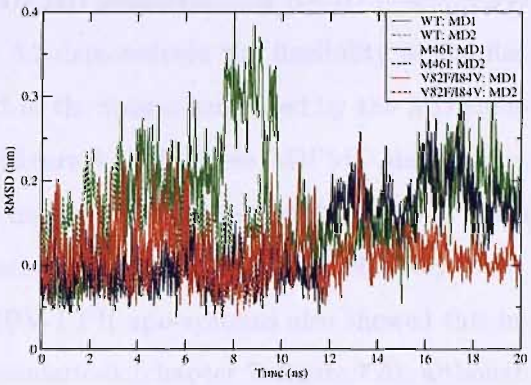


(c) Non-flap residues

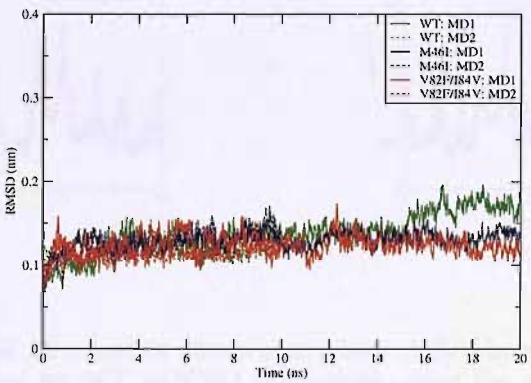
FIGURE 8.5: RMSD of α-carbon atoms of residues of IDV-bound HIV-1 PR calculated for WT and mutant MD simulations (MD1: 20 ns, MD2: 10 ns), using the equilibrated structure as the reference. Figure (a) RMSD of flap 1, fitting on non-flap residues, (b) RMSD of flap 2, fitting on non-flap residues, (c) RMSD of non-flap residues, fitting on non-flap residues. All calculations used the equilibrated starting structure as the reference structure.



(a) Flap 1



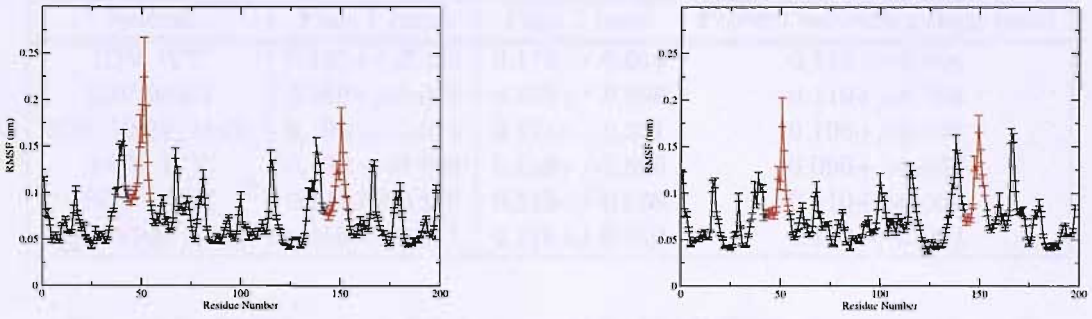
(b) Flap 2



(c) Non-flap residues

FIGURE 8.6: RMSD of α-carbon atoms of residues of SQV-bound HIV-1 PR calculated for WT and mutant MD simulations (MD1: 20 ns, MD2: 10 ns), using the equilibrated structure as the reference. Figure (a) RMSD of flap 1, fitting on non-flap residues, (b) RMSD of flap 2, fitting on non-flap residues, (c) RMSD of non-flap residues, fitting on non-flap residues. All calculations used the equilibrated starting structure as the reference structure.

Representative RMSF plots generated by RDFMD simulations of the IDV- and SQV-bound HIV-1 PR systems, shown in figures 8.7 demonstrate the flexibility of the flap residues of interest to be increased compared to the values generated by the MD simulations of the same inhibitor-bound systems (figure 8.4). In these RDFMD simulations, the flap residues are clearly shown to be the most flexible regions of the protein, which is not consistently the case in the MD simulations, where the RMSF is relatively lower. Analysis of the RDFMD simulations of the HIV-1 PR apo-systems also showed this increase in flap mobility compared with MD simulations (chapter 7, figure 7.5), although these regions are the target of the digital filter, so an increase in flexibility is expected.



(a) IDV

(b) SQV

FIGURE 8.7: RMSF of α-carbon atoms of residues of IDV-bound and SQV-bound HIV-1 PR calculated for WT RDFMD simulations, using the equilibrated starting structure (starting structure 1) as the reference structure for each system. RMSF values calculated as an average for all simulations, with the standard error calculated from each individual simulation marked. The flap residues are marked in red.

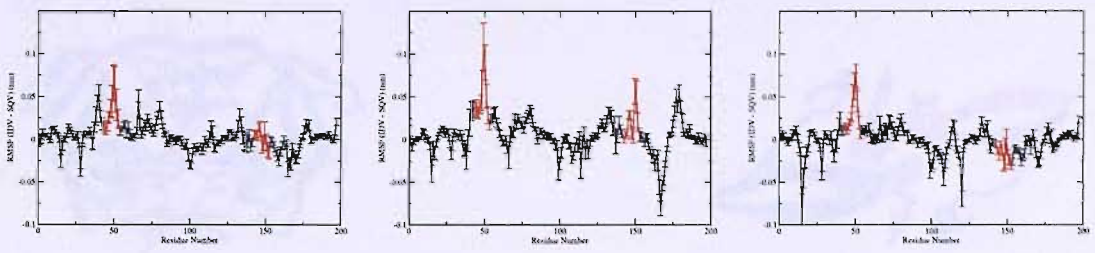
Comparison of the RMSF values obtained from the simulations of the WT and mutant IDV-bound HIV-1 PR systems do not show any consistent differences in the flexibility of the flaps, and only one representative RMSF plot of each inhibitor-bound systems is represented here. A full set of RMSF plots can be found in the appendix (figures C.2 and C.3). Some of the simulations observe slightly reduced flexibility in one flap compared with the other systems, which is not consistently observed in the other flap. Therefore, no conclusions can be made on variations in flexibility between the WT and mutant structures. This is also evident in the RMSD results (table 8.3).

The RMSD values shown in table 8.3 are averages of the RDFMD simulations for each system, with the standard error calculated, based on the individual simulations. As with the MD simulations, the RMSD values for the non-flap residues are low, indicating the majority of the HIV-1 PR structure to be relatively immobile and stable, with no notable differences observed between the WT and mutants of the IDV and SQV-bound structures and no consistent differences are seen in the RMSD values for both flap 1 and 2 between the WT and mutant simulations.

RMSD Calculation for RDFMD Simulations			
System	Flap 1 (nm)	Flap 2 (nm)	Protein excluding flaps (nm)
IDV WT	0.127+/-0.020	0.179+/-0.015	0.116+/-0.004
IDV M46I	0.210+/-0.019	0.169+/-0.009	0.119+/-0.004
IDV V82F/I84V	0.196+/-0.014	0.171+/-0.011	0.108+/-0.003
SQV WT	0.138+/-0.009	0.138+/-0.006	0.096+/-0.002
SQV M46I	0.141+/-0.110	0.145+/-0.008	0.110+/-0.003
SQV V82F/I84V	0.159+/-0.011	0.176+/-0.012	0.102+/-0.003

TABLE 8.3: RMSD calculated as an average for the RDFMD simulations for each inhibitor-bound system. The calculation of the RMSD of the flaps and the protein excluding the flaps was based on the α-carbon atoms, fitting on residues 1-43, 55-143 and 155-198 (non-flap residues,) using the two equilibrated starting structures as the reference structures. The standard error of the mean is calculated from the individual simulations.

Figure 8.8 displays the differences in RMSF between the IDV- and SQV-bound WT and mutant HIV-1 PR simulations. The differences in RMSF are consistently positive for flap 1 of the the IDV-bound simulations for the WT and mutant structures and flap 2 of the M46I SQV-bound HIV-1 PR. The differences in RMSF for flap 2 between the WT and V82F/I84V bound simulations are much smaller, demonstrating the flexibility of this flap to be more similar and indicates flap 2 of the SQV-bound flap to be slightly more flexible. This indicates the flaps of these HIV-1 PR systems to be more flexible than those of the SQV-bound WT and mutant structures. This overall increased mobility of the IDV-bound simulations is also demonstrated in the RMSD values, as shown in table 8.3, where the RMSD values of the flap residues (and non-flap residues) of the IDV-bound HIV-1 enzymes are generally higher.



(a) RDFMD: WT (b) RDFMD: M46I (c) RDFMD: V82F/I84V

FIGURE 8.8: The difference between the RMSF calculated for the WT and mutant RDFMD simulations of IDV- and SQV-bound HIV-1 PR systems. The average RMSF value has been calculated over the concatenated RDFMD simulations for each of the systems. The combined standard errors are marked. The flap residues are marked in red.

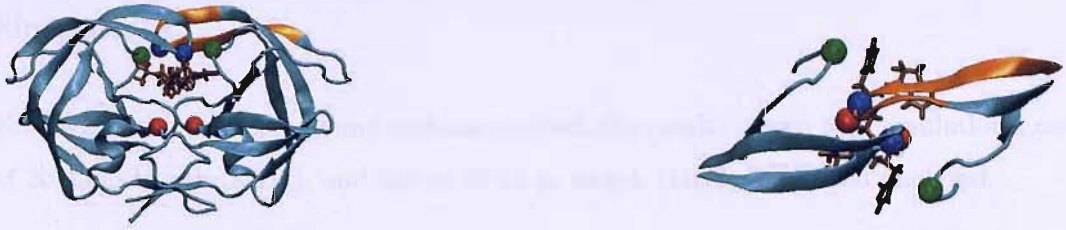
Comparison of the mobility of the MD and RDFMD simulations of the systems studied in this chapter has been judged on the RMSF values. As with the RDFMD simulations of the apo HIV-1 PR, comparison between the RMSD values obtained for the MD and RDFMD has not been made due to the number of RDFMD simulations carried out, and the complexity of the flap motion over the duration of the RDFMD simulations.

As observed in the comparison of the MD and RDFMD simulations of the apo HIV-1 PR systems, the increased RMSF of the RDFMD simulations compared with the MD simulations of the inhibitor bound structures clearly demonstrates the RDFMD methodology to increase the mobility of the flap residues, thus showing the method to consistently enhance the conformational dynamics of the flaps.

Summary

The analysis carried out in this section demonstrates the IDV- and SQV-bound HIV-1 PR systems to be stable over the duration of the MD and the RDFMD simulations.

The simulations of all the inhibitor-bound HIV-1 PR systems produce RMSD and RMSF values which indicate lower mobility of the flap residues compared with results of the equivalent apo-systems. This is due to the increased interaction between the flaps and between the flaps and bound-inhibitor of these structures. Visualisation of the trajectories confirms this, showing the flaps to remain in their closed conformation, close to their position at the start (figure 8.9).



(a) Side view (b) Top view

FIGURE 8.9: Conformations of IDV-bound HIV-1 PR showing the flaps in their closed conformation when bound to IDV. Flap 2 is in orange. Red corresponds to α-carbon atoms of the catalytic aspartate residues (residue 25 and 125 in monomers A and B respectively). Blue corresponds to α- carbon residues of the flap-tips (residues 50 and 150 in monomer A and B respectively) and green corresponds to α-carbon atoms of residues of the P1 loop (residues 81 and 181 in monomers A and B respectively)

As observed with the apo HIV-1 PR systems, the use of the RDFMD methodology increases the conformational dynamics of the flap residues compared with the use of conventional MD.

Owing to the significantly reduced mobility of these inhibitor-bound structures, it is harder to obtain information from the simulations which provides evidence of any variation of the flexibility of the flaps between the WT and mutant inhibitor-bound structures, and therefore, no conclusions have been made on the effect of mutation on the flap dynamics. However, the RMSD and RMSF data does show the flaps of the IDV-bound HIV-1 PR structures to be more flexible than those of the SQV-bound HIV-1 PR structures.

As with the apo-systems, the flaps are the focus of the analysis in the remainder of this chapter, since the chosen mutations are postulated to have an effect on their dynamics.

8.4.2 Principal Component Analysis

PCA has been used to study the dynamics of the flaps of the IDV- and SQV-bound HIV-1 PR enzymes. The dynamics of the rest of the protein has been disregarded for this analysis, to avoid the flap motion of interest being obscured by the motions of the rest of the protein. For consistency with the analysis of the apo HIV-1 PR systems, flap 2 is the main focus of analysis carried out here.

Analysis of Eigenvectors generated from inhibitor-bound HIV-1 PR MD Simulations.

For each of the inhibitor-bound systems studied, the results of two MD simulations, one of 20 ns in length (MD1), and one of 10 ns in length (MD2) have been analysed.

PCA has been carried out on each of the two trajectories generated by the MD simulations for each of the inhibitor-bound systems. In each case, the covariance matrix and resulting eigenvectors and eigenvalues have been generated using the α-carbon atom trajectories of flap 2 (residues 144-154), whilst fitted onto the α-carbon atoms of the non-flap residues (residues 1-43, 55-143, 155-198). This is the same method as used for the simulations of the apo HIV-1 PR systems (section 7.3.2 of chapter 7).

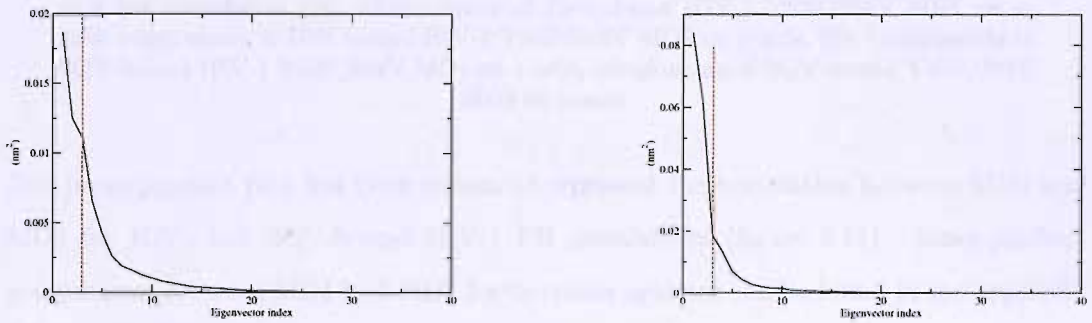
Figure 8.10 displays representative scree plots as a result of the PCA carried out on the MD1 simulation of WT IDV- and SQV-bound HIV-1 PR systems. The scree plots of the WT and mutant forms of IDV-bound HIV-1 PR are similar, and the same is the case for the SQV-bound systems, and so have not been included in this chapter.

The scree plots, generated as a result of the PCA on the MD simulations of the IDV-bound WT and mutant HIV-1 PR simulations are shown to capture 64-67 % of the total motion in the first three eigenvectors, whereas 68-81 % of the total motion is captured in the first three eigenvectors of the SQV-bound HIV-1 PR simulations. The relatively larger proportion of the total motion captured in the first three eigenvectors of the SQV-bound HIV-1 PR simulations compared with the IDV-bound simulations may indicate this structure to be able to undergo larger amplitude motions compared with the IDV-bound structure, as also indicated by the RMSD analysis. It may also be an indication that the motions associated with the first three eigenvectors perhaps being intrinsically more correlated.

The relatively lower RMSF and RMSD values of the inhibitor-bound simulations compared with the apo HIV-1 PR simulations indicates the flaps of these systems to have reduced mobility, which would result in the principal components capturing smaller and less defined motions. Therefore, the relatively low proportion of the total motion captured in the first three eigenvectors of the simulations of the IDV-bound simulations can be explained by the presence of the large number of small fluctuations.

Correlation Coefficients along the Diagonal	
Simulations compared	Correlation Coefficient (Eigenvector)
IDV WT MD1 and MD2	0.308 (1), 0.036 (2), 0.000(3)
IDV M46I MD1 and MD2	0.000 (1), 0.066 (2), 0.048 (3)
IDV V82F/I84V MD1 and MD	0.000 (1), 0.105 (2), 0.077 (3)
SQV WT MD1 and MD2	0.304 (1), 0.392 (2), 0.438 (3)
SQV M46I MD1 and MD2	0.262 (1), 0.229 (2), 0.267 (3)
SQV V82F/I84V MD1 and MD2	0.773 (1), 0.638 (2), 0.146 (3)

TABLE 8.4: Correlation coefficients along the diagonal for the first three eigenvectors of SQV- and IDV-bound HIV-1 PR WT and mutant MD1 and MD2 simulations. Matrix representation demonstrating levels of correlation between the first 10 eigenvectors are shown in figures 8.11 and C.4

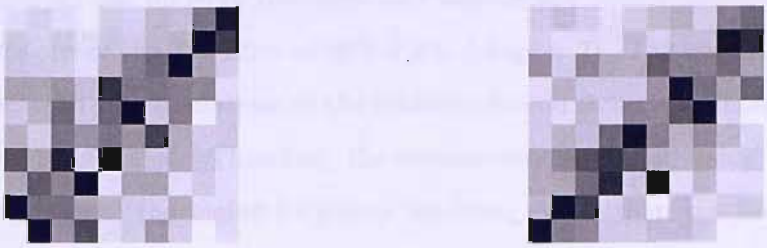


(a) IDV

(b) SQV

FIGURE 8.10: Scree plot depicting the proportion of motion captured in each of the eigenvectors of flap 2 of MD1 of WT IDV- and SQV-bound HIV-1 PR systems. The red line depicts the eigenvalues of the first three eigenvectors.

Figure 8.11 compares the inner-products of the eigenvectors generated by MD1 and MD2 for each of the inhibitor-bound HIV-1 PR systems. As described in the background section on PCA (in chapter 2, and in the analysis of the apo HIV-1 PR simulations, chapter 7), inner-product analysis is used to compare the level of correlation between eigenvectors, with perfect correlation depicted by a dark block, and no correlation depicted by a white block. Correlation coefficients (see table 8.4) have also been used to give a quantitative description of the correlation between two eigenvectors, with a value of 1 demonstrating the two eigenvectors to be identical, and a value of 0 meaning the eigenvectors are orthogonal.



(a) IDV (b) SQV

FIGURE 8.11: Comparison of inner products of the first ten eigenvectors of IDV and SQV-bound MD1 and MD2 simulations of flap 2. Eigenvectors are plotted sequentially along the x- and y-axis, with 1 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Shading corresponds to the correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation ([a]: Components of IDV-bound HIV-1 V82F/I84V MD1 on x-axis, components of IDV-bound HIV-1 V82F/I84V MD2 on y-axis, [b]: Components of SQV-bound HIV-1 V82F/I84V MD1 on x-axis, components of SQV-bound V82F/I84V MD2 on y-axis.

One inner-product plot has been chosen to represent the correlation between MD1 and MD2 for IDV- and SQV-bound HIV-1 PR simulations (figure 8.11). Inner-product comparisons between MD1 and MD2 for the other systems can be found in the appendix (chapter C, figure C.4).

Overall, the MD simulations of the IDV-bound systems show little correlation between each other, and where correlation is evident, it is off-diagonal, demonstrated by the poor correlation coefficients displayed in the table 8.4. In the case of the SQV-bound simulations, slightly improved diagonal correlation is evident, especially in the comparison of the MD simulations of the V82F/I84V mutant. The improved correlation between the mutant and WT simulations of the SQV-bound HIV-1 PR system compared with the IDV-bound HIV-1 PR system, along with the results of the scree plots may indicate the SQV-bound HIV-1 PR simulations to observe larger and relatively more defined motions compared with the IDV-bound HIV-1 PR simulations. The improved correlation may also be an indication of the motions of the SQV-bound simulations to be intrinsically more correlated.

This general overall lack of diagonal correlation between MD1 and MD2 can be attributed to the motions observed in these simulations being small. As the analysis of the flexibility of the systems indicates, in addition to the visualisation of the trajectories,

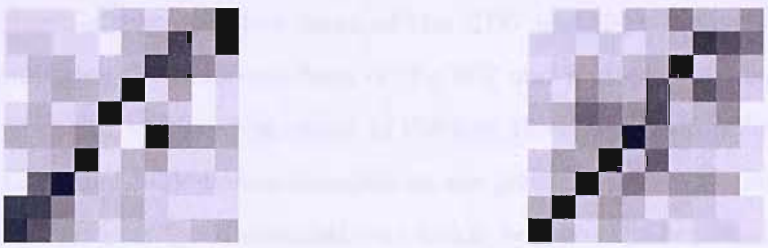
these structures do not undergo the large and well-defined conformational changes as seen in simulations of the apo-form of HIV-1 PR (chapter 7). The motions are significantly smaller, with the simulations of the inhibitor-bound structures showing the flaps to be far more constrained. Therefore, the eigenvectors generated using these trajectories capture more of the higher-frequency motions, rather than the slower motions associated with conformational change.

As a consequence of the poor correlation between the eigenvectors of the simulations, no further analysis has been carried out using the eigenvectors of these MD simulations. The RMSD and RMSF analysis of the RDFMD simulations indicates the application of the RDFMD methodology to promote mobility in the flap regions of these systems and therefore, PCA has been carried out using the RDFMD trajectories.

Analysis of Eigenvectors generated from inhibitor-bound HIV-1 PR RDFMD Simulations

PCA analysis has been carried out in the same manner for the RDFMD simulations as for the MD simulations.

Trajectories for analysis have been created from the concatenation of the individual RDFMD simulation trajectories, thus producing six trajectories, one for each inhibitor-bound system (IDV-bound: WT, M46I and V82F/I84V, SQV-bound: WT, M46I and V82F/I84V). The α-carbon atom coordinates of each of the trajectories for flap 2 constitutes the data-set for the calculation of the covariance matrix, created whilst fitting on the non-flap residues of the structure. The reference structure used for PCA must be the same if comparison between trajectories is required. Therefore, the reference structure used in each case was the α-carbon coordinates of the equilibrated WT IDV-bound structure.



(a) IDV (b) SQV

FIGURE 8.12: Comparison of inner products of the first ten eigenvectors of IDV- and SQV bound WT and mutant RDFMD simulations of flap 2. Eigenvectors are plotted sequentially along the x- and y-axis, with 1 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Shading corresponds to correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation (Components of WT on x-axis, components of V82F/I84V on y-axis)

Correlation Coefficients along the Diagonal	
Simulations compared	Correlation Coefficient (Eigenvector)
IDV WT and M46I	0.928 (1), 0.924 (2), 0.889(3)
IDV M46I and V82F/I84V	0.585 (1), 0.576 (2), 0.598 (3)
IDV WT and V82F/I84V	0.538 (1), 0.398 (2), 0.696 (3)
SQV WT and M46I	0.972 (1), 0.879 (2), 0.113 (3)
SQV M46I and V82F/I84V	0.981 (1), 0.924 (2), 0.067 (3)
SQV WT and V82F/I84V	0.933 (1), 0.858 (2), 0.540 (3)

TABLE 8.5: Correlation coefficients along the diagonal for the first three eigenvectors of SQV- and IDV-bound HIV-1 PR WT and mutant RDFMD simulations. Eigenvectors generated from the concatenation of the trajectories for each system. Matrix representation demonstrating levels of correlation between the first 10 eigenvectors are shown in figures 8.12

Figure 8.12 shows a representative plots of the IDV- and SQV-bound HIV-1 PR simulations, comparing the inner-products of the WT and mutant systems. Correlation coefficients comparing the inner-products of the first three eigenvectors for all the WT and mutant IDV- and SQV-bound simulations are given in table 8.5. For both IDV- and SQV-bound systems, good diagonal correlation between the principal motions are demonstrated, indicating that the WT and mutant simulations to be similar.

Since the eigenvectors of the simulations are shown to be similar, the trajectories of the WT and mutant IDV-bound HIV-1 PR systems have been concatenated together for further PCA analysis. This has also been carried out for the trajectories of SQV-bound HIV-1 PR WT and mutants.

The resulting inner-products of the two sets of eigenvectors have been compared in figure 8.13, and the correlation coefficients of the first three eigenvectors are given in table 8.6. They show that little correlation exists between the eigenvectors representing the major motions in the IDV and SQV-bound simulations. The highest similarity is shown to be between the third eigenvector of the IDV-bound RDFMD simulations and the second eigenvector of the SQV-bound RDFMD simulations, with a correlation coefficient of 0.508.

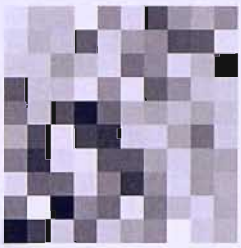
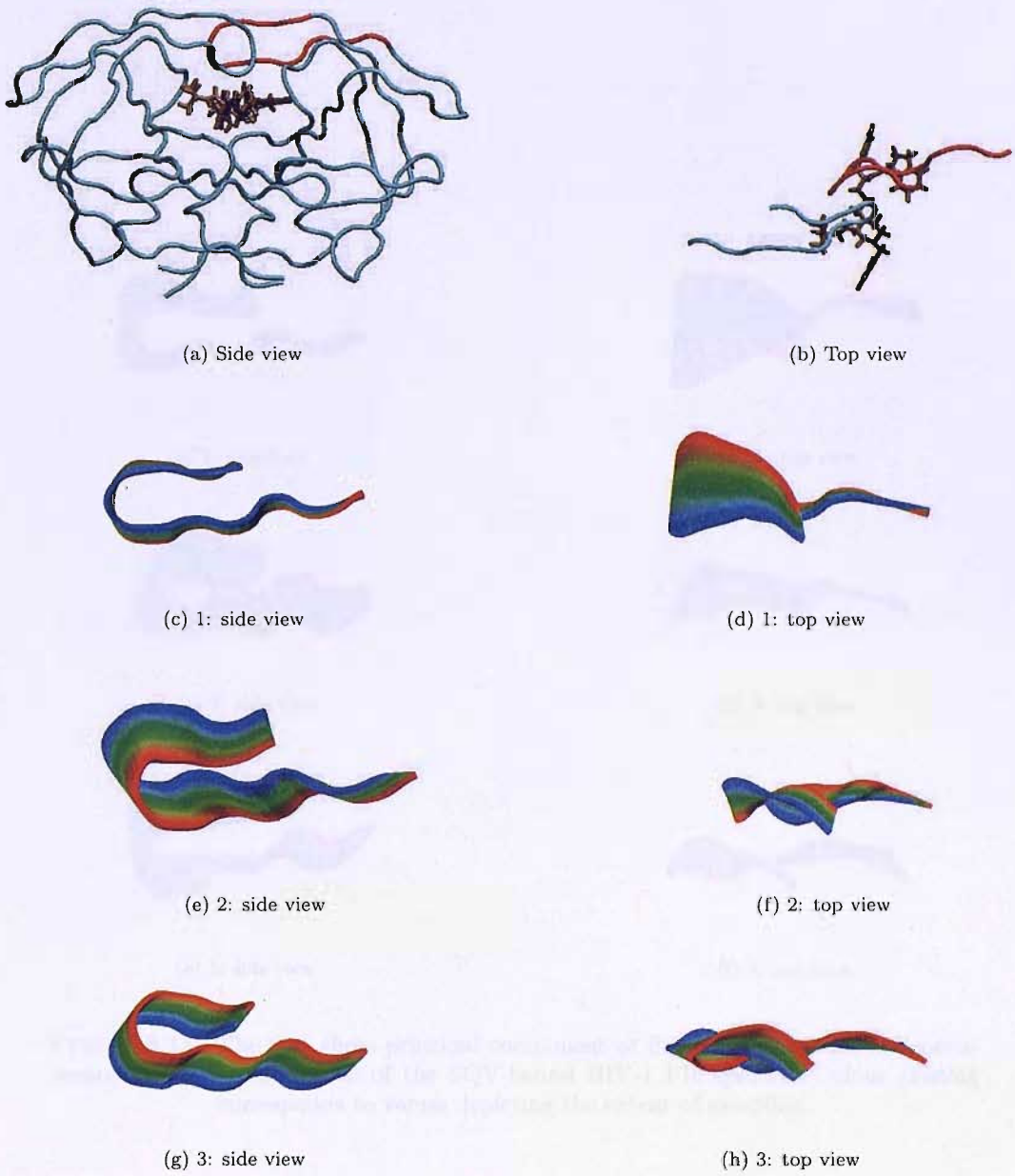


FIGURE 8.13: Comparison of inner products of the first ten eigenvectors of SQV structures and IDV-bound RDFMD simulations of flap 2. Shading corresponds to similarity between eigenvectors, with dark shading corresponding to high similarity and lighter shading to a low similarity. Eigenvectors are plotted sequentially along the x- and y-axis, with 1 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Components of IDV-bound structures on x-axis, components of SQV-bound structures on y-axis

Correlation Coefficients along the Diagonal	
Simulations compared	Correlation Coefficient (Eigenvector)
IDV and SQV	0.399 (1), 0.000 (2), 0.109(3)

TABLE 8.6: Correlation coefficients along the diagonal for the first three eigenvectors of SQV- and IDV-bound HIV-1 PR RDFMD simulations. Eigenvectors generated from the concatenation of the SQV WT and mutant HIV-1 PR and IDV WT and mutant trajectories for each system. Matrix representation demonstrating levels of correlation between the first 10 eigenvectors are shown in figures 8.13

Visualisation of the motion of flap 2 captured by the first three eigenvectors of the IDV- and SQV-bound HIV-1 PR simulations is displayed in figures 8.14 and 8.15. These figures have been generated through interpolation between the two extreme structures for each eigenvector. The motion depicted by the first principal component (figures 8.14 (c) and (d) and figures 8.15 (a) and (b)) shows an element of horizontal motion in the simulations of the IDV and SQV-bound HIV-1 PR systems. However, the low diagonal correlation seen between the inner-products of these eigenvectors is due to the subtle vertical motion also incorporated in the first eigenvector of the SQV-bound HIV-1 PR simulations. The second and third principal components are shown to display variations of small flap motions, which explains the low correlation between the inner-products of the IDV and SQV HIV-1 PR simulations.



(a) Side view

(b) Top view

(c) 1: side view

(d) 1: top view

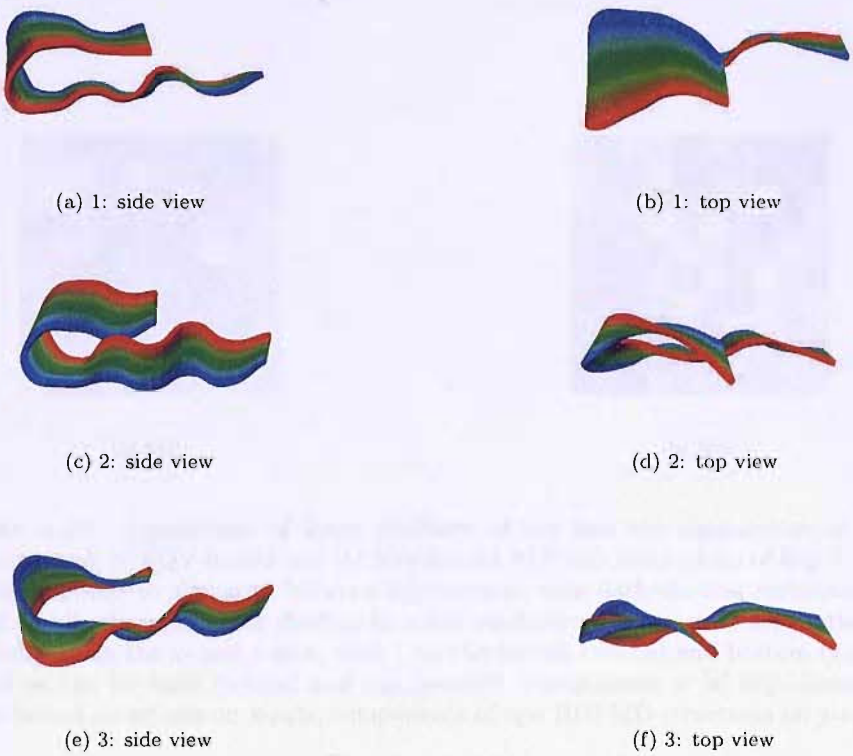
(e) 2: side view

(f) 2: top view

(g) 3: side view

(h) 3: top view

FIGURE 8.14: The first three principal components of flap 2 created from the concatenated RDFMD simulation of the IDV-bound HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling. (a) and (b) show the inhibitor-bound HIV-1 PR structure with the flap 2 residues, the target region for the PCA shown in red.



(a) 1: side view

(b) 1: top view

(c) 2: side view

(d) 2: top view

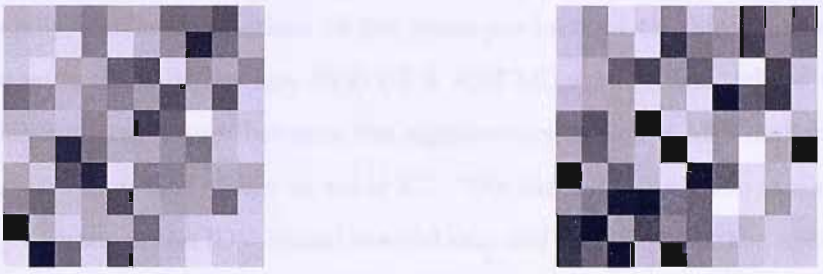
(e) 3: side view

(f) 3: top view

FIGURE 8.15: The first three principal component of flap 2 created from the concatenated RDFMD simulation of the SQV-bound HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling.

Owing to the restricted motion of the flap structures, PCA analysis based on the eigenvectors generated by the inhibitor-bound simulations has been shown to capture only small, poorly-defined and subtle flap 2 movements as the principal motions of the inhibitor-bound simulations.

To compare the extent of sampling of the important motions associated with the opening, closing and flap-tip curling, further PCA analysis has been carried out using the eigenvectors generated by the concatenation of the RDFMD simulations of the apo HIV-1 PR systems.



(a) SQV (b) IDV

FIGURE 8.16: Comparison of inner products of the first ten eigenvectors of apo-structures and [a] SQV-bound and [b] IDV-bound RDFMD simulations of flap 2. Shading corresponds to similarity between eigenvectors, with dark shading corresponding to high similarity and lighter shading to a low similarity. Eigenvectors are plotted sequentially along the x- and y-axis, with 1 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Components of [a] SQV-bound [b] IDV-bound structures on x-axis, components of apo RDFMD structures on y-axis.

Correlation Coefficients along the Diagonal	
Simulations compared	Correlation Coefficient (Eigenvector)
IDV and apo RDFMD	0.207 (1), 0.409 (2), 0.284(3)
SQV apo RDFMD	0.000 (1), 0.022 (2), 0.01 (3)

TABLE 8.7: Correlation coefficients along the diagonal for the first three eigenvectors of SQV- and IDV-bound HIV-1 PR WT and mutant RDFMD simulations and eigenvectors generated from apo RDFMD simulations. Eigenvectors generated from the concatenation of the trajectories for each system. Matrix representation demonstrating levels of correlation between the first 10 eigenvectors are shown in figure 8.16.

Figure 8.16 compares the inner-products of the eigenvectors generated from the concatenated SQV and IDV inhibitor-bound RDFMD simulations with those generated from the concatenated RDFMD trajectories of the apo HIV-1 PR. Overall, correlation is poor, especially for the comparison of the inner-products of the SQV-bound RDFMD trajectories with those of the apo HIV-1 PR RDFMD simulations, where there is almost no diagonal correlation between the eigenvectors, which is also demonstrated by the low correlation values shown in table 8.7. The highest correlation is seen between the first eigenvector of the IDV-bound simulations and the second of the apo HIV-1 PR simulations (correlation coefficient 0.743). This motion corresponds to the horizontal gating motion of flap 2, towards and away from the active site (figure 8.14 (a) and (b), and figure 7.18 in chapter 7). Additionally, the vertical motion, captured as the first eigenvector of the apo HIV-1 PR simulations, is shown to be present to some extent in the second eigenvector of the IDV-bound HIV-1 PR simulations (correlation coefficient 0.561).

Slightly improved correlation is seen between the inner-products of the eigenvectors generated by the concatenation of the WT and mutant IDV-bound HIV-1 PR trajectories with those of the concatenated WT and mutant apo HIV-1 PR trajectories. Again, the horizontal gating motion is observed to be captured as one of the principal motions, demonstrated by the good diagonal correlation with the inner-product of the second eigenvector of the apo HIV-1 PR simulations (correlation coefficient: 0.409). The vertical motion of the flap (eigenvector 1) is shown to be present to a small extent in the third eigenvector of the SQV-bound HIV-1 PR simulations (correlation coefficient: 0.349).

Overall, comparison of the inner-products between the apo HIV-1 PR simulations and the inhibitor-bound simulations show low correlation between the motions, although

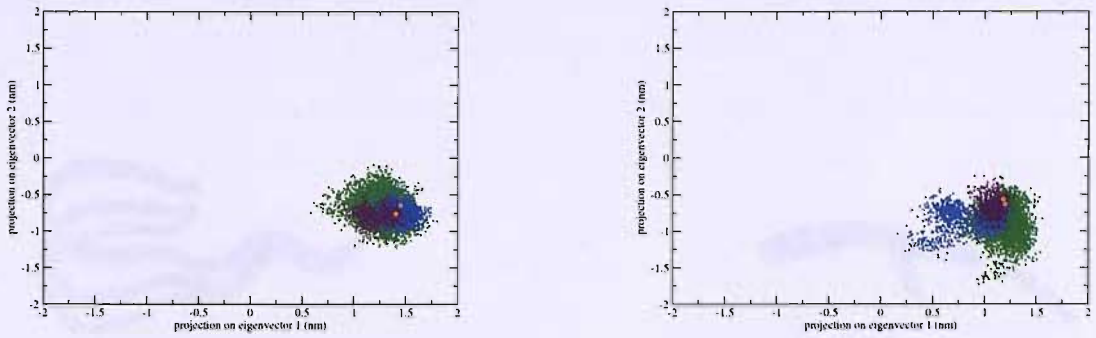
some similarity between the eigenvectors is evident, especially the horizontal gating motion. This low correlation is to be expected owing to the low mobility of the flap structures of the inhibitor-bound systems, resulting in the absence of the large defined motions observed in the apo HIV-1 PR simulations.

Projection onto Eigenvectors of Apo HIV-1 PR RDFMD Simulation Eigenvectors: Analysis of Eigenvectors 1 and 2.

To directly compare the conformational sampling of the flaps of the mutants, inhibitor-bound and the apo-structures, the trajectories of each must be projected onto the same eigenvectors. In this case those generated by the apo RDFMD simulations have been chosen to investigate the extent to which the important conformations associated with flap opening and closing have been sampled by the inhibitor-bound simulations.

Figure 8.17 represents the sampling of eigenvectors 1 and 2 with respect to each other. As described in the apo HIV-1 PR chapter, the motions captured in the first two eigenvectors describe the overall conformation of the flaps as being semi-open, closed or open. The first eigenvector captures the vertical motion of flap 2, towards and away from the active site. The projection on this eigenvector is shown on the x-axis of the two-dimensional plot. Positive projection values correspond to the flap moving towards the active site, and negative values correspond to the flap moving vertically away from the active site (figures 8.18 (a), (b), (e) and (f)). Sampling of the second eigenvector represents a horizontal gating motion, towards and away from the active site. The projection onto this eigenvector is shown along the y-axis, with negative values corresponding to flap 2 moving away from the active site, and positive values corresponding to the flap moving towards the active site (note: the positive and negative values for eigenvector 2 of the inhibitor-bound HIV structures have the opposite meaning to those of the apo HIV structures due to the flaps being swapped in position in the closed structure compared with the apo-structure).

As shown in figure 8.17, the sampling of the MD simulations covers a smaller area of the plot compared to that sampled by the RDFMD simulations, and the area of sampling

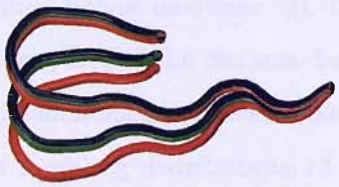


(a) WT IDV-bound HIV-1 PR

(b) WT SQV-bound HIV-1 PR

FIGURE 8.17: Sampling of the first two eigenvectors, generated from the projection of the inhibitor-bound WT RDFMD (green) ([a]: IDV-bound HIV-1 PR, [b]: SQV-bound HIV-1 PR) and the two MD (MD1: blue and MD2: maroon) simulations onto the eigenvectors of the second flap created from the concatenation of all apo RDFMD trajectories. The two starting structures are represented by an orange and brown circle.

covered by the RDFMD simulation encompasses the sampling by the MD simulations. This is the case for the majority of the RDFMD and MD sampling distributions (with one exception), thus the two-dimensional plots for all the systems are not shown here. The exception of this is shown in figure 8.17(b), where MD1 of the WT SQV-bound HIV-1 PR system is shown to sample conformations where the flap lifts vertically away from the active site to a greater extent than what is seen in the RDFMD simulations. However, owing to the highly constrained motion of the flaps, although the area of sampling is notably different from the RDFMD simulations, on visual inspection of conformations where this is observed, this actually corresponds to only a very small movement of the flap. The reason for the unexpected result of the MD1 simulation may be a consequence of the unusual dynamics of the N-terminal residues, as previously discussed in an earlier section 8.4.1. The unusual behaviour noted may have propagated to the flaps, affecting their motion. This is demonstrated by distance measurements.



(a) IDV: Eigenvector 1: side

(b) IDV: Eigenvector 1: top

(c) IDV: Eigenvector 2: side

(d) IDV: Eigenvector 1: top

(e) SQV: Eigenvector 1: side

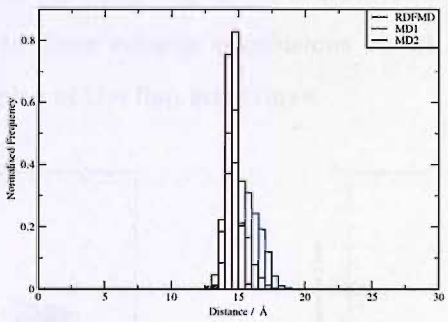
(f) SQV: Eigenvector 1: top

(g) SQV: Eigenvector 2: side

(h) SQV: Eigenvector 1: top

FIGURE 8.18: Extreme conformations of flap 2 of inhibitor-bound HIV-1 from the projection of RDFMD simulations onto apo HIV-1 PR eigenvectors 1 and 2.

Distances were measured between the α-carbon atom residues of the flap tips (residue 50 of monomer A, and residue 150 of monomer B) and the corresponding α-carbon atom of the catalytic aspartate of the same monomer (residue 25 of monomer A and residue 125 of monomer B). Figure 8.19 displays a histogram showing the distribution of sampling of the distance between the tip of flap 2 and the catalytic aspartate of the same monomer for the simulations of WT SQV-bound HIV-1 PR. Histograms displaying the sampling distributions of the described distances of all the IDV- and SQV-bound HIV-1 PR systems studied do not observe any significant differences between the WT and mutants, and between the distances measured for flap 1 and flap 2, and so are not shown here. As also demonstrated by the sampling of the first eigenvector, the RDFMD simulations sample a slightly larger range of flap-tip to active site distances, similar to that shown between the distances ranges sampled by the RDFMD and MD2 simulation in 8.19. However, there one exception to this, the MD1 simulation of WT SQV-bound HIV-1 PR, where the flap is shown to sample larger distances between the flap tip and the active site. This result is consistent with the sampling of the first eigenvector and as noted previously, this be may a consequence of the unusual dynamics of the N-terminal residues.



(a) Residues 150-125: SQV

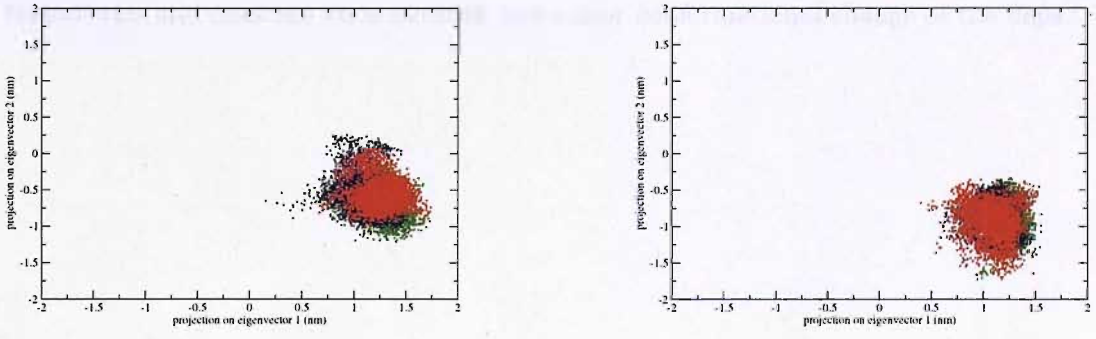
FIGURE 8.19: Sampling distributions of the distances sampled by the WT SQV-bound HIV-1 PR simulations of the α-carbon atom of flap-tip 2 residue (residue 150) to α-carbon atom of the catalytic aspartate residue of the same monomer (residue 125). For each of the MD simulations and the concatenated WT RDFMD simulations, the frequency of distances has been normalised to a sum of 1.

Comparison of the sampling of the WT and mutants of the IDV- and SQV-bound HIV-1 PR RDFMD simulations is shown in figure 8.20. The area of sampling is dense, demonstrating that only a narrow range of conformations are observed, which correspond to the

closed conformation. Differences between the sampling of the WT and mutants is shown to be negligible for the IDV- and SQV-bound simulations. The extreme conformations generated from the projection of the concatenated WT and mutant trajectories onto the eigenvectors are shown in figure 8.18, which also demonstrate the restricted conformational dynamics of flap 2. For the IDV-bound simulations 8.18 (a)-(d), the projection onto the first eigenvector shows the V82F/I84V mutant's flap to reach further into the active site compared with the other WT and M46I mutant HIV-1 PR simulations. On the two-dimensional plot, this flap conformation is represented by the sampling of the most positive values of eigenvector 1 (x-axis), but as the plot shows, the sampling of this region is very sparse, and as a consequence, the existence of this difference is arguably insignificant.

The same is true for the distribution of the sampling of the first two eigenvectors of the SQV-bound HIV-1 PR simulations. Where there are shown to be variations between the extreme conformations sampled by the WT and mutants 8.18 (e)-(h), the variation is slight and the two-dimensional plot shows the sampling to be sparse in this region.

Overall, the motions of the inhibitor-bound simulations are shown to be restricted, and although small variations in the sampling of the WT and mutants exists, the differences are too slight to be able to draw reliable conclusions about any effect of mutation on the conformational dynamics of the flap structures.



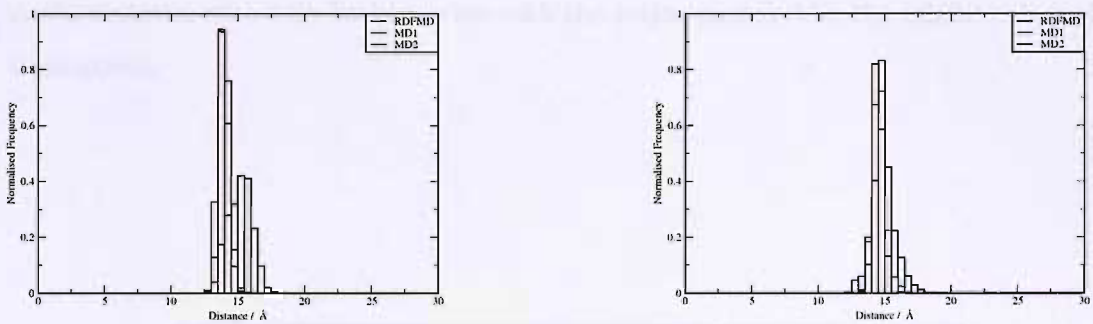
(a) IDV

(b) SQV

FIGURE 8.20: Sampling of the first two eigenvectors, generated from the projection of the inhibitor-bound ([a]: IDV-bound HIV-1 PR, [b]: SQV-bound HIV-1 PR) α-carbon RDFMD trajectories (green: WT, black: M46I, red: V82F/I84V HIV-1 PR) onto the eigenvectors of the second flap created from the concatenation of all apo RDFMD trajectories.

Comparing the overall areas of sampling of the first two eigenvectors in the two- dimensional plots between the SQV- and IDV-bound simulations (figure 8.20) shows a distinct difference in the overall areas sampled. The size of the area of sampling is approximately the same for the simulations of IDV- and SQV-bound HIV-1 PR simulations, however, the area of sampling of simulations of the IDV-bound HIV-1 PR corresponds to conformations where flap 2 is relatively vertically higher than observed in the simulations of the SQV-bound simulations, and also samples conformations where flap 2 is closer to the active site in a direction perpendicular to that of the vertical motion. These differences correspond to only small variations in flap movement and is likely to be due to the positioning of the inhibitor within the active site affecting the location of the flaps.

This difference in also evident in distance measurements. Figure 8.21 displays representative histograms showing the distribution of sampling of the distance between the tip of flap 2 and the catalytic aspartate of the same monomer for the simulations of M46I IDV- and SQV-bound HIV-1 PR (figures 8.21 (a) and (b)) As previously mentioned, no significant differences are observed between the WT and mutants, and between the distances measured for flap 1 and flap 2. However, the MD and RDFMD simulations of IDV-bound HIV-1 PR WT and mutant are shown to consistently sample relatively more values which correspond to the flaps being slightly vertically higher than those of the SQV-bound HIV-1 PR simulations. It must be noted that although a difference is apparent, it actually corresponds to only subtle variations on visual inspection of the trajectories and does not coincide with any major conformational change of the flaps.



(a) Residues 150-125: IDV

(b) Residues 150-125: SQV

FIGURE 8.21: Comparison of the sampling distributions of the distances sampled by the SQV- and IDV-bound M46I HIV-1 PR simulations of the α-carbon atom of flap-tip 2 residue (residue 150) to α-carbon atom of the catalytic aspartate residue of the same monomer (residue 125). For each of the MD simulations and the concatenated RDFMD simulations, the frequency of distances has been normalised to a sum of 1.

Comparison of Sampling of Eigenvectors 1 and 2 of the Inhibitor-bound HIV-1 PR simulations with apo HIV-1 PR simulations

Comparison of the sampling of the first two eigenvectors of the inhibitor-bound simulations, with the sampling by the apo HIV-1 PR simulations is demonstrated in the two-dimensional plot in figure 8.22. The sampling of the inhibitor-bound simulations lies in a confined region, with the projection of their simulations onto the first two eigenvectors resulting in sampling solely in the region corresponding to the closed conformations. These conformations are depicted by the positive values along the x-axis (eigenvector 1) and negative values along the y-axis (eigenvector 2). The sampling of the apo HIV-1 PR RDFMD simulations is far more diffuse, sampling a significantly larger variety of conformations, with very little overlap with the region sampled by the inhibitor-bound simulations.

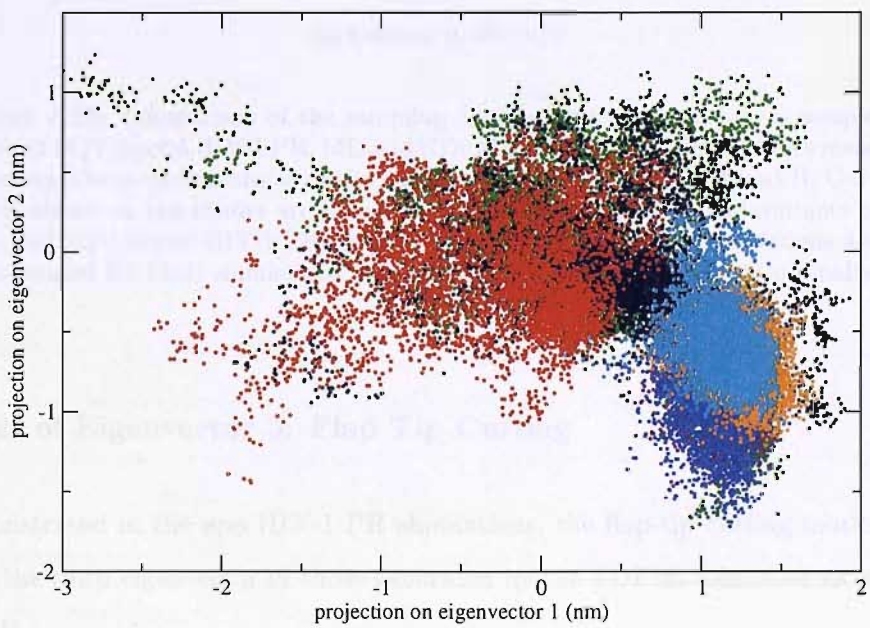
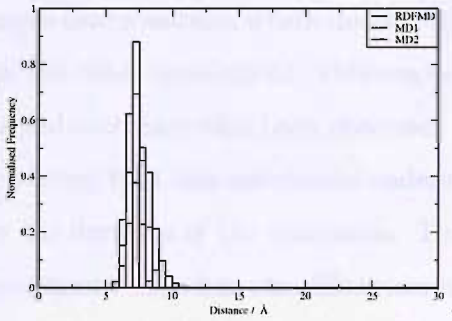


FIGURE 8.22: Sampling of the first two eigenvectors, generated from the projection of the apo and inhibitor-bound HIV-1 PR α-carbon RDFMD trajectories onto the eigenvectors of the second flap created from the concatenation of all apo RDFMD trajectories. Colour coding: APO (WT: light green, M46I: black, V82F/I84V: red) IDV-bound (WT: orange, M46I: cyan, V82F/I84V: brown) and SQV-bound (WT: dark green, M46I: violet, V82F/I84V: blue)

The relatively restricted motion of flap 2 of the inhibitor-bound simulations is also demonstrated by the distance measurements between the flap tip residues (α-carbon atoms of residue 50 and 150 of flap 1 and flap 2 respectively) and between the flap-tips and the catalytic aspartate (figure 8.23). The range and magnitude of the distances sampled by the inhibitor-bound simulations are far reduced compared with the apo HIV-1 PR simulations, whose flaps are shown to be able to separate over 25 Å vertically above the active site, and over 15 Å horizontally away in a gating motion figure (figures 7.24 7.28 and 7.31). Consequently, the ranges sampled by the WT and mutant structures are more similar, making the judgement of differences between the WT and mutants more difficult.



(a) Residues 50-150: SQV

FIGURE 8.23: Comparison of the sampling distributions of the distances sampled by the M46I SQV-bound HIV-1 PR. MD and RDFMD simulations. The distance measured is between the α-carbon atoms of the flap-tip residues of monomer A and B. Only one plot is shown as the results are shown to be similar for the WT and mutants of the IDV- and SQV-bound HIV-1 PR simulations. For each of the MD simulations and the concatenated RDFMD simulations, the frequency of distances has been normalised to a sum of 1.

Analysis of Eigenvector 3: Flap Tip Curling

As demonstrated in the apo HIV-1 PR simulations, the flap-tip curling motion is captured in the third eigenvector of those generated by the RDFMD simulations of the apo HIV-1 PR.

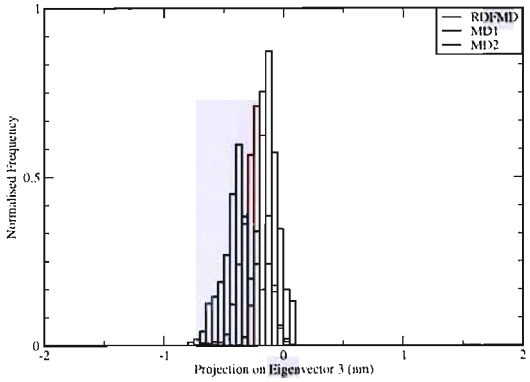
Figure 8.24 and 8.25 shows the projection of the SQV- and IDV-bound HIV-1 PR MD and RDFMD simulations onto the third eigenvector, displayed as a histogram. Sampling of positive values along the x-axis are indicative of the flap-tip residues being curled

towards the active site, and negative values correspond to conformations where the flap-tip residues are curled away from the active site (note: the positive and negative values for eigenvector 3 of the inhibitor-bound HIV structures have the opposite meaning to those of the apo HIV structures due to the flaps being swapped in position in the closed structure compared with the apo-structure).

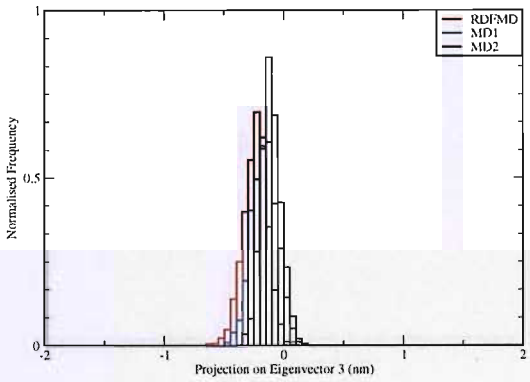
In all but one simulation, the sampling of this eigenvector is shown to be enhanced using the RDFMD methodology compared with conventional MD, with the sampling of the RDFMD simulations shown to incorporate the range of conformations sampled by the MD simulations, as also observed in the sampling of eigenvectors 1 and 2. The one exception to this is the WT SQV-bound MD1 simulation (figure 8.24 (a)), which is shown to preferentially sample conformations where the flap-tip is curled away the active site more than observed in the other simulations. Differences in sampling of the other eigenvectors for this MD simulation have also been observed. At the start of the result chapter (figure 8.3), it was noted that this simulation underwent abnormal changes in the N-terminal region over the duration of the simulation. This may have had an effect on the flaps also, which may have resulted in the differences in sampling seen.

Figures 8.24 (a)-(c) show the WT and mutant forms of the SQV-bound RDFMD HIV-1 PR systems to display similar sampling distributions of eigenvector 3. This is also observed for the WT and M46I IDV-bound RDFMD simulations. However, the extreme conformations (figure 8.26) and the sampling distribution of eigenvector 3 (figure 8.25) indicates that the V82F/I84V HIV-1 PR mutant system is able to sample conformations where the flap-tip is able to curl towards the active site to a greater extent than the WT and M46I SQV-bound HIV-1 PR enzymes, shown by the more positive values sampled. However, as the histogram of the sampling of the third eigenvector shows 8.25 (c), only a small percentage of the total conformations sampled is of this conformation. Furthermore, the difference in the conformation sampled between the WT and this mutant, demonstrated by visualisation of the eigenvector extremes (figure 8.26) is small, so the result possibly not significant.

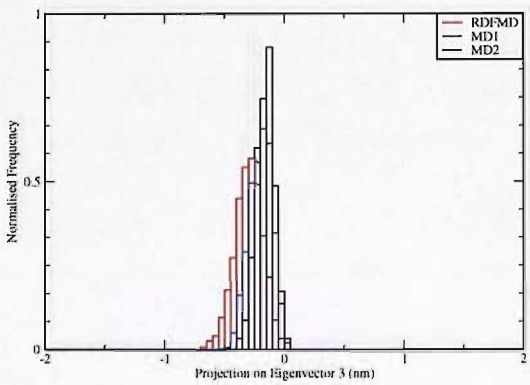
Comparison with the apo HIV-1 PR simulations shows an overall reduction in the flexibility of the flap-tips. The simulations of the apo HIV-1 PR systems are able to sample conformations where the flap-tip is curled-back away from the active site (figure 7.35),



(a) WT

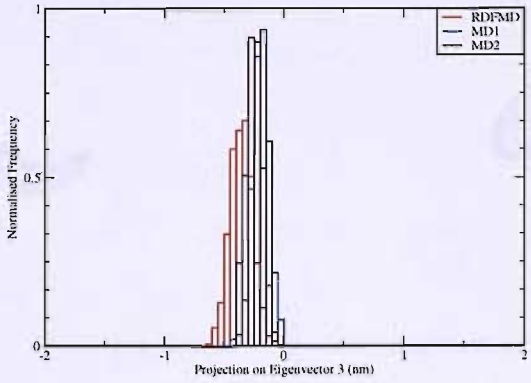


(b) M46I

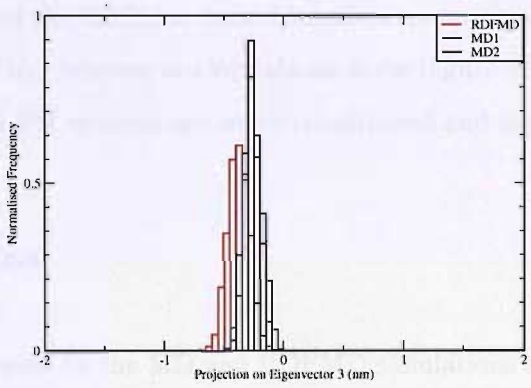


(c) V82F/I84V

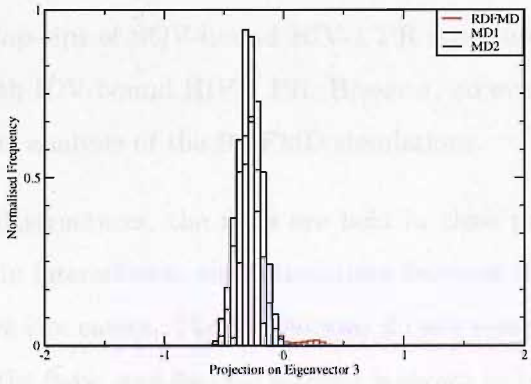
FIGURE 8.24: Sampling of the third principal component by RDFMD simulations of the WT, M46I and V82F/I84V SQV-bound HIV-1 PR systems. Eigenvectors generated from a covariance matrix from the concatenation of the all the apo HIV-1 PR simulations. For each of the MD simulations and the concatenated RDFMD simulations, the projection onto eigenvector 3 has been normalised to a sum of 1.



(a) WT



(b) M46I



(c) V82F/I84V

FIGURE 8.25: Sampling of the third principal component by RDFMD simulations of the WT, M46I and V82F/I84V IDV-bound HIV-1 PR systems. Eigenvectors generated from a covariance matrix from the concatenation of the all the apo-HIV-1 PR simulations. For each of the MD simulations and the concatenated RDFMD simulations, the projection onto eigenvector 3 has been normalised to a sum of 1.

(a) SQV (b) IDV

FIGURE 8.26: Extreme conformations sampled by the first three principal components of flap 2 of HIV-1 PR IDV-bound simulations. Figures generated by the projection of the inhibitor-bound trajectories onto the eigenvectors created by the RDFMD simulations of the apo HIV-1 PR systems (WT: green, M46I: black, V82F/I84V: red).

but as the projection of the inhibitor-bound simulations on eigenvector 3 (figure 8.24), and the visulisation of the extreme conformations show (figure 8.26), the flap-tips of the inhibitor-bound HIV-1 PR systems are more constrained and possess less mobility.

8.5 Conclusions

The trajectories generated by the MD and RDFMD simulations show flap 2 of the IDV- and SQV-bound HIV-1 PR structures to have remained in its closed conformation, with only minor fluctuations of the flap observed. Some of the RDFMD simulations observed *trans-cis* isomerism of the peptide bonds in the target flap-tip residues, whereas *trans-cis* isomerism was not observed in any of the IDV-bound HIV-1 PR simulations. This would perhaps suggest the flap-tips of SQV-bound HIV-1 PR structures to possess increased mobility compared with IDV-bound HIV-1 PR. However, no evidence to this effect has been observed from the analysis of the RDFMD simulations.

In the inhibitor-bound structures, the flaps are held in close proximity to each other through flap-tip-flap-tip interactions, and interactions between the flaps and the inhibitor bound in the active site cavity. The simulations do not sample the large vertical or horizontal motions of the flaps, and flap-tip activity is shown to be reduced compared to the motions observed in the apo HIV-1 PR simulations. In the apo HIV-1 PR RDFMD simulations, these interactions are not present, and differences in the dynamics between

the WT and mutants is apparent, which is postulated to have an effect on inhibitor-binding either by stabilising the closed conformation (M46I) or increasing flap activity (V82F/I84V).

In an attempt to increase the conformational sampling of the inhibitors, an increased range of flap residues were included in the target of the digital filter in the RDFMD simulations, including residues which may possess interactions with the inhibitor. However, these simulations failed to enhance the range of conformations sampled. Therefore, for the study of the IDV- and SQV-bound HIV-1 PR systems, the RDFMD simulations have failed to sample any major conformational change, and as a consequence, no conclusive differences between the WT and mutants of the inhibitor-bound structures have been observed.

PCA carried out on the trajectories of the RDFMD simulations suggest a difference between the conformations achieved by SQV- and IDV-bound simulations, where the HIV-1 PR IDV-bound simulations appear to show flap 2 to be able to move away from the protein in a horizontal gating motion, to a further extent than seen in the HIV-1 PR SQV-bound simulations. But owing to the small scale of the motions and the subtlety of difference between the results, it is unclear whether this is of any siguificance.

The timescale for major conformational change which occurs in the inhibitor-bound HIV-1 PR enzymes may be too large to be seen in the simulations reported here, and RDFMD may be an inappropriate technique to be able to enhance the dynamics of the systems studied in this chapter.

Further work into optimising the parameters of the RDFMD method may lead to improved results, or the use of other enhanced sampling methods, such as those mentioned in chapter 3 may be more successful in the study of these inhibitor-bound systems.

Chapter 9

Human Immunodeficiency Virus-1 Integrase (HIV-1 IN)

Aims

The aim of this chapter is to use conventional MD and the enhanced sampling technique of RDFMD to study the conformational dynamics of the catalytic domain of the apo HIV-1 IN enzyme in its WT and mutant (G140A/G149A) forms. As mentioned in the background section to this enzyme (chapter 6), the length structure comprises three domains, but owing to the lack of structural information, the domains are often studied individually. Studies of the activity of HIV-1 IN have been focused on the catalytic domain since residues 50-190 are sufficient to promote disintegration *in vitro* (reverse reaction to integration -see chapter 6) [317] [318]. The mechanism of action of this enzyme is not fully understood, but the mobility of the catalytic loop (residues 140-149) is postulated to play a fundamental role in the functioning of this enzyme [351].

The double G140A/G149A mutation is reported to deleteriously affect the integration activity of the enzyme, by affecting the conformational dynamics of the loop. Experimental and theoretical studies postulate this mutant to operate through a change in the conformational dynamics of the loop causing it to be less mobile compared with the WT HIV-1 IN. [351] [111].

9.1 Simulation Details

9.1.1 System Preparation and Equilibration

The starting structure 1BL3 [343], an apo-form of the wild type catalytic domain of HIV-1 IN, was taken from the Protein Data Bank [8]. This structure was chosen as it was the only crystal structure available at the time which possessed all the residues of the loop and included the catalytically important Mg^{2+} ion. However, three end residues, 210-212 are missing from this structure, but this was not considered significant, as they are located away from the active loop, the focus of this study. The WHATIF [171] program was used to add polar hydrogens and to check the structure. The AMBER utility XLEAP [172] was used to add other hydrogen atoms, and to solvate the system, with a minimum distance of 12 Å from the protein, in a box of 8693 TIPS3P [386] water molecules. One chloride counterion was added to the system to neutralise the overall charge of the system.

This structure contains two solubility enhancing mutants, F185K and W131E, which were mutated back to their native forms using the SCAP [378][379] software, as it has been suggested that they may cause a deformation of the native structure. A previous experimental study has suggested that the F185K mutant resulted in the mutant protein being more active than the WT [387].

All simulations, unless otherwise stated, have been carried out using the NAMD [170] molecular dynamics package and the CHARMM27 forcefield [16].

Minimisation was carried out in stages, starting with the protein only (5,000 steps), followed by solvent (30,000 steps), ions (1,000 steps), solvent and ions (20,000 steps) and finally the entire system (40,000 steps), giving a total of 96,000 steps. Two minimisation algorithms are used, initially the steepest descent algorithm, followed by the conjugate gradient method.

The minimised system was heated to 300 K in the NVT ensemble. The procedure employed a Langevin thermostat with a 10 ps^{-1} damping parameter and a 2 fs timestep. The heating was carried out gradually in stages at 50 K intervals, each interval being 20,000 steps long.

Equilibration simulations, using a Nose-Hoover barostat in the NPT ensemble were then carried out for 50,000 steps, with a target pressure of 1 atm. A decay parameter of 100 fs and a piston period of 200 fs were used. A further 50,000 steps were run, with a decay parameter of 300 fs and a piston period of 500 fs.

The final equilibrated system had box dimensions of 66.57, 68.17 and 61.88 Å.

The apo mutant system, possessing the G140A/G149A double mutant has also been studied. Since there are no complete crystal structures of this double mutant which possess the Mg^{2+} ion, the equilibrated apo WT structure was taken and the appropriate residues mutated using SCAP [378] [379]. Careful minimisation and equilibration of this mutated structure was carried out before use in simulations. Initially, the mutated residues were minimised for 1,000 steps whilst restraining the rest of the system, followed by the minimisation of the rest of the protein for 5,000 steps, the solvent and counterion for 1000 steps, and lastly, the entire system for 10,000 steps. Heating and equilibration of the system was then carried out as for the apo WT structure.

The final equilibrated system had box dimensions of 66.63, 68.27 and 61.91 Å.

For each system, one MD production simulation, 20 ns in length has been carried out. All production MD simulations were run in the NVT ensemble, using a Langevin thermostat with a 5 ps^{-1} damping parameter at a temperature of 300 K. Periodic boundary conditions were used, along with a particle mesh Ewald treatment of electrostatic interactions, using a interpolation order of 6, and switching function applied to the Lennard-Jones interactions between 9 Å and the 10.5 Å cutoff. PME gridsizes of 69 x 72 x 64 Å were used, similar values to those of the boxsizes. SHAKE [35] was applied to all bonds containing hydrogen, using a tolerance of 10^{-8} Å.

9.1.2 RDFMD Simulation Details

Since the dynamics of the catalytic loop is thought to play a fundamental role in the activity of the enzyme, the residues of this region (140-149) have been selected as the target region of the filter in the RDFMD simulations.

A protocol of parameters for use on regions of interest in proteins has been optimised previously [231], and the parameters, with the exception of the temperature cap (700 K

used in this study) have been used. These include a 0-100 cm^{-1} digital filter, with 201 coefficients, an amplification factor of 2, filter delays of 50 and 100 steps and 4 ps of MD run in between the filter applications. The temperature cap was determined from a number of trial and error simulations, determining the maximum temperature which could be used without affecting the secondary structure. The final results are taken from piecing together the individual 4 ps MD runs. Each RDFMD simulation produces 100 4 ps MD sections, totalling 400 ps.

Simulations were run in the NVT ensemble use the Langevin thermostat with a 5 ps^{-1} damping parameter.

A total of 12 RDFMD simulations have been carried out using six different starting structures with different velocities. The last stage of the equilibration process was extended for a further 60,000 steps, taking the velocities and starting coordinates after every 10,000 steps, to generate each of the six different starting structures.

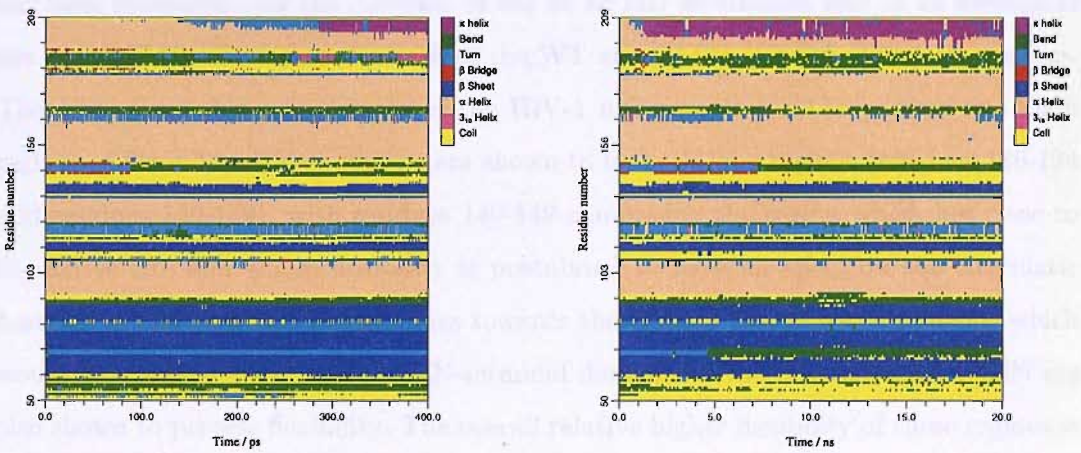
9.2 Results

For both the WT and G140A/G149A mutant HIV-1 IN systems, twelve RDFMD simulations (parameters used shown in table 9.1) and one MD simulation of 20 ns in length have been carried out. The simulations reveal conformations of the catalytic loop in this domain which may be of biological significance to the mechanism of this enzyme.

RDFMD Simulations	
Start Structure No.	Filter Delay
1	50
2	50
3	50
4	50
5	50
6	50
1	50
2	100
3	100
4	100
5	100
6	100

TABLE 9.1: RDFMD simulations: parameters used for simulations of each of the WT and G140A/G149A systems. A total of 12 simulations carried out for each system, using 6 different starting structures. An amplification factor of 2 and temperature cap of 700 K were used in all RDFMD simulations.

9.2.1 Structural Stability and Flexibility



(a) RDFMD: WT

(b) MD:WT'

FIGURE 9.1: Representative secondary structure plots for the WT HIV-1 PR structures. Parameters of plots generated by RDFMD: filter delay of 100 steps, temperature cap of 700 K, amplification of 1.8, using starting structure 1. Plots generated by MD simulation have been produced from 20 ns of simulation.

Representative plots showing the secondary structure over the duration of the RDFMD and MD simulations are shown in figure 9.1. The secondary structure plots of the other MD and RDFMD simulations of both WT and G140A/Gl149A HIV-1 IN systems are similar and thus have not been included here.

These plots demonstrate the secondary structure to be stable over the duration of the simulations and to possess the 5 β-sheet regions flanked by α-helices, as reported by structural studies [343]. The residues positioned at the end of the catalytic domain, where, in the full-length HIV-1 IN, connection to the C-terminal exists, is shown to possess some π-helical character in some of the simulations, a class of secondary structure which is similar to α-helices. This region is in the vicinity of a known proteolytic site, and is proposed to act as a flexible elbow to reorientate the domains during the process of integration [304]. Therefore, this change in secondary structure is likely to be due to this high flexibility.

Additionally, the plots show the input of energy in the filter application stages of the RDFMD method has not caused any degradation of the structures.

The root mean squared fluctuations (RMSF) of the α-carbon atoms of all the residues has been measured over the duration of the 20 ns MD simulation, and as an average of the 12 RDFMD simulations for each of the WT and G140A/G149A HIV-1 IN systems. The plots show this core domain of the HIV-1 integrase enzyme to possess two main regions of flexibility. These regions are shown to be both loop regions (residues 186-194 and residues 140-149), with residues 140-149 comprising the region which lies close to the active site and whose flexibility is postulated to have an effect on the enzymatic function of the protein. The residues towards the ends of the catalytic domain, which would be connected to the C- and N-terminal domains in the full-length HIV-1 IN are also shown to possess flexibility. The overall relative higher flexibility of these regions is consistent with the profile of the B-factors calculated from MD simulations carried out by Lee *et al.* [111].

Comparison of the RMSF of the MD simulations of the WT and G140A/G149A enzyme (figure 9.2) show the mutant HIV-1 IN to possess slightly higher flexibility of the catalytic loop region, and the WT to have slightly greater flexibility of the other mobile regions.

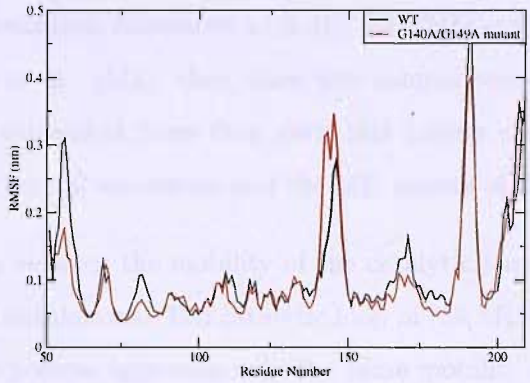


FIGURE 9.2: RMSF of the α-carbon atoms of all residues of the core domain for the WT and G140A/G149A mutant MD simulations. The catalytic loop residues are marked as thicker lines (residues 91-100).

The measurement of the RMSF over the length of the 12 RDFMD simulations of each system (figure 9.3) also displays the relative higher flexibility of the catalytic loop region, and the smaller loop comprising of residues 186-194, although, this is be expected since the digital filters are targeting these regions of the proteins.

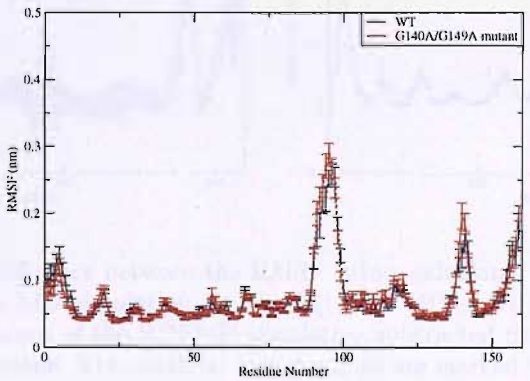


FIGURE 9.3: RMSF of the α-carbon atoms of all residues of the core domain for the WT and G140A/G149A mutant RDFMD simulations. (The standard error has been calculated for each residue for each RDFMD simulation). The catalytic loop residues are marked as dotted lines (residues 91-100).

As demonstrated in figure 9.4, which highlights the difference in mobility between the MD and RDFMD simulations of the WT and mutant simulations, the non-catalytic loop residues are significantly more mobile in the MD simulations, especially at each of the terminals and the second loop (residues 186-194), which maybe a result of the longer

length of the MD simulations compared with the RDFMD simulations. Additionally, in the study by Lee *et al.* [111], they show the comparison of B-factors of crystal structures and those calculated from their own MD results and demonstrate a large variation between the crystal structures and the MD results of these regions also.

Only small differences between the mobility of the catalytic loop are observed between the MD and RDFMD simulations. The catalytic loop in the MD simulations of the WT HIV-1 IN is shown to possess approximately the same mobility as what is observed of the loop in the RDFMD simulations (figure 9.4 (a)). Whereas, the RDFMD simulations of the mutant HIV-1 IN observe a slightly more mobile loop compared with the MD simulations. Any significance of this result is unclear since the difference is only slight.

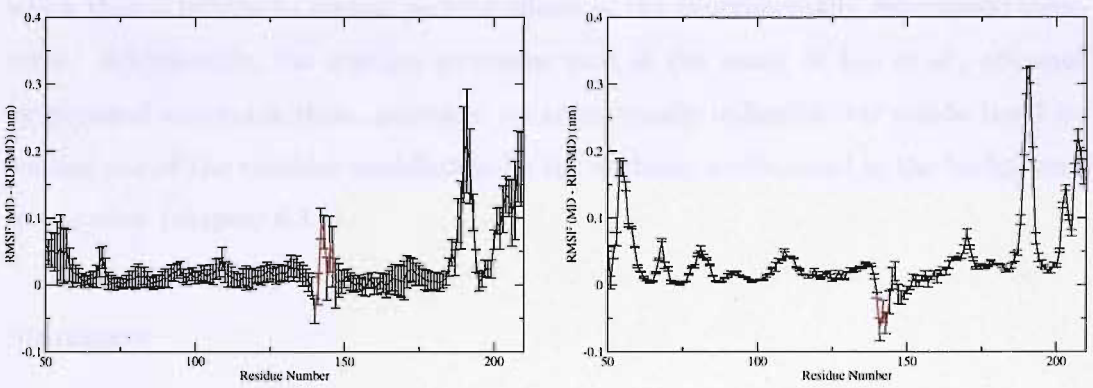


FIGURE 9.4: The difference between the RMSF values calculated for the 12 RDFMD simulations and the MD simulation for the [a]: WT HIV-1 IN [b]: G140A/G149A HIV-1 IN. RMSF values of the RDFMD simulation subtracted from those of the MD simulations. The catalytic loop residues are marked in red.

These results are in contrast to conclusions reached by previous experimental and experimental work. Previous experimental and theoretical studies [351] [111] have reported that the glycine to alanine substitution rigidifies the loop, thus reducing its mobility.

Greenwald *et al.* [351] determine the crystal structures of G149A and the double mutant G140A/G149A (the crystal structure of the G140A could not be resolved to a high enough quality for analysis) and report a reduction in temperature factors for the mutants, with the loop of the double mutant being virtually stationary. However, there maybe some ambiguity in the crystal structures used due to the use of cacodylate in the

crystallisation process, which has been reported to significantly affect the structure in the vicinity of the active site, which has been attributed to the covalent modification of C65 and C130 by the arsenic originating from the cacodylate buffer [326]. The use of this chemical also prevents the binding of the required Mg^{2+} ion, which is also known to affect the conformation of the loop. MD studies have investigated the loop in the presence and absence of the ion, and conclude that in the absence of this ion, this loop exhibits significantly increased flexibility [344].

Theoretical work carried out by Lee *et al.* [111], as described previously (chapter 6.3.4), study the conformational dynamics of the catalytic loop of the WT and G140A/G149A HIV-1 IN using conventional MD and local enhanced sampling. The authors report a reduced mobility of the mutant structures when compared to the WT. However, their calculated B-factors for the structures do not consistently compare well with experiment, which they attribute to crystal packing effects of the experimentally determined structures. Additionally, the starting structure used in the study of Lee *et al.*, obtained by personal communication, possesses an energetically unfeasible *cis* amide bond involving one of the residues modelled-in by the authors, as discussed in the background information (chapter 6.3.4).

Summary

The structural analysis shows the WT and G140A/G149A HIV-1 IN structures to remain stable over the duration of the MD and RDFMD simulations carried out in this study.

Analysis of the flexibility of the systems shows the catalytic loop to be the most flexible region of the WT and G140A/G149A HIV-1 IN systems. Other regions of the protein which also shown to possess flexibility are loop residues 186-194 and the residues at the either end of the domain which would be connected to N- and C-terminal domains in the full-length HIV-1 IN. Differences in flexibility of the overall system are seen between the MD and RDFMD simulations, with the MD simulations showing slightly higher flexibility. However, only one MD simulation of each system has been carried out, so the consistency of the MD results is not known.

In both the MD and RDFMD simulations, the catalytic loop is shown to be slightly more flexible in the G140A/G149A HIV-1 IN simulations compared with the WT, although this difference is subtle in the case of the RDFMD simulations. These results are in contrast to those of the theoretical studies of Lee *et al.* [111] and X-ray crystallography studies [351], which conclude the double mutant to posses significantly reduced flexibility, although there is some doubt as to the realiability of these structures.

9.2.2 Principal Component Analysis (PCA)

PCA has been carried out on the catalytic loop residues (residues 140-149) to investigate the conformational dynamics of this region, and to investigate any effect the G140A/G149A mutation may have on the dynamics of the catalytic loop. The non-catalytic loop residues have been disregarded for this analysis, to avoid the dynamics of region of interest being obscured by the dynamics of the rest of the protein.

The α-carbon atoms of the catalytic loop in the trajectories of each of the systems has been used as the data-set in the calculation of the covariance matrix for PCA analysis, with fitting carried out using the non-loop residues. To compare the eigenvectors of different trajectories, the same reference structure must be used. In this case the α-carbon atoms of the equilibrated WT HIV-IN system (starting structure 1) has been used as the reference structure.

Molecular Dynamics Simulations

PCA has been carried out on the catalytic loop region for the 20 ns trajectories resulting from the WT and G140A/G149A HIV-1 IN MD simulations. Ordering of the eigenvalues according to their contribution to the total motion of the protein is shown in the scree plots in figure 9.5.

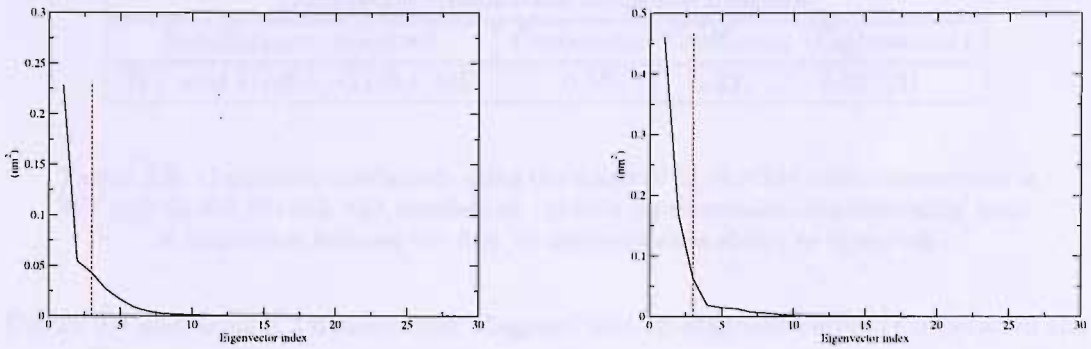


FIGURE 9.5: Scree plot depicting the proportion of motion captured in each of the eigenvectors of the loop residues. The red line depicts the eigenvalues for the first three eigenvectors.

The scree plot shows that a significant proportion of the total motion is captured within the first three eigenvectors (WT (82.8 %) and G140A/G149A (90.1 %) HIV-1 IN), indicating the loop to have large defined motions.

Comparison of the inner-products of the first 10 eigenvectors of the catalytic loop (figure 9.6) for the WT and G140A/G149A HIV-1 IN gives the level of correlation between the major motions seen in each of the eigenvectors between the MD simulations of each system. Additionally, correlation coefficients have been calculated to provide quantitative information describing the correlation between two eigenvectors, with a value of 1 demonstrating the two eigenvectors to be identical, and a value of 0 meaning the eigenvectors are orthogonal (table 9.2).

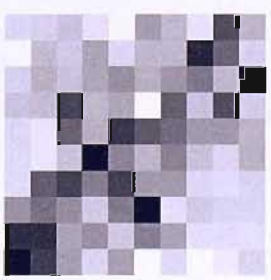
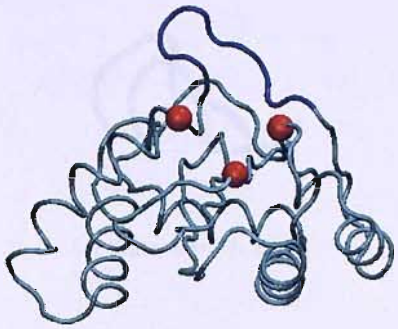


FIGURE 9.6: Comparison of inner products of the first ten eigenvectors of apo WT and G140A/G149A MD simulations of the catalytic loop. Shading corresponds to correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation (components of WT MD on x-axis, components of G140A/G140A HIV-1 IN MD on y-axis).

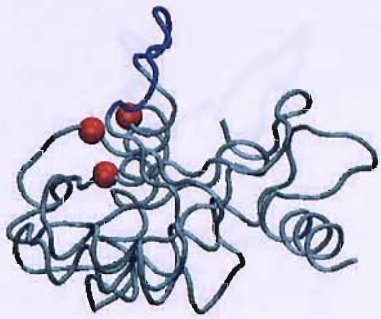
Correlation Coefficients along the Diagonal	
Simulations compared	Correlation Coefficient (Eigenvector)
WT and G140A/G149A MD	0.574 (1), 0.327 (2), 0.072(3)

TABLE 9.2: Correlation coefficients along the diagonal for the first three eigenvectors of WT and G140A/G149A MD simulations. Matrix representation demonstrating levels of correlation between the first 10 eigenvectors is shown in figure 9.6.

Figure 9.6 and table 9.2 demonstrate diagonal and off-diagonal correlation between the first few eigenvectors of the WT and G140A/G149A HIV-1 IN simulations. Visual comparison of these motions is shown in figures 9.7 and 9.8. They show the direction of motion generated by the projection of the MD simulations onto the first three eigenvectors of the WT and G140A/G149A HIV-1 IN MD simulations.

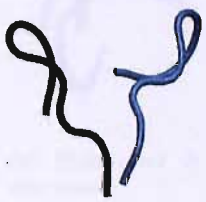


(a) HIV-1 IN: front view



(b) HIV-1 IN: side view

(c) Eigenvector 1: front view



(d) Eigenvector 1: side view

(e) Eigenvector 2: front view

(f) Eigenvector 2: side view

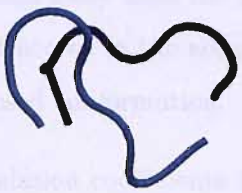
(g) Eigenvector 3: front view

(h) Eigenvector 3: side view

FIGURE 9.7: Extreme conformations sampled from the projection of the WT HIV-1 IN MD simulation on the first three eigenvectors (figures (c)-(h)). Eigenvectors of the catalytic loop generated from the WT HIV-1 IN 20 ns simulation. Figures (a) and (b) show the catalytic domain of HIV-1 IN with the atoms in red showing the location of the α-carbon atoms of the three catalytic residues (D64 D116 E152), and the region marked in blue showing the catalytic loop residues shown in figures (c)-(h).

(a) Eigenvector 1: front view

(b) Eigenvector 1: side view



(c) Eigenvector 2: front view

(d) Eigenvector 2: side view

(e) Eigenvector 3: front view

(f) Eigenvector 3: side view

FIGURE 9.8: Extreme conformations of the catalytic loop (blue and black) sampled from the projection of the G140A/G149A HIV-1 IN MD simulation on the first three eigenvectors (figures (a)-(f)). Eigenvectors of the catalytic loop generated from the G140A/G149A HIV-1 IN 20 ns simulation.

Comparison of the first eigenvectors of the WT and mutant HIV-1 IN MD simulations shows a degree of diagonal correlation (0.574) and off-diagonal correlation with the second eigenvectors of both MD simulations (eigenvector 1 of WT and eigenvector 2 of G140A/G149A HIV-1 IN: 0.319, eigenvector 2 of WT and eigenvector 1 of G140A/G149A HIV-1 IN: 0.283). Visualisation of this motion (figures 9.7 (c) and (d) and 9.8 (a) and (b)) for both the WT and G140A/G149A HIV-1 IN MD simulations shows the first eigenvector to capture the gating motion of the catalytic loop, moving towards and away from the active site. This motion is associated with the opening and closing of access to the active site, with the loop predicted to overhang the active site in the closed conformation.

The correlation coefficients demonstrate less diagonal correlation exits between the second and third eigenvectors (correlation coefficients in table 9.2), although a degree of off-diagonal correlation exits between the second and first eigenvectors of the WT and mutant simulations. The motion captured by the second eigenvector in the WT simulations appears to show the loop moving to from the side with a twisting motion incorporated. For the G140A/G149A mutant MD simulation, the second eigenvector captures the motion where the loop moves from a compact loop conformation, which leans to one side, to one where the loop becomes more extended and moves backwards, away from the active site (WT: figures 9.7 (e) and (f), G140A/G149A: 9.8 (c) and (d)). Similarity between the third eigenvectors of the WT and G140A/G149A HIV-1 IN MD simulations is very low, demonstrated by the very low diagonal correlation (0.072) and almost absence of off-diagonal correlation, indicating this motion to be less defined compared with those captured by the first two eigenvectors. Visualisation of this motion (figures 9.7 (g) and (h), G140A/G149A: 9.8 (e) and (f)) also demonstrates the motions to be quite different. The WT HIV-1 IN MD simulation shows this eigenvector to show the loop moving from a position where it leans to one side, to one where the loop points upwards. Whereas, the G140A/G149A HIV-1 IN MD simulation appears to shows the loop to move forwards, towards the active site with an incorporated twisting motion.

RDFMD Simulations

The 12 RDFMD trajectories of each system have been concatenated together to form two single trajectories (WT and mutant) on which PCA has been performed in the same manner as for the MD simulations.

The dataset for the calculation of the covariance matrix involved in this method of analysis comprises the concatenated RDFMD trajectory coordinates of the α-carbon atoms of residues 140 to 149 (loop residues) of each of the systems whilst fixing the rest of the domain. The same reference structure as used for the MD simulations has been applied here (i.e. the equilibrated WT HIV-1 IN starting structure 1).

Figure 9.9 shows the correlation between the inner-products of the first 10 eigenvectors of the WT and mutant RDFMD simulations, and table 9.3 gives the correlation coefficients representing the level of diagonal correlation between the first three eigenvectors. Since the eigenvectors are shown to be highly comparable for the first few eigenvectors, further PCA analysis has been carried out on a single trajectory incorporating all the mutant and WT trajectories together (i.e. 24 RDFMD trajectories concatenated into one).

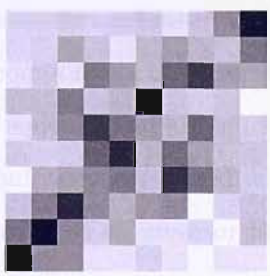


FIGURE 9.9: Comparison of inner products of the first ten eigenvectors of WT and G140A/G149A mutant HIV-1 IN RDFMD simulations. Shading corresponds to level of correlation between the eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation. Eigenvectors are plotted sequentially along the x- and y-axis, with 1 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Components of WT plotted on the x-axis and components of G140A/G149A of the y-axis.

Owing to the good correlation between the eigenvectors of the WT and G140A/G149A mutant HIV-1 IN RDFMD simulations, the total 24 RDFMD trajectories have been concatenated together to form one long trajectory, on which PCA has been performed. The concatenation of all the RDFMD trajectories to form one was also carried out for

Correlation Coefficients along the Diagonal	
Simulations compared	Correlation Coefficient (Eigenvector)
WT and G140A/G149A RDFMD	0.814 (1), 0.686 (2), 0.468(3)

TABLE 9.3: Correlation coefficients along the diagonal for the first three eigenvectors of WT and G140A/G149A RDFMD simulations. Matrix representation demonstrating levels of correlation between the first 10 eigenvectors are shown in figure 9.9.

the PCA of the apo HIV-1 systems (section 7.3.2 of the apo HIV-1 PR chapter), and as described previously, this allows the differences of sampling between the WT and mutant HIV-1 IN simulations to be displayed more clearly. Again, the covariance matrix was calculated using the trajectory coordinates of the α-carbon atoms of the catalytic loop, using the same equilibrated WT reference structure (starting structure 1) as previously used for the previous PCA carried out on this system in this chapter.

The resultant scree plot is shown in figure 9.10. The number of eigenvectors to study was based on the proportion of the total motion captured and visual inspection of the motions which they represented. As a result, the first two eigenvectors have been selected (68.2 % of total motion), and the third eigenvector has been disregarded. Although the scree plot shows the motion of this eigenvector to account for approximately 10% of the total motion, visualisation of the trajectories show the motions of the first two eigenvectors to clearly define the overall main conformations of the loop. Further motions, captured in the third and other principal components incorporate several twisting motions of the loop. These motions are higher frequency and are harder to define, demonstrated by the lower correlation between the simulations compared with the first two eigenvectors. For this reason, the motion defined by third eigenvector has not been considered further in this study.

First ten principle components of loop			
Eigenvector	Eigenvalue	% total variance	Cumulative %
1	0.347	37.1	37.1
2	0.292	31.1	68.2
3	0.095	10.1	78.3
4	0.039	4.17	82.5
5	0.035	3.74	86.2
6	0.030	3.17	89.4
7	0.029	3.11	92.5
8	0.018	1.92	94.5
9	0.010	1.04	95.5
10	0.007	0.734	96.2

TABLE 9.4: Motion captured in the first ten principal components of the catalytic loop.

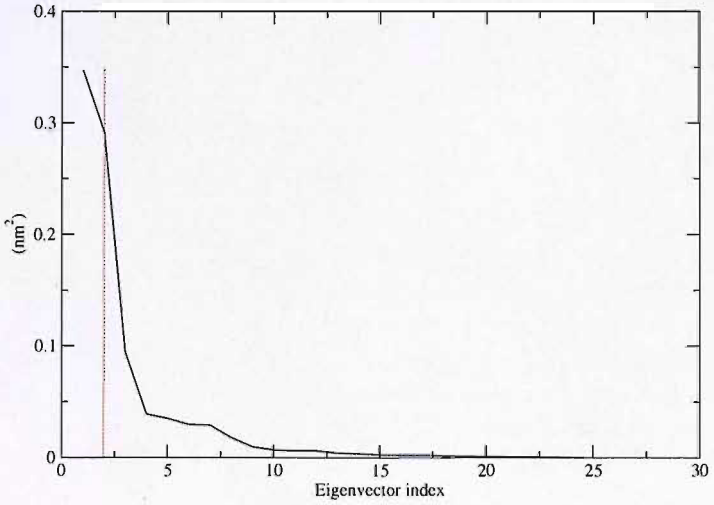
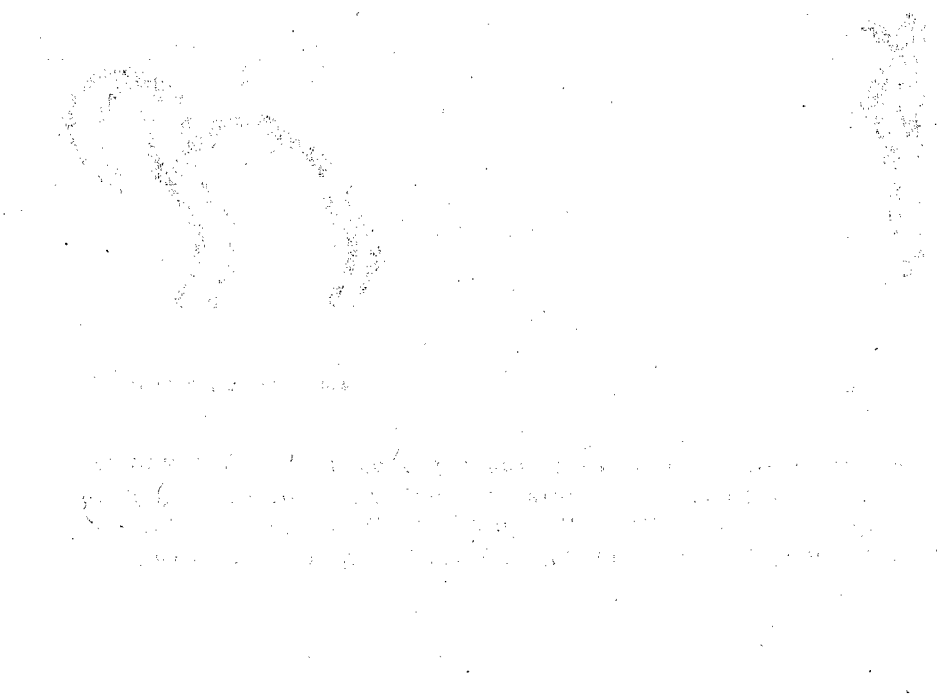


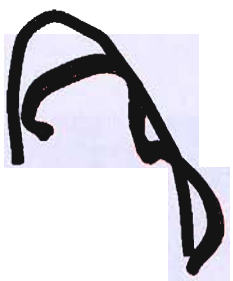
FIGURE 9.10: Scree plot depicting the proportion of motion captured in each of the eigenvectors of the loop of the concatenated WT and G140A/G149A HIV-1 IN trajectories The red line depicts the eigenvalues of the first two eigenvectors.

The WT and G140A/G149A RDFMD trajectories have been projected on the first two eigenvectors, and the motions which they represent are demonstrated by the two extreme conformations sampled in the RDFMD simulations of the WT and G140A/G149A HIV-1 IN systems (figure 9.11).

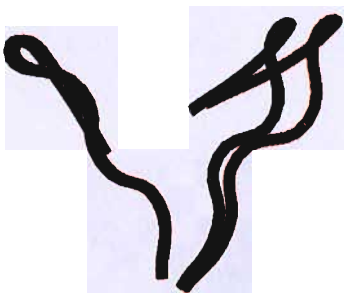
The motion represented in the first eigenvector shows the opening/closing gating movement of the loop towards and away from the active site (figure 9.11 (a) and (b)). The second principal component shows the loop moving from a more compact structure, which leans to one side, to a more extended conformation of the loop which spans a larger area (figure 9.11 (c) and (d)).

The differences between the extremes sampled by the WT and G140A/G149A HIV-1 IN simulations when projected on these eigenvectors will be investigated later in this section.

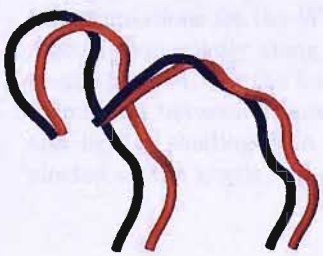




(a) Eigenvector 1: front view



(b) Eigenvector 1: side view



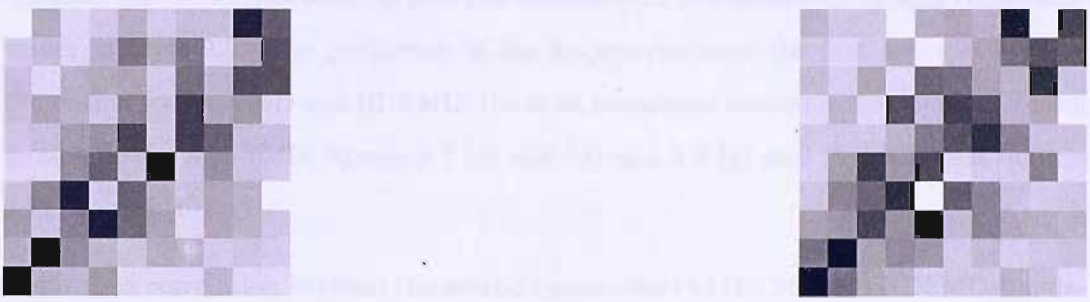
(c) Eigenvector 2: front view

(d) Eigenvector 2: side view

FIGURE 9.11: Extreme conformations sampled by the concatenated WT and G140A/G149A HIV-1 IN RDFMD trajectories projected on the first two eigenvectors (black: WT, red: G140A/G149A HIV-1 IN). Eigenvectors generated from the concatenation of the WT and G140A/G149A HIV-1 IN RDFMD trajectories.

Comparison of Eigenvectors Generated by MD and RDFMD Simulations

To compare the major motions of the MD and RDFMD simulations, comparison of the inner-products of the first 10 eigenvectors of the WT and G140A/G149A simulations has been carried out (figure 9.12) with values corresponding to the level of diagonal correlation of the first three eigenvectors shown in table 9.5.



(a) WT (b) G140A/G149A

FIGURE 9.12: Comparison of inner products of the first ten eigenvectors of RDFMD and MD simulations for the WT and G140A/G149A HIV-1 IN structures. Eigenvectors are plotted sequentially along the x- and y-axis, with 0 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Shading corresponds to the correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation. Components of RDFMD simulations are plotted on the x-axis and components of the MD simulations are plotted on the y-axis.

Correlation Coefficients along the Diagonal	
Simulations compared	Correlation Coefficient (Eigenvector)
RDFMD and WT MD	0.870 (1), 0.766 (2), 0.101(3)
RDFMD and G140A/G149A	0.754 (1), 0.600 (2), 0.279 (3)

TABLE 9.5: Correlation coefficients along the diagonal for the first three eigenvectors of WT and G140A/G149A RDFMD simulations. Matrix representation demonstrating levels of correlation between the first 10 eigenvectors are shown in figure 9.12.

Comparison of both the inner-products of the WT and G140A/G149A HIV-1 IN MD simulations with those generated by the concatenated RDFMD simulations show strong diagonal correlation between the first two eigenvectors, demonstrated by high correlation values in table 9.5. The projection of the trajectories onto the first two eigenvectors shows that for both MD and RDFMD, the most prominent motion is the opening/closing motion of the loop (MD: figures 9.7 (c) and (d) and 9.8 (a) and (b), RDFMD: figures 9.11 (a) and (b)).

The strong correlation between the second eigenvectors of the MD and RDFMD simulations demonstrates that simulations resulting from the use of both methods demonstrate the second most major motion of the simulations to show a degree of the motion where the loop moves from the compact conformation which is orientated to one side, to the more extended conformation which spans a larger area (MD: figures 9.7 (e) and (f) and 9.8 (c) and (d), RDFMD: figures 9.11 (c) and (d)).

The diagonal correlation between the third eigenvector of the MD and RDFMD simulations is relatively lower for both the WT and G140A/G149A HIV-1 IN systems. As mentioned previously, this motion is one of the higher frequency twisting motions of the loop in the RDFMD simulations, which is not believed to affect the overall motion of the loop. The poor definition of the motion which is captured by this eigenvector is demonstrated by the very low level of diagonal and off-diagonal correlation of this eigenvector.

Overall, comparison of the important eigenvectors demonstrates the loop in simulations carried out using the RDFMD technique to undergo the same types of movement (i.e loop opening/closing and the compact/extended loop conformations) as seen in the MD

simulations, thus confirming that the RDFMD method samples reasonable conformations of the catalytic loop.

Sampling of the Eigenvectors Generated by RDFMD Simulations

To compare the sampling of the MD and RDFMD simulations rigorously, the trajectories must be projected onto the same eigenvectors. For this purpose, the eigenvectors generated by the concatenated RDFMD simulations have been chosen. Although, the eigenvectors of the MD and RDFMD simulations have been shown to be similar, the eigenvectors of the RDFMD simulations have been generated through the concatenation of several simulations using several different starting structures, rather than based on just a single simulation, as with the MD simulations. Therefore, the WT and G140A/G149A HIV-1 IN MD and RDFMD trajectories have all been projected on the first two eigenvectors generated from the concatenation of the 24 WT and G140A/G149A HIV-1 IN RDFMD trajectories.

The RDFMD and MD simulations have been projected in the first two eigenvectors, with the results shown in two-dimensional plots. Sampling along the x-axis demonstrates the sampling of eigenvector 1 (motion depicted in figures 9.11 (a) and (b)), with positive values corresponding to the loop moving over the active site (closed conformation), and negative values corresponding to the loop moving backwards, away from the active site (open conformation). Values along the y-axis demonstrate the sampling of the second eigenvector, with positive values correspond to the compact loop conformation which leans to one side, and negative values corresponding to the more extended catalytic loop structure which spans a larger area (motion shown in figures 9.11 (c) and (d)).

Sampling of this two-dimensional plot will demonstrate any correlation between these two eigenvectors, and will also highlight differences in the conformations sampled by the MD and RDFMD simulations.

Figure 9.13 shows the sampling of eigenvectors 1 and 2 for the MD and RDFMD simulations of the WT HIV-1 IN system. It is clearly shown that the RDFMD simulations have sampled a greater range of conformations of the catalytic loop compared with the MD simulation, demonstrated by the larger area of the plot the RDFMD sampling covers. The sampling by the RDFMD simulations is shown to largely encompass the area

of sampling of the MD simulation, with the exception of a small region of sampling of the more positive values of eigenvector 1 (x-axis) which is not sampled by any of the 12 RDFMD simulations of this HIV-1 structure. This corresponds to the MD simulation sampling a few conformations where the loop opens to a greater extent than that seen in the 12 RDFMD simulations. Since only a relatively small amount of the conformations sampled by the MD simulations correspond to the loop being so open, with no sampling at all of this region of the two-dimensional plot by the RDFMD simulations, it is likely that the loop attaining this conformation is a rare event.

The PCA results are not consistent with RMSF data. The RMSF results did not indicate any significant difference between the mobility of the RDFMD and MD simulations. This is owing to the calculation of RMSF not giving any indication of the conformations associated with mobility.

The regions marked on the two-dimensional plot show the extremes of sampling of the first two eigenvectors. Conformations of the loop representing the sampling of these regions are shown in figure 9.14. Sampling of regions 1 and 2 (with the corresponding loop conformations shown in figure 9.14 (a), (b) and (c), (d)) represent the sampling of the extremes of the open (positive values of x-axis) and closed loop (negative values of x-axis). Sampling in regions 3 and 4 (with the corresponding loop conformations shown in figure 9.14 (e), (f) and (g), (h)) represent the conformations of the loop when the extreme values of eigenvector 2 are sampled. The positive values along the y-axis correspond to the loop in its most compact conformation, where it leans to one side (figures 9.14 (e) and (f)), and negative values correspond to the loop in its extended conformation (figures 9.14 (g) and (h)).

Generally, the sampling distribution of the first two eigenvectors shows that, when the loop is in its most closed conformation (negative values of x-axis), it is mostly in its extended conformation (negative values of y-axis), rather than its compact conformation. In contrast, when the loop is in its more open conformation (positive values of x-axis), the loop becomes less extended, assuming a more compact conformation (positive values of y-axis).

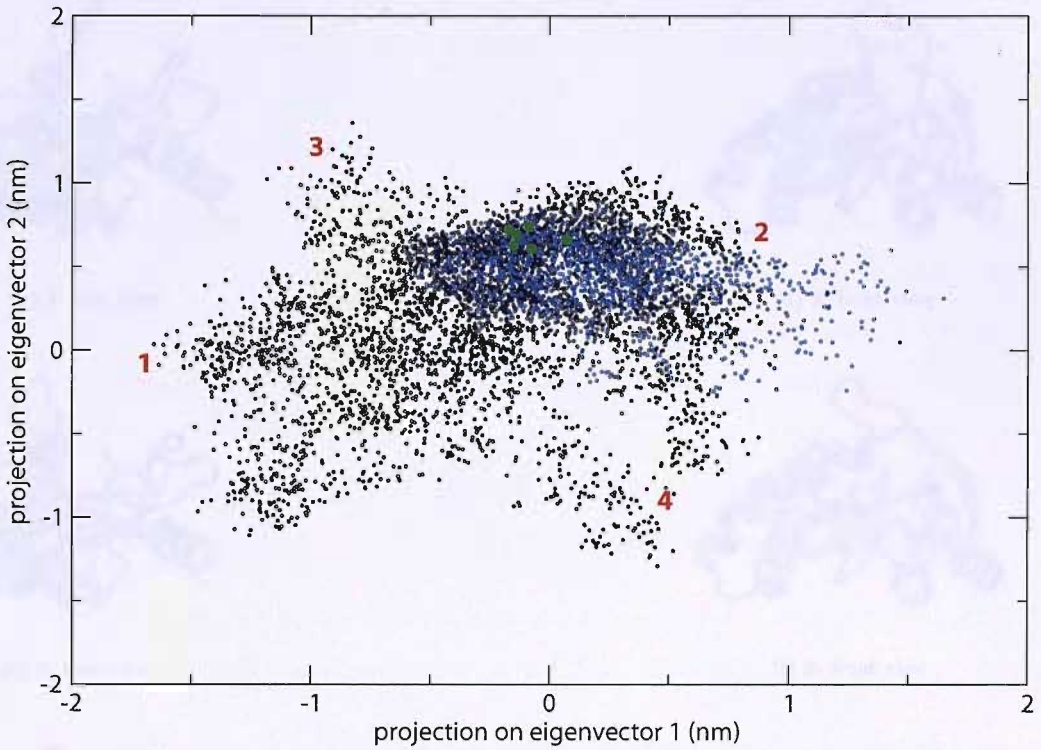
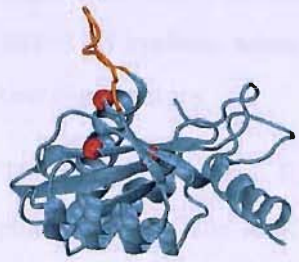
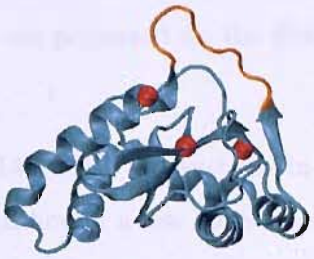


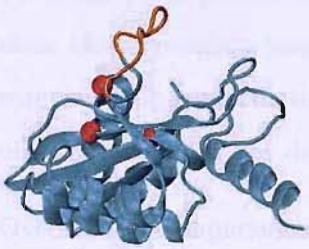
FIGURE 9.13: Sampling of the first two eigenvectors, generated from the projection of the WT RDFMD (black) and the MD (blue) simulations onto the eigenvectors of the catalytic loop created from the concatenation of all RDFMD trajectories. (The six different starting structures are represented by green circles). The numbers correspond to conformations of the catalytic loop represented in figure 9.14.



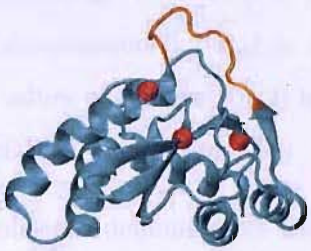
(a) 1: side view



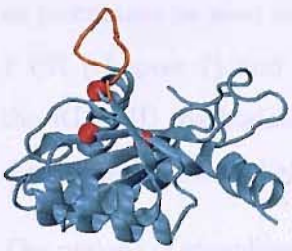
(b) 1: front view



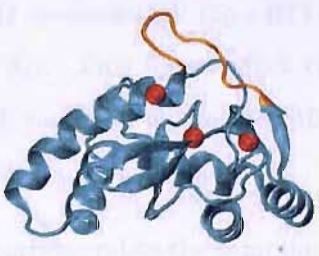
(c) 2: side view



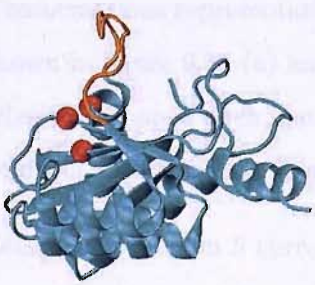
(d) 2: front view



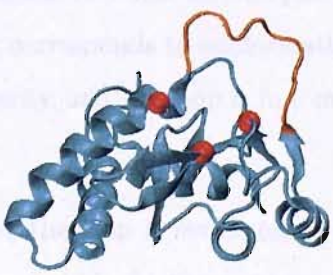
(e) 3: side view



(f) 3: front view



(g) 4: side view



(h) 4: front view

FIGURE 9.14: Conformations of the catalytic loop (shown in orange) sampled by projection of the WT HIV-1 IN RDFMD simulations on the first two eigenvectors. Eigenvectors generated from the concatentation of the WT and G140A/G149A HIV-1 IN RDFMD trajectories. Atoms shown in red represent the location of the α carbon atoms of the active site residues, D64, D116 and E152. Numbers correspond to labelling in figure 9.13.

Figure 9.15 shows the sampling of the MD and RDFMD simulations of the G140A/G149A HIV-1 IN systems when the trajectories of these simulations are projected on the first two eigenvectors.

The sampling by the RDFMD simulations of the G140A/G149A HIV-IN systems includes most of the area of sampling by the MD simulation, although a few conformations corresponding to the loop opening, moving away from the active site to a greater extent than sampled by the RDFMD simulations are observed. Additionally, a low level of sampling (above region 6) by the MD simulation corresponding to the loop in a more closed position (negative values of eigenvector 1) whilst simultaneously being in a compact loop conformation, which leans to one side (positive values of eigenvector 2) is observed, which is not demonstrated in the sampling of the RDFMD simulations.

Overall, the comparison of the sampling by the RDFMD simulations demonstrates the methodology to be able to enhance the range of conformations of the loop, compared with conventional MD. The difference in the range of conformations sampled is not as prominent as seen in previous applications of the RDFMD methodology (apo HIV-1 PR (chapter 7) and inhibitor-bound HIV-1 PR (chapter 8)). This lesser effect of the RDFMD methodology is also demonstrated in the RMSF analysis, where the MD simulations showed slightly increased flexibility of the catalytic loop (figure 9.4).

The regions of sampling marked on the two-dimensional plot correspond to the sampling of the extremes of the first two eigenvectors by the RDFMD simulations.

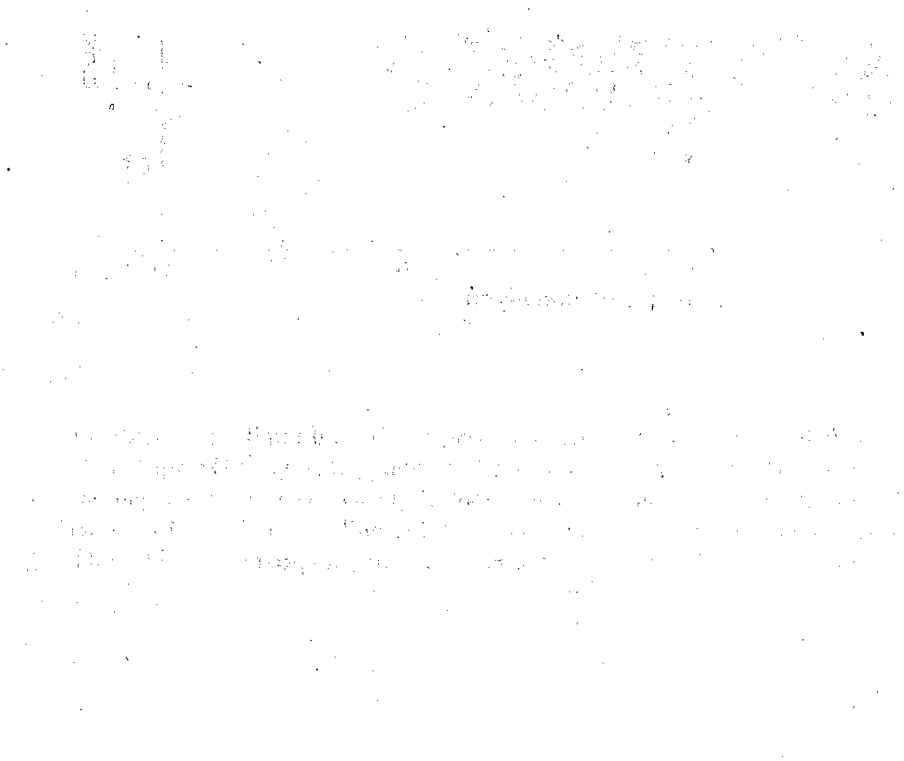
Conformations representative of the sampling of region 5 of the two-dimensional plot are shown in figure 9.16 (a) and (b). This region of sampling corresponds to conformations where the loop is open, gated away from the active site cavity, and the loop is in a more compact conformation (postive values of the y-axis).

Sampling of region 6 corresponds to conformations where the loop is overhanging the active site at its extreme (most negative values of eigenvector 1). At times where this occurs, simultaneously sampling of the more negative values of eigenvector 2 occurs, corresponding to conformations where the loop is in a conformation where is neither very compact or extended. (figures 9.16 (c) and (d)). Sampling of the less extreme negative values of eigenvector 1, where the loop is is forwards, towards the active site, the loop

is shown to be more likely to be in the more extended conformation, demonstrated by the preference for the sampling of the negative values of eigenvector 2.

Sampling of region 7 and figures 9.16 (e) and (f) (extreme sampling of eigenvector 2) show the loop in its most compact conformation, leaning to one side (positive values of eigenvector 2) whilst in a open-type conformation, bending away from the active site (positive values of eigenvector 1).

The sampling in region 8 (extreme sampling of eigenvector 2) corresponds to conformations where the loop is in its extended conformation (negative values of eigenvector 2), again where the loop is in a more open conformation, away from the active site (positive values of eigenvector 1). Visualisation of this type of conformation is shown in figures 9.16 (g) and (h).



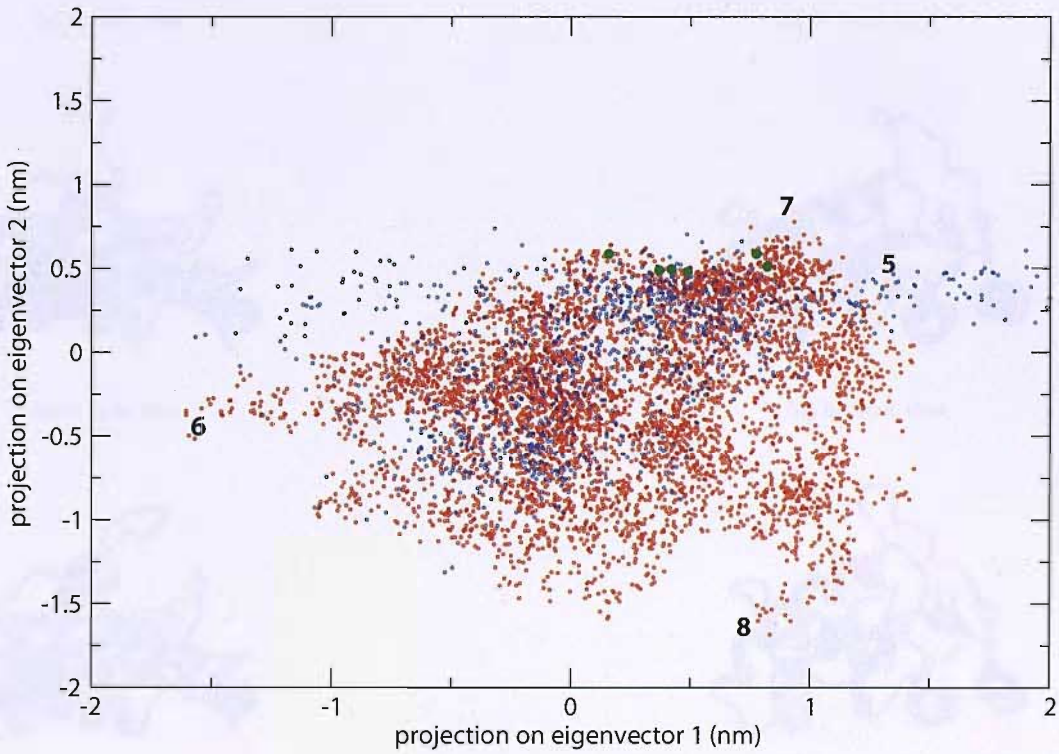
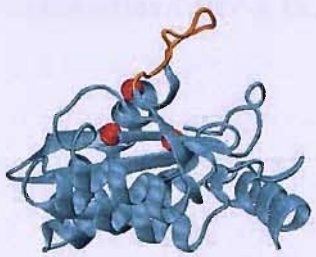
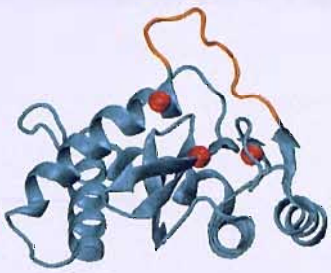


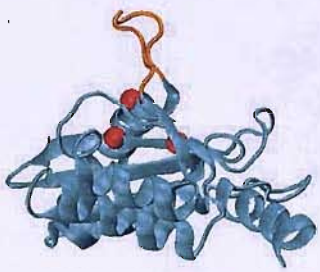
FIGURE 9.15: Sampling of the first two eigenvectors, generated from the projection of the apo G140A/G149A mutant RDFMD (red) and the MD (blue) simulations on the eigenvectors of the catalytic loop created from the concatenation of all RDFMD trajectories. (The six different starting structures are represented by green circles). The numbers correspond to conformations of the catalytic loop represented in figure 9.16.



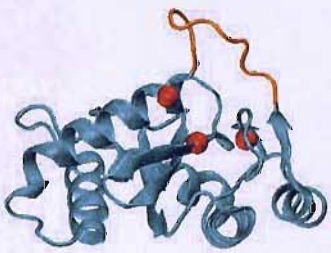
(a) 5: side view



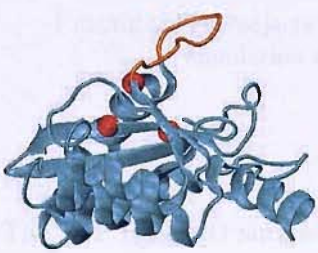
(b) 5: front view



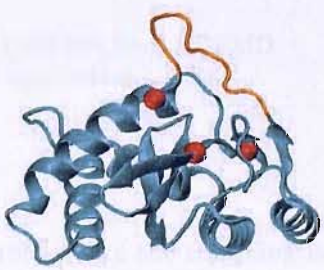
(c) 6: side view



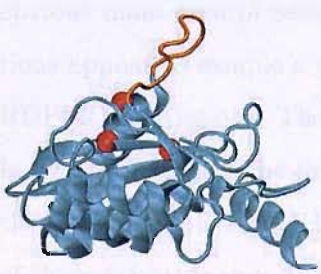
(d) 6: front view



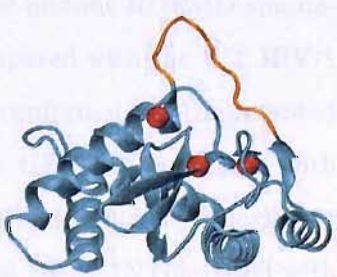
(e) 7: side view



(f) 7: front view



(g) 8: side view



(h) 8: front view

FIGURE 9.16: Conformations of the catalytic loop sampled by the projection of the G140A/G149A HIV-1 IN RDFMD simulations on the first two eigenvectors. Numbering corresponds to labelling in figure 9.15. The catalytic loop is shown in orange and the α-carbon atoms of the active site residues (D64 D116 E152) are shown as red balls.

Figure 9.17 overlays the sampling of the WT and mutant RDFMD simulations, demonstrating a difference in the sampling patterns of these two eigenvectors for the WT and G140A/G149A HIV-1 IN RDFMD simulations.

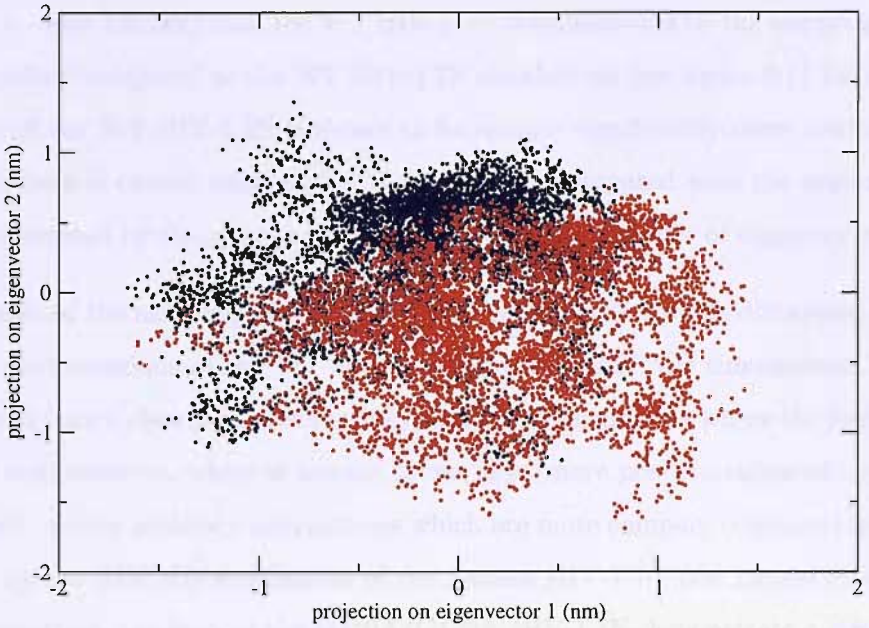


FIGURE 9.17: Projection of concatenated WT (black) and G14A/G149A (red) RDFMD simulation trajectories of the loop onto the second eigenvector.

The WT RDFMD simulations show a clear main sampling area, where the sampling is shown to be denser on the two-dimensional plot. This is not the case in the mutant RDFMD simulations, whose sampling distribution appears to be more diffuse, with no obvious main area of sampling. Additionally, the loop of the mutant RDFMD simulations appear to sample a greater range of conformations compared with the WT HIV-1 RDFMD simulations. The increased range of accessible loop conformations (represented by the sampling of the first two eigenvectors) of the mutant HIV-1 IN compared with the WT RDFMD HIV-1 IN simulations correlates with the slightly increased flexibility of the catalytic loop in the RDFMD simulations of the mutant HIV-1 IN compared with the simulations of the WT HIV-1 IN (figure 9.3).

Comparison of the sampling of the WT and G140A/G149A HIV-1 IN simulations along

eigenvector 1 clearly shows the mutant HIV-1 IN simulations to preferentially sample the positive values, and less of the negative values compared with the WT HIV-1 IN simulations. This corresponds to the loop of the mutant HIV-1 IN sampling more open-type conformations compared to the WT HIV-1 IN simulations. Additionally, the mutant HIV-1 IN is shown to be able to bend away from the active site to a greater extent (i.e. open further) than the WT HIV-1 IN, demonstrated by the sampling of more positive values compared to the WT HIV-1 IN simulations (see figure 9.11 (a) and (b)). The loop of the WT HIV-1 IN is shown to be sample significantly more conformations where the loop is closed, overhanging the active site compared with the mutant HIV-1 IN, demonstrated by the greater sampling of the negative values of eigenvector 1.

Comparison of the sampling of the second eigenvector also reveals differences between the loop conformations of the WT and G140A/G149A RDFMD simulations. The WT HIV-1 IN is shown clearly to preferentially sample conformations where the loop is in its compact conformation, where is leaning to one side (more positive values of eigenvector 2 sampled), and to achieve conformations which are more compact compared with those sampled by the RDFMD simulations of the mutant HIV-1 IN (see figures 9.11 (c) and (d)). In contrast, the loop of the G140A/G149A HIV-1 IN demonstrate a significantly greater amount of sampling of the extended conformation (negative values of eigenvector 2).

Generally, where the WT RDFMD simulations sample the most compact conformations of the catalytic loop, where it leans to one side, the loop is more likely to be in a more upright position, intermediate between fully open or closed. When the mutant RDFMD simulations are shown to sample the conformations where the loop is fully open, away from the active site (positive values of eigenvector 1), the compact conformations seen in the WT RDFMD simulations are not seen to the same extent. Additionally, where the loop is seen to overhang the active site in the closed position (negative values of eigenvector 1), the loop is shown to more likely sample the second eigenvector corresponding to slightly extended loop conformations. This is shown in the figures 9.14 (b) and (c). This correlation of sampling indicates that, in order for the loop to move into the very closed conformation, which is postulated to be an active conformation, in closer proximity to the active site residues (D116, D64 and E152), the loop cannot exist in the

very compact conformations, which are associated with the loop leaning to one side; instead the loop must be at least slightly extended.

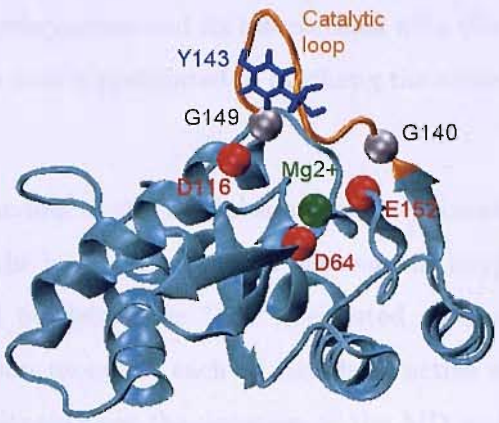
Similarly, when the loop is sampling the more open conformations, as observed for the G140A/G149A HIV-1 IN simulations, again the loop cannot be in the compact conformation, which leans to one side, the loop must become extended.

In summary, the loop of the WT G140A/G149A RDFMD simulations can open further than the WT and preferentially samples this conformation, whereas the loop of the WT HIV-1 IN RDFMD simulations is able to close over the active to a greater extent.

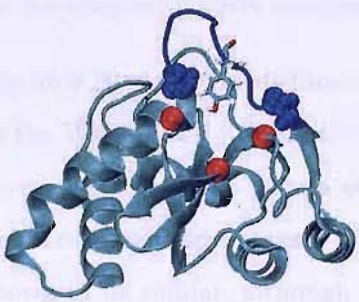
The ability of the G140A/G149A HIV-1 IN loop to bend back, away from the active site to a greater extent than seen in WT HIV-1 IN loop is also noted in the MD study of this mutant by Lee *et al.* [111].

9.2.3 Role of Tyrosine 143 (Y143)

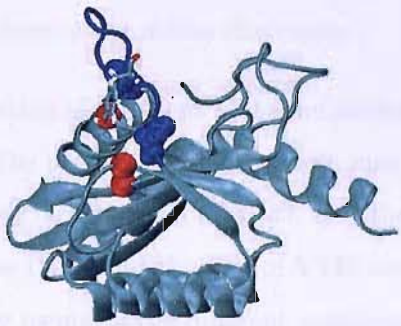
It has been predicted that the Y143 residue plays an important role in the catalytic activity of this core domain. Based on the activity of 3'-5' exonuclease of *E. coli polymerase I* [352], it has been suggested that this residue may be involved in the stabilisation and in guiding the nucleophile through a hydrogen-bond interaction, thus assisting in the catalysis of the hydrolytic and phosphoryl transfer reactions. The different conformations of the Y143 sidechain in the active and inactive forms of the enzyme are demonstrated, with the inactive form displaying the sidechain to point away from the active site to a greater extent. The assumed active sidechain conformation of this residue points downwards towards the active site (figure 9.18 (a) and (b)).



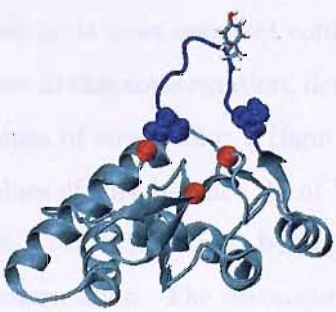
(a)



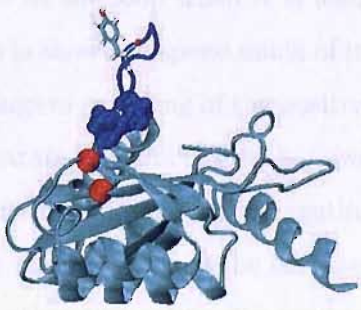
(b) WT: front view



(c) WT: side view



(d) G140A/G149A: front view



(e) G140A/G149A: side view

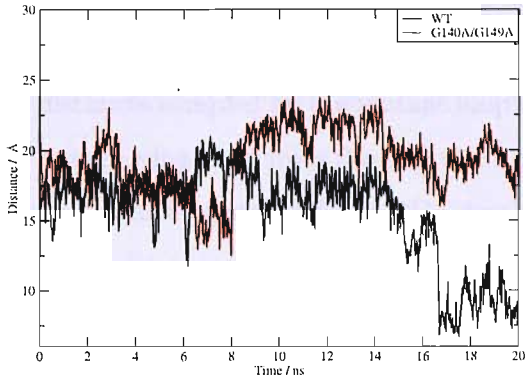
FIGURE 9.18: Conformations where the Y143 side chain lies closest and furthest from the active site. Figure [a]: HIV-1 IN. α-carbon atoms of active site (DDE) residues: red, catalytic Y143: blue, loop: orange, Mg^{2+}: green, α-carbon atoms of mutated G140A/G149A residues: silver; [b] and [c]: the sidechain of Y143 pointing towards the active site in WT HIV-1 IN; [d] and [e]: the sidechain of Y143 pointing away from the active site in the G140A/G149A HIV-1 IN.

As mentioned previously, the catalytic mechanism of HIV-1 IN is not well understood and the active loop conformation and its interactions with the substrate, are not known in detail, although the loop is postulated to overhang the active site in the active conformation.

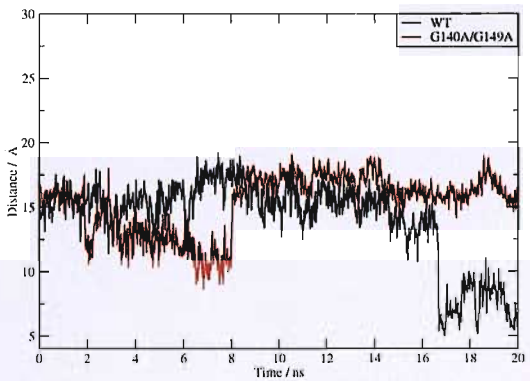
To investigate the behaviour of the Y143 residue and any correlation between its position and the dynamics of the loop, the distances between the oxygen atom of the hydroxyl group (this atom will be defined as "OH"), situated on the ring sidechain group of Y143, and the α-carbon atoms of each of the three active site residues (D64, D116, E152) have been monitored over the duration of the MD and RDFMD simulations of the WT and G140A/G149A HIV-1 IN simulations. The distances between Y143 and have been measured separately for each of the active site residues in an attempt to gain a more detailed understanding of the behaviour of the position of the Y143 sidechain, as the three active site residues span the β-sheet structure of the active site cavity.

Figure 9.19 shows the distances measured over the duration of the 20 ns MD simulations of the WT and G140A/G149A HIV-1 IN structures. The profiles of the distances measured between the α-carbon atoms of residues D64 and the OH atom of Y143, and the distance measured between α-carbon atoms of residues D116 and the OH of Y143 are shown to be similar, although the actual distances vary owing to the different positions of the active site residues (figures 9.19 (a) and (b)). The profile of the distance measured between the α-carbon atom of residues E152 and Y143 is less similar (figure 9.19 (c)). This is expected as the E152 residue is on the same side as the loop when it is leant over in its more compact conformation and the simulation is shown to spend much of its time in this conformation, demonstrated by the narrow range of sampling of the positive values of eigenvector 2 (figure 9.13). This is also demonstrated by the relatively lower values of this distance (O of Y143 and α-carbon atom of residues E152), demonstrating the Y143 sidechain to be pointed towards this active site residue when in the compact conformation. The distances between the Y143 and the other two active site residues are representative of the distances where the loop is in its more extended conformation, where it is not leaning to one side.

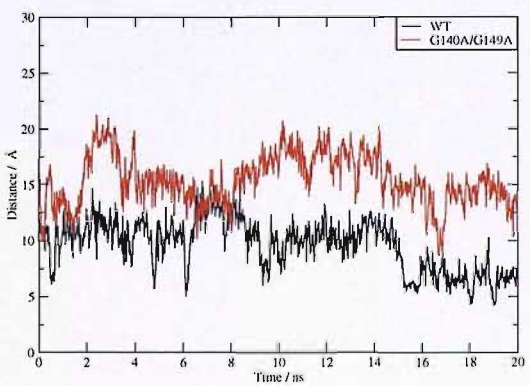
The MD simulation of the WT HIV-1 IN is clearly shown to sample smaller distances between the α-carbon atoms of the active site residues and the Y143 OH atom compared



(a)



(b)

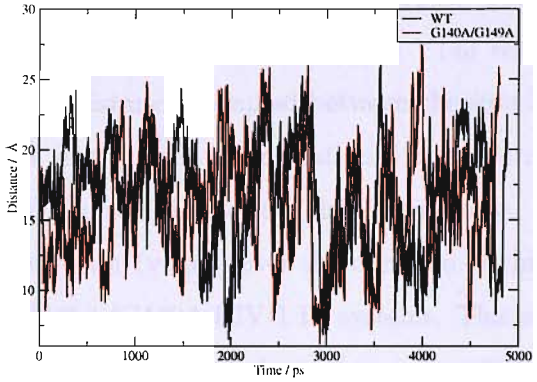


(c)

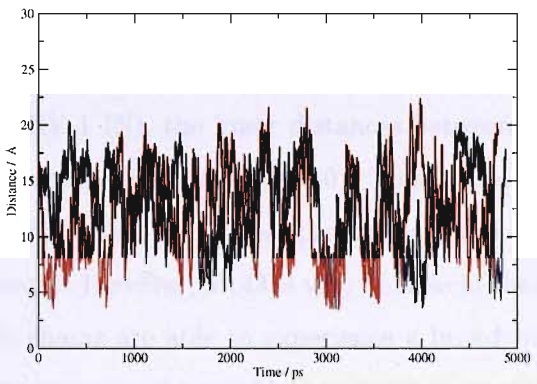
FIGURE 9.19: Distance measured between the OH atom of Y143 and α-carbon atom of the active site residues [a] D64 [b] D116 and [c] E152 over the 20 ns MD simulations of the WT and G140A/G149A mutant simulations

with the G140A/G149A HIV-1 IN indicating that the loop of the G140A/G149A HIV-IN enzyme is not able to close to the extent seen in the simulations of the WT HIV-1 IN . Additionally, the larger distances sampled by the mutant loop demonstrates its ability to open further, with the loop moving away from the active site to a greater extent than the WT HIV-1 IN simulations which agrees with the PCA results.

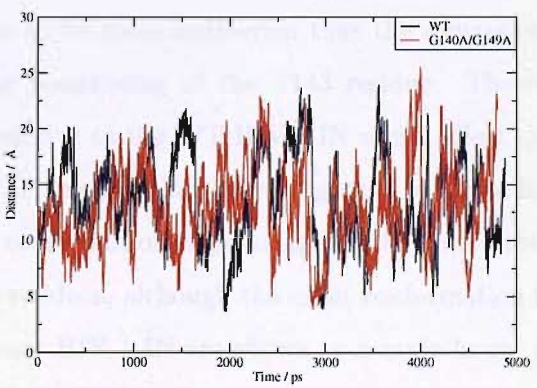
Figure 9.20 shows the distances measured over the 12 RDFMD simulations carried out for each of the WT and G140A/G149A HIV-1 IN systems. A similar trend is seen when comparing the distances of the different active site residues and the catalytic Y143 in terms of the times where the smallest (figures 9.18 [b] and [c]) and largest distances (figures 9.18 [d] and [e]) are measured. As mentioned previously, differences in the values are seen due to the different locations of the active site residues. The results are harder to interpret compared to the MD simulations since the loop undergoes significantly increased frequency of conformational change with the use of the RDFMD methodology. No conclusions can be reached concerning differences in the orientation of the Y143 residue between the WT and mutant HIV-1 IN simulations since the differences are subtle and the exact active conformation of the loop is unknown.



(a)



(b)



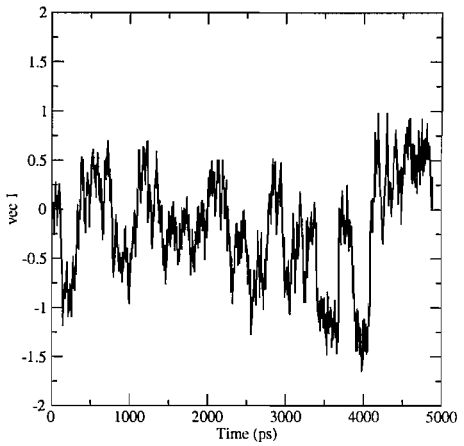
(c)

FIGURE 9.20: Distance measured between the OH atom of Y143 and α-carbon atom of the active site residues [a] D64 [b] D116 and [c] E152 over the 12 RDFMD simulations of the WT and G140A/G149A HIV-1 IN simulations

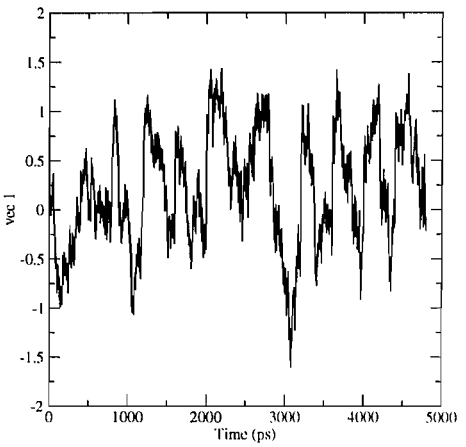
The ability of the loop of the mutant HIV-1 IN to open, away from the active site can be seen to have an effect on the proximity of the Y143 residue to the active site, demonstrated by the larger distances sampled between the sidechain of Y143 and each of the active site residues in the open conformations. Some correlation exists between the distance sampled and the sampling of the first eigenvector. Figure 9.21 shows the sampling of the first eigenvector (y-axis) over the duration of the RDFMD simulations (x-axis) of the WT and G140A/G149A HIV-1 IN systems. This eigenvector was created from the concatenation of the RDFMD trajectories, as described earlier in this chapter (section 9.2.2). Sampling of positive values of this eigenvector corresponds to open conformations of the loop, and negative values to closed conformations, where the loop is overhanging the active site.

Figure 9.21 shows that at times where the loop is closed, demonstrated by sampling of negative values of eigenvector 1 (e.g. ~4000 ps and ~3500 ps for the WT, and ~3000 ps for the G140A/G149A HIV-1 IN), the lower distances between the active site residues and the Y143 residue are sampled (see figures 9.20). This suggests that the closing of the flap is able to bring the Y143 residue in closer proximity to the active site, where it may interact with the substrates. However, Y143 is very flexible in the simulations of the apo HIV-1 IN, and the side chains are able to experience a broad range of conformations, and perhaps requires the presence of a substrate to assist in its stabilisation of a suitable conformation.

Overall, there is shown to be some indication that the dynamics of the catalytic loop may play a role in the positioning of the Y143 residue. Therefore, a change in the dynamics of the loop relative to the WT HIV-1 IN would affect the catalytic activity of the domain, since the Y143 residue has been suggested to play a fundamental role in the mechanism. Its active conformation is postulated to be where the sidechain is pointing towards the active site residues, although the exact conformation is not yet known. The simulations of the mutant HIV-1 IN are shown to sample larger distances between the active site residues and the Y143 sidechain compared with the WT simulations. This correlates with the PCA results, which demonstrate the loop of the G140A/G149A HIV-1 IN to be able to open to a greater extent, away from the active site, thus positioning



(a) WT HIV-1 IN



(b) G140A/G149A HIV-1 IN

FIGURE 9.21: Projection of WT and G140A/G149A HIV-1 IN RDFMD simulations on the first eigenvector. Eigenvectors generated from the concatenation of the WT and G140A/G149A HIV-1 IN RDFMD trajectories.

the sidechain of the Y143 residue further from the active site. Some correlation is also seen between the closing of the flap and smaller distances measured between the Y143 residue and the active site residues, indicating that the closing of the loop plays a role in positioning the catalytic Y143 nearer to the active site.

9.3 Conclusions

The RDFMD technique is shown to enhance the range of conformations sampled for the HIV-1 IN apo-enzyme compared with MD simulations. The results highlight differences in the conformational dynamics of the catalytic loop between the WT and G140A/G149A HIV-1 IN which may explain the diminished disintegration activity observed in the presence of the G140A/G149A mutant. The results indicate agreement with previous suggestions proposing the importance of the Y143 residue in the catalytic mechanism, and the function of the loop to position this residue in the correct orientation for the functional form [353] [354] [111] [355]. However, unlike previous studies, the mobility of the mutant loop is not reduced compared to the WT, RMSF results indicate the mobility to be largely similar, with a slightly increased mobility seen in the residues of the catalytic loop.

Previous experimental [351] and theoretical studies [111] have concluded that the reduced catalytic activity of the disintegration process carried out by the core domain in the presence of the G140A/G149A mutation is due to significantly reduced mobility of the loop. However, there is some doubt concerning the reliability of the structures used in these studies. The theoretical study by Lee *et al.* [111] uses a starting structure possessing an energetically unfeasible *cis* amide bond involving one of the loop residues modelled in the author (between residues G140 and I141). In addition, the conclusions reached in the experimental study of Greenwald *et al.* [351] are based on comparisons between the B-factors of X-ray crystal structures with poor resolution of the loop owing to the use of cacodylate in the cystallisation process. The use of this chemical is known to affect the conformation of the active site residues and prevent the binding of the catalytically important metal ion [326]. The dynamics of the loop has also been demonstrated to be affected as a consequence [344] [345] [346].

The results of this study indicate the mechanism for the diminished catalytic activity could be due to a difference in the equilibrium between the conformations for the different stages of the activity of the enzyme. As mentioned, the active conformation of the enzyme involves the Y143 sidechain being positioned close to the active site, with the hydroxyl group pointing downwards. The PCA results show the G140A/G149A HIV-1 IN system to sample a larger number of conformations where the catalytic loop is open. Additionally, the loop is able to open further and is not able to close to the same extent as seen in the RDFMD simulations of the WT HIV-1 IN. These differences in conformational sampling of the loop may result in the decreased likelihood of the Y143 sidechain being in a suitable location and conformation for the catalytic mechanism to take place.

The RDFMD method has provided an efficient way to produce several simulations with varying starting structures whilst enhancing the conformational change compared with the use of the MD methodology.

Chapter 10

Conclusions

The study of the conformational dynamics of biological systems is of great interest and the use of computer simulation allows proteins to be studied on an atomistic scale.

Conventional methods of computational study fail to sample the full-range of accessible conformations and many enhanced sampling methods have been developed to overcome this problem. In this thesis, the two enhanced sampling methods, parallel tempering (PT) [3] and reversible digitally filtered molecular dynamics (RDFMD) [4] [5] [6] have been applied to biological molecules.

The enhanced sampling technique of parallel tempering (PT) involves several independent simulations of the same system being prepared at a range of different temperatures, and exchanges with a replica of neighbouring temperature being attempted, according to the Metropolis criterion, to ensure detailed balance is conserved. The use of the higher temperature simulations overcomes the energy barriers, whereas the simulations at the lower temperatures enable a thorough sampling of the minima, thus overcoming the energy barriers separating the energy minima, allowing all of phase space to be sampled. In chapter 5, this method has been applied to the study of *trans-cis* isomerism of two small proline-containing cyclic peptide systems, cPAPAA and cFFAiLP. Small cyclic peptides are of interest to the pharmaceutical industry as their constrained nature results in improved binding to receptor interfaces and enhanced specificity compared with linear molecules [152]. However, the time-scale of the isomerism event is long (typically $10 - 10^3$ s), corresponding to a barrier height of 15-20 kcal/mol [147]

[148] [149]. Therefore, this event proves difficult to study experimentally and virtually impossible to study by conventional MD.

For comparison purposes, two MD simulations of each of the cyclic peptides were carried out in addition to the PT simulation. In each case, the PT simulation demonstrated a significantly greater amount of sampling of conformational phase space compared with the MD simulations, although the technique was found to be computationally expensive.

For both cyclic peptides, the prolyl peptide bonds were found to isomerise from their *trans* starting conformations, to *cis* at 300 K, owing to the nature of the proline residue, whereas the other peptide bonds were found to predominately remain in the energetically more stable *trans* conformation. A single conformation of cPAPAA was shown to exist in DMSO solvent at 300 K, which is in agreement with experimental NMR data [165]. In the case of cFFAilP in MeCN, two conformations were shown to exist, one where the prolyl bond was in a *cis* conformation, and one where is was *trans*. NMR data [166] also shows the presence of the two conformations, determining them to be in almost equal proportions at 300 K. These relative proportions detected by experiment, were not demonstrated in the results of the PT simulation, which displayed the presence of only a very small proportion of the conformations where the prolyl bond was in the *trans* form. However, the good mobility and acceptance probability shown between the replicas indicates that the PT simulation was not at fault and perhaps deficiencies in the force field may be the reason for the poor agreement between the experimental and PT data for cFFAiLP.

RDFMD is able to selectively enhance the low-frequency motions of proteins, motions associated with the large conformational changes of interest. Frequency selective kinetic energy is applied to specific structural regions of the protein, amplifying the internal velocities of a few selected residues. The method has been optimised and applied to two of the HIV-1 enzymes, HIV-1 protease (HIV-1 PR) and HIV-1 integrase (HIV-1 IN). In each case, optimisation of the amplification factor and temperature cap was carried out using a number of trial and error simulations to identify suitable values which did not lead to unfeasible disruption of the structure.

HIV-1 PR is a homodimer comprising 198 residues in total. Each dimer contributes a β-hairpin flap-type structure, both of which cover the active site, their function being

to control access to and from the active site cavity. Numerous mutations have been reported for this enzyme which decrease the efficacy of current FDA approved inhibitors. These mutations are described as primary or secondary (compensatory), with primary mutations often abolishing the catalytic/inhibitory activity of the enzyme, whilst secondary mutations act by restoring the activity to a level suitable for catalytic function, whilst preventing inhibition. The mechanism by which several of these mutations operate has been attributed to a shift in the equilibrium between conformations [276]. The V82F/I84V mutation is located near the active site and deleteriously affects the activity of all current inhibitors [277] [278] [279]. Theoretical studies have demonstrated this mutant to shift the conformational equilibrium towards the semi-open conformation and for the flap to lift further from the active site than what is observed in the WT [221] [234]. This would result in the inhibitor having to pay a greater cost to close the flaps on binding. M46I is a compensatory mutation located in the flaps of the enzyme and is reported to increase catalytic activity, whilst unaffecting the affinity for inhibitors [254] [284] [255]. Experimental [256] and theoretical [218] results demonstrate the increased catalytic activity to be achieved through stabilising the flaps in the closed conformation.

Three systems of the apo HIV-1 PR enzyme have been investigated in chapter 7, these being the WT HIV-1 PR, M46I HIV-1 PR and V82F/I84V HIV-1 PR. In addition to two MD simulations on each system, 12 RDFMD simulations, starting from two different starting structures and using a variety of amplification factors and filter delays have been carried out for each system. The main method of analysis used was PCA, concentrating on the dynamics of flap 2.

In the RDFMD simulations, three flap-tip residues of each flap have been selected as the target for the digital filter and the results reveal several biologically relevant conformations, which include sampling of flap-tip curling, a variety of semi-open conformations, flap opening, and conformations near to the closed conformation, where the flap-tips are almost swapped. This range of conformational sampling was not observed in the two MD simulations. Additionally, the RDFMD simulations demonstrated significant differences between the conformational sampling of the WT and mutant apo HIV-1 PR systems, which show agreement with experimental data. The V82F/I84V enzyme was shown to display sampling of a larger range of conformations, and to preferentially sample the semi-open conformation compared with the WT simulations. Prior MD simulations

performed by Perryman *et al.* [221] demonstrate the flap-tip residues to curl more frequently in the V82F/I84V mutant HIV-1 PR compared with the WT. However, this was not observed in the RDFMD simulations carried out in this thesis, but these residues were shown to be able to curl back, away from the active site to a greater extent than the WT, an observation which, to the author's knowledge, has not been commented on before. The RDFMD simulations of the M46I mutant HIV-1 PR reveal the flaps of the mutant to preferentially sample the closed conformation compared to the WT and also to display less mobility, demonstrated by the reduced incidences where the flap moves from its starting position, away from the active site vertically or horizontally.

In chapter 8, the RDFMD method has been applied to the WT and mutant (M46I and V82F/I84V) HIV-1 PR systems when bound to saquinavir (SQV) and indinavir (IDV), targeting the same flap-tip residues as in the RDFMD simulations of the apo HIV-1 PR. The aim of this chapter was to investigate differences between the conformational sampling between the WT and mutants and between HIV-1 PR when bound to different inhibitors.

The presence of the M46I and V82F/I84V mutations have been reported when HIV-1 PR is bound to the inhibitor SQV [277] [278] [279], whereas in the case of IDV binding, the existence of the V82F/I84V double mutant, but not the M46I mutant has been recorded. In addition, studies have shown the V82F/I84V mutant to affect the binding affinity of the SQV-bound HIV-1 PR enzyme to a greater extent than the binding of IDV [269]. The inhibitor IDV has been selected for study to demonstrate whether there is any variation between the dynamics of the flaps when the inhibitors are in the presence of effective and ineffective mutations. As with the apo HIV-1 PR, 12 RDFMD simulations and two MD simulations were carried out for each system, focusing on the dynamics of the flap 2, using PCA as the main method of analysis.

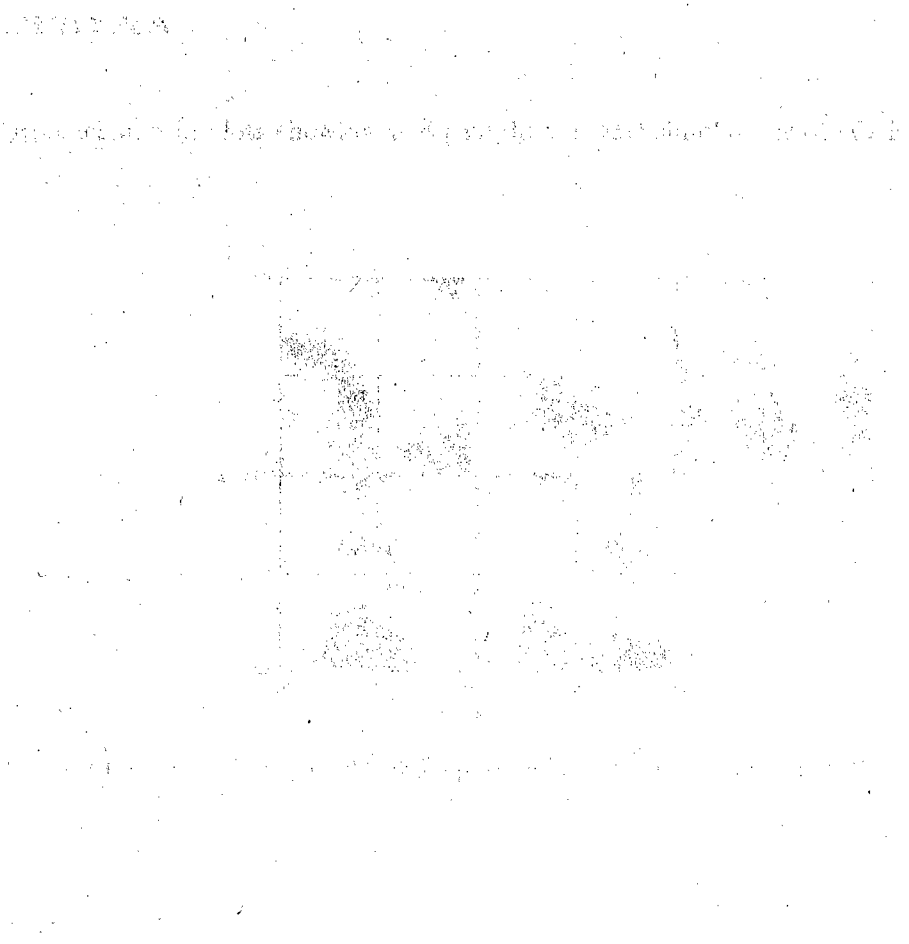
The RDFMD simulations display significantly reduced conformational change compared with the apo-structures, owing to the interactions which exist between the inhibitor in the active site and the flaps. Although a greater amount of conformational sampling is demonstrated by the RDFMD simulations compared with the MD simulations, the limited motion of the flaps has resulted in difficulty in concluding differences between the conformational sampling of the WT and mutants. However, PCA has demonstrated

an overall difference in the dynamics of the flaps depending on whether SQV or IDV is bound, with simulations of IDV-bound HIV-1 PR indicating flap 2 to be able to move away from the protein in a horizontal gating motion, to a further extent than seen in the HIV-1 PR SQV-bound simulations. This difference is due to the differences in the position and binding of the two inhibitors.

In chapter 9, RDFMD has been applied to a second HIV-1 enzyme, HIV-1 integrase (HIV-1 IN) in the WT and mutant (G140A/G149A) apo-forms. Less is known about this enzyme compared with HIV-1 PR since more limited structural information exists. The complete enzyme comprises three domains, but no reliable structure of the complete enzyme exists. Instead, the focus of study has been on the core domain, which has been demonstrated to initiate the process of disintegration in isolation [317] [318]. This domain possesses a catalytic loop which overhangs the active in the closed position and is bent away in the open conformation. A tyrosine residue exists at the top of the loop which has been predicted to play a fundamental role in the catalytic mechanism, with the conformation of the loop linked to the correct orientation of this residue with respect to the active site [111].

RDFMD simulations, targeting the loop residues, have revealed a difference in the sampling between the WT and mutant forms of this enzyme which may explain the loss of catalytic activity in this enzyme. Like the apo HIV-1 PR simulations, these simulations demonstrate increased amount of conformational phase space sampled compared with MD experiments, reversibly sampling open and closed conformations of the loop. The mutation, G140A/G149A involves the two hinge residues of the loop and experimental and theoretical studies have shown the mutation to significantly decrease the catalytic activity by rigidifying the important catalytic loop compared with the WT structure [351] [111]. However, there is some doubt as to the reliability of the structures used in these studies. The results of the RDFMD simulations in this thesis do not suggest the mutation to rigidify the loop. Instead, the mutant HIV-1 IN is demonstrated to preferentially sample open conformations, which would result in the fundamental Y143 residue being further from the active site, and also unable to sample the closed conformation to the same extent as the WT, a conformation predicted to be that of the active form.

In summary, the enhanced sampling techniques used in this thesis have provided access to a greater amount of conformational phase space compared with conventional methods. The PT methodology is a computationally expensive method, which is perhaps not suited in application to large protein systems. The RDFMD methodology has provided an efficient way to generate biologically relevant conformations, which have provided an insight into the effect of mutation on protein dynamics.



Appendix A

Cyclic Peptides

cPAPAA

Ramachandran plots showing ψ/ϕ sampling of MD simulations of cPAPAA.

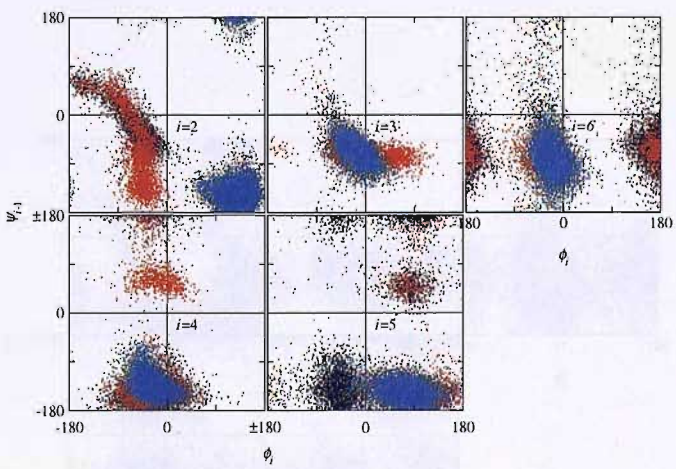


FIGURE A.1: Backbone dihedral space sampled by replicas at 611.8 K over the 20 ns of MD simulation

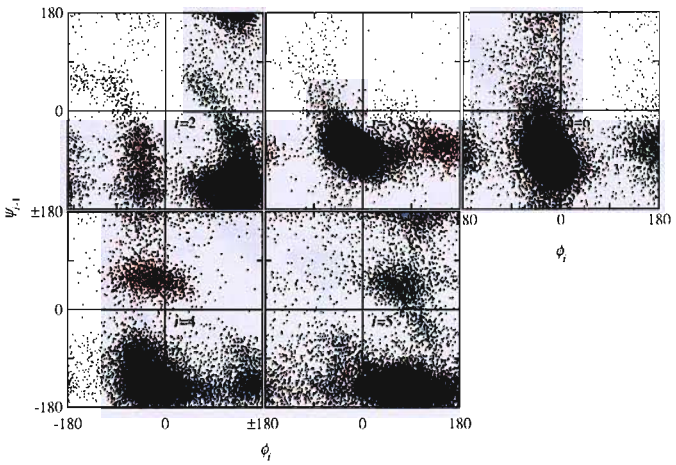


FIGURE A.2: Backbone dihedral space sampled by replicas at 933.8 K over the 20 ns of MD simulation

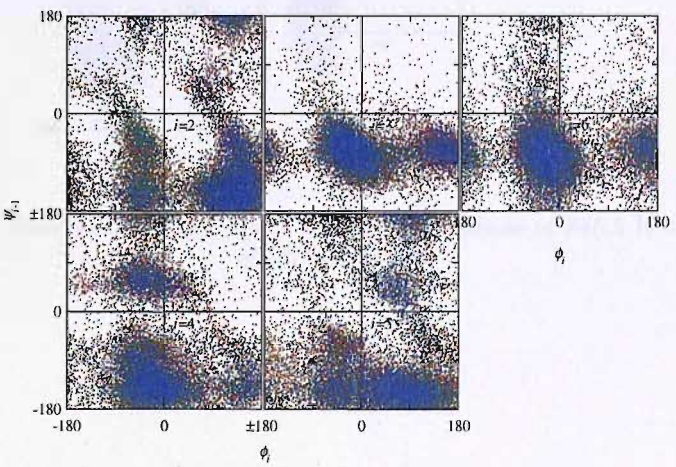


FIGURE A.3: Backbone dihedral space sampled by replicas at 1210.6 K over the 20 ns of MD simulation

cFFAiLP

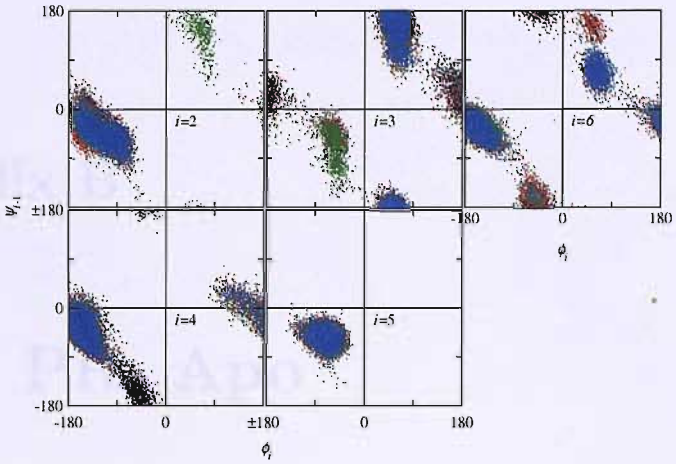


FIGURE A.4: Backbone dihedral space sampled by replicas at 501.9 K over the 20 ns of MD simulation

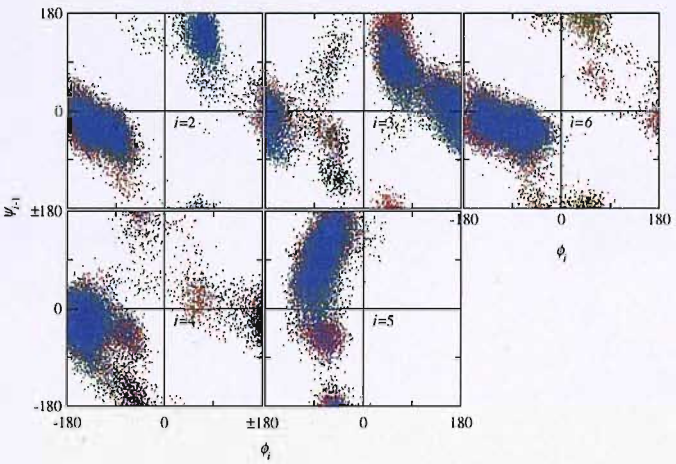


FIGURE A.5: Backbone dihedral space sampled by replicas at 846.5 K over the 20 ns of MD simulation

Appendix B

HIV-1 PR: Apo

PCA: Analysis of Eigenvectors

Visualisation of the MD trajectories onto the first three eigenvectors.

(a) First principal component - side view

(b) First principal component - top view

(c) Second principal component - side view

(d) Second principal component - top view

(e) Third principal component - side view

(f) Third principal component - top view

FIGURE B.1: The first three principal components of flap 2 created from the MD 1 simulation of the WT HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling. The region coloured in blue in (a) and (b) are the flap residues, shown in figures (c)-(h).

(a) First principal component - side view

(b) First principal component - top view

(c) Second principal component - side view

(d) Second principal component - top view

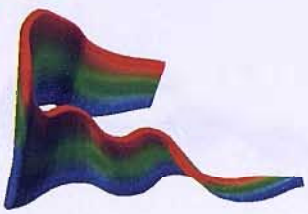
(e) Third principal component - side view

(f) Third principal component - top view

FIGURE B.2: The first three principal components of flap 2 created from the MD 2 simulation of the WT HIV-1 PR system WT. Colour grading corresponds to values depicting the extent of sampling.

(a) First principal component - side view

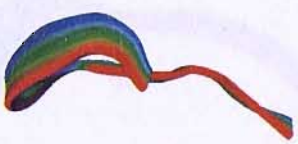
(b) First principal component - top view



(c) Second principal component - side view

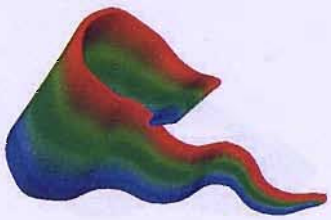
(d) Second principal component - top view

(e) Third principal component - side view



(f) Third principal component - top view

FIGURE B.3: The first three principal components of flap 2 created from the MD1 simulation of the apo M46I mutant HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling.



(a) First principal component - side view

(b) First principal component - top view

(c) Second principal component - side view

(d) Second principal component - top view

(e) Third principal component - side view

(f) Third principal component - top view

FIGURE B.4: The first three principal components of flap 2 created from the MD2 simulation of the apo M46I mutant HIV-1 PR system. Colour grading corresponds to values depicting the extent of sampling.

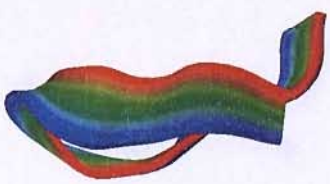
(a) First principal component - side view

(b) First principal component - top view

(c) Second principal component - side view

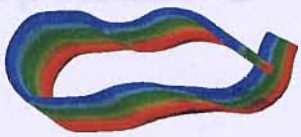
(d) Second principal component - top view

(e) Third principal component - side view



(f) Third principal component - top view

FIGURE B.5: The first three principal components of flap 2 created from the MD1 simulation of the apo V82F/I84V mutant HIV-1 PR system. Colour grading corresponds to values depicting the amplitude of motion, red corresponds to the conformation of the flap at high values, blue corresponds to the conformation of the flap at low values, and green corresponds to intermediate conformations



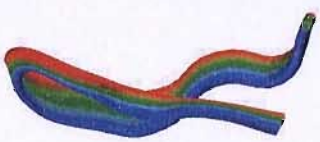
(a) First principal component - side view

(b) First principal component - top view

(c) Second principal component - side view

(d) Second principal component - top view

(e) Third principal component - side view

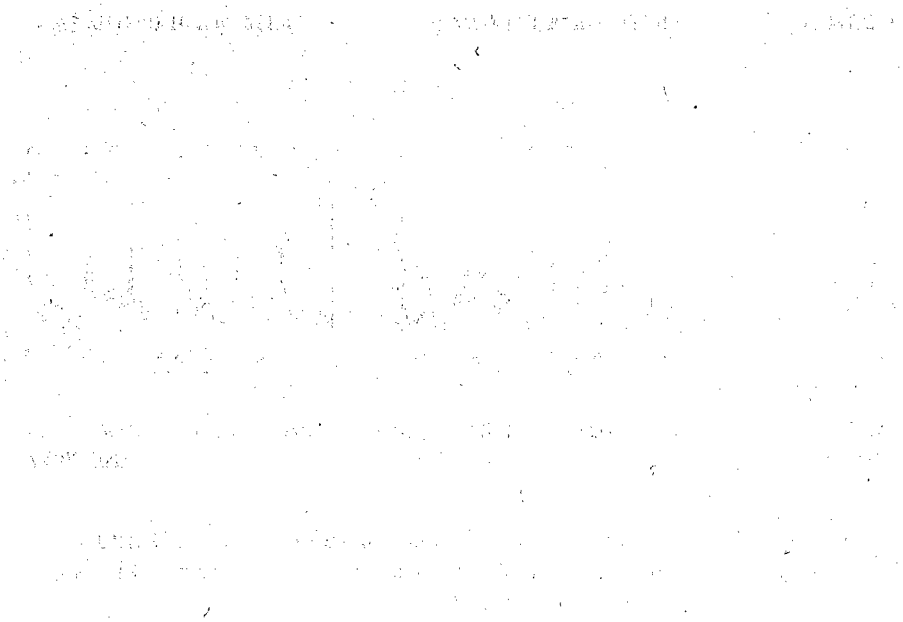


(f) Third principal component - top view

FIGURE B.6: The first three principal components of flap 2 created from the MD2 simulation of the apo V82F/I84V mutant HIV-1 PR system. Colour grading corresponds to values depicting the amplitude of motion, red corresponds to the conformation of the flap at high values, blue corresponds to the conformation of the flap at low values, and green corresponds to intermediate conformations

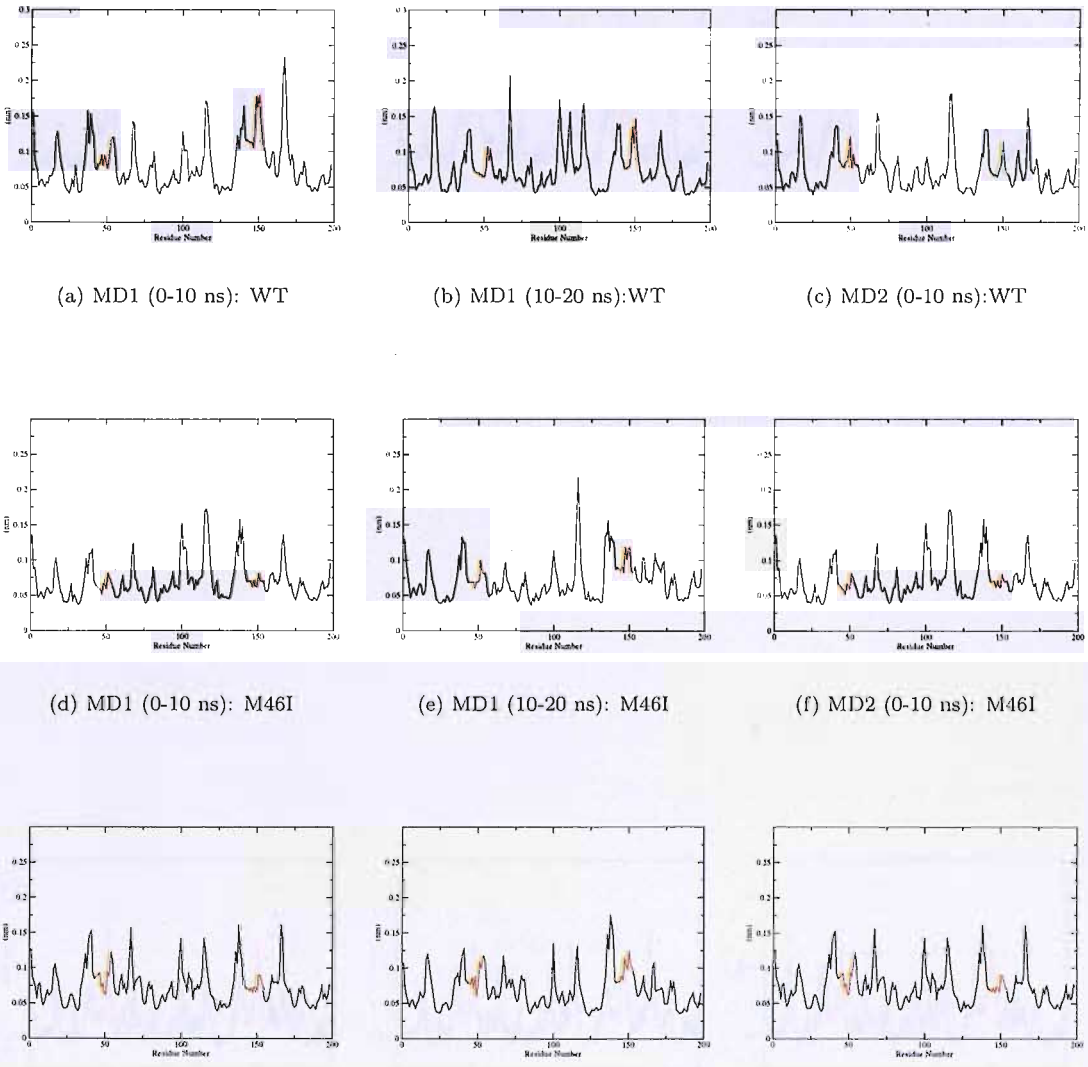
Appendix C

HIV-1 PR: Inhibitor Bound



Structural Stability and Flexibility

RMSF

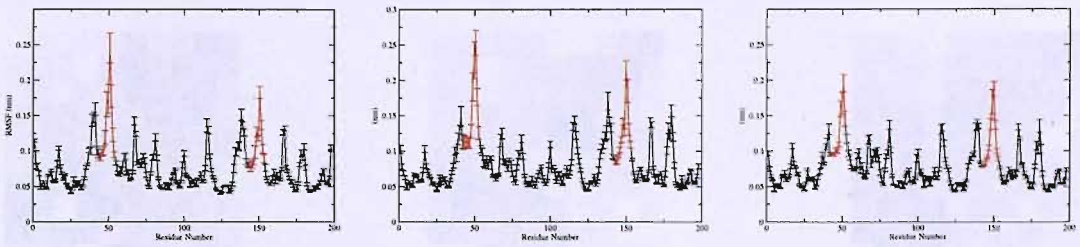


(a) MD1 (0-10 ns): WT (b) MD1 (10-20 ns):WT (c) MD2 (0-10 ns):WT

(d) MD1 (0-10 ns): M46I (e) MD1 (10-20 ns): M46I (f) MD2 (0-10 ns): M46I

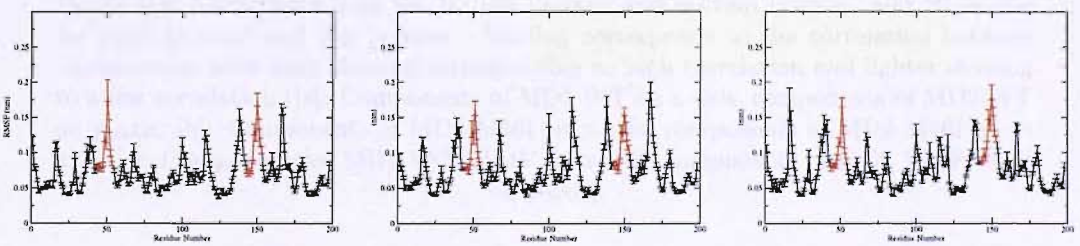
(g) MD1 (0-10 ns): V82F/I84V (h) MD1 (10-20 ns): V82F/I84V (i) MD2 (0-10 ns): V82F/I84V

FIGURE C.1: RMSF of α-carbon atoms of all residues of SQV-bound HIV-1 PR calculated for the two MD simulations of the WT and mutant apo-systems. The flap residues are marked in red.



(a) RDFMD: WT (b) MD:M46I (c) RDFMD: V82F/I84V

FIGURE C.2: RMSF of α-carbon atoms of residues of IDV-bound HIV-1 PR calculated for WT and mutant RDFMD simulations, using the equilibrated structure as the reference. RMSF values calculated as an average for all simulations, with the standard error calculated from each individual simulation marked. The flap residues are marked in red.

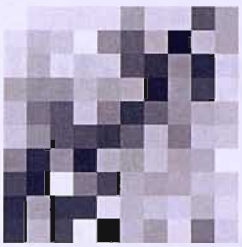


(a) RDFMD: WT (b) MD:M46I (c) RDFMD: V82F/I84V

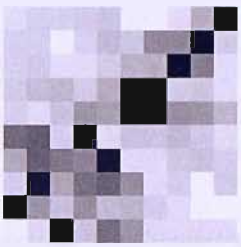
FIGURE C.3: RMSF of α-carbon atoms of residues of SQV-bound HIV-1 PR calculated for WT and mutant RDFMD simulations, using the equilibrated structure as the reference. RMSF values calculated as an average for all simulations, with the standard error calculated from each individual simulation marked. The flap residues are marked in red.

PCA Analysis

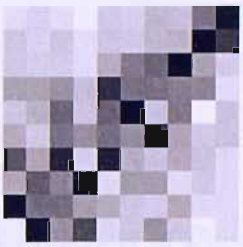
Inner-product Analysis



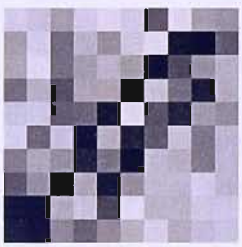
(a) MD1 and MD2: WT



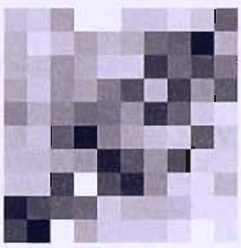
(b) MD1 and MD2: M46I



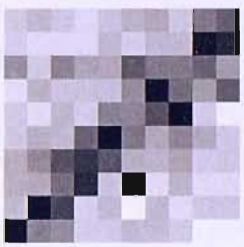
(c) MD1 and MD2: V82F/I84V



(d) MD1 and MD2: WT



(e) MD1 and MD2: M46I



(f) MD1 and MD2: V82F/I84V

FIGURE C.4: Comparison of inner products of the first ten eigenvectors of apo WT and mutant MD simulations of flap 2. Eigenvectors are plotted sequentially along the x- and y-axis, with 0 on the far left (x-axis) and bottom (y-axis), and 10 on the far right (x-axis) and top (y-axis). Shading corresponds to the correlation between eigenvectors, with dark shading corresponding to high correlation and lighter shading to a low correlation ([a]: Components of MD1 WT on x-axis, components of MD2 WT on y-axis, [b]: Components of MD1 M46I on x-axis, components of MD2 M46I on y-axis, [c]: Components of MD1 V82F/I84V on x-axis, components of MD2 V82F/I84V on y-axis)

Bibliography

[1] L. A. Eriksson, editor. *Theoretical Biochemistry-Processes and Properties of Biological Systems.* Elsevier, first edition, 2001.

[2] J. Janin and S. J. Wodak. *Prog. Biophys. Mol. Biol.*, 42:21–78, 1983.

[3] R. H. Swendsen and J. S. Wang. *Phys. Rev. Lett.*, 57:2607–2609, 1986.

[4] S. C. Phillips, M. T. Swain, A. P. Wiley, J. W. Essex, and C. M. Edge. *J. Phys. Chem. B*, 107:2098–2110, 2003.

[5] S. C. Phillips, J. W. Essex, and C. M. Edge. *J. Chem. Phys.*, 112:2586–2597, 2000.

[6] A. P. Wiley, M. T. Swain, S. C. Phillips, and J. W. Essex. *J. Chem. Theory Comput.*, 1:24–35, 2005.

[7] A. P. Wiley. *The Computational Investigation of Conformational Change.* PhD thesis, University of Southampton, 2004.

[8] H. M. Berman, Z. Westbrook, J.; Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. *Nucleic Acids Res.*, 28:235–242, 2000.

[9] G. Wagner. *J. Biomol. NMR*, 3:375–385, 1993.

[10] J. R. Lumb. *Mol. Cell Biochem.*, 73:91–98, 1987.

[11] H. Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans, and W. Yang. *Proteins*, 44:484–489, 2001.

[12] M. P. Jacobson, R. A. Friesner, Z. Xiang, and B. Honig. *J. Mol. Biol.*, 320:597–608, 2002.

[13] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiang, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, et al. *J. Comput. Chem.*, 24:1999–2012, 2003.

[14] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swiminathan, and M. Karplus. *J. Comput. Chem.*, 4:187–217, 1983.

[15] A. D. MacKerell, D. Bashford, M. Bellot, and M. Karplus *et al. J. Phys. Chem. B.*, 102:3586–3616, 1998.

[16] A. D. MacKerell, D. Bashford, D. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joesph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[17] J. Hermans, H. J. C. Berendsen, W. F. van Gunsteren, and J. P. M. Postma. *Biopolymers*, 23:1513–1518, 1984.

[18] K. H. Ott and B. Meyer. *J. Comp. Chem.*, 17:1068–1084, 1996.

[19] W. F. van Gunsteren, X. Daura, and A. E. Mark. *The Encyclopedia of Computational Chemistry*, volume 2. Wiley & Sons, 1998. 1211-1216.

[20] D. M. York, T. A. Darden, L. G. Pedersen, and M. W. Anderson. *Biochemistry*, 32:1443–1453, 1993.

[21] Jr. MacKerell, A.D., B. Brooks, III Brooks, C. L., L. Nilsson, B. Roux, Y. Won, and M. Karplus. *The Encyclopedia of Computational Chemistry.* John Wiley Sons, first edition.

[22] A. D. MacKerrell. Website, date accessed: Mar 2007. `http://www.wiley.com/legacy/wileychi/ecc/samples/sample07.pdf`.

[23] W. L. Jorgensen and J. Tirado-Rives. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[24] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[25] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard. *J. Phys. Chem. A*, 105:9396–9409, 2001.

[26] A. C. T. van Duin, A. Strachan, S. Stewman, Q. Zhang, X. Xu, and W. A. Goddard. *J. Phys. Chem. A*, 107:3803–3811, 2003.

[27] J. Wang, P. Cieplak, and P. A Kollman. *J. Comput. Chem.*, 21:1049–1074, 2000.

[28] W. Hastings. *Biometrika*, 57:97–109, 1970.

[29] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and E. Teller. *J. Chem. Phys.*, 21:1087–1092, 1953.

[30] A. R. Leach. *Molecular Modelling; Principles and Applications.* Addison Wesley Publishing Company, 2nd edition, 1997. 414-415.

[31] L. Verlet. *Phys. Rev.*, 159:98–103, 1967.

[32] R. W. Hockney. *Methods Comp. Phys.*, 9:136–211, 1970.

[33] D. Beeman. *J. Comp. Phys.*, 20:130–139, 1976.

[34] W. C. Swope, H. C. Anderson, P. H. Berens, and J. Wilson. *J. Chem. Phys.*, 76:637–649, 1982.

[35] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. *J. Comp. Phys.*, 23:327–341, 1977.

[36] D. Quigley. Website, date accessed: Apr 2007. `www.cmt.york.ac.uk/cmd/dq_talks/adiabatic.ps.gz`.

[37] H. C. Anderson. *J. Chem. Phys.*, 72:2384–2393, 1980.

[38] S. Nose. *Mol. Phys.*, 1984.

[39] W. G. Hoover. *Phys. Rev. A*, 31:1695–1697, 1985.

[40] S. Nose. *J. Chem. Phys.*, 81:511–519, 1984.

[41] G. S. Grest and K. Kremer. *Phys. Rev. A*, 33:3628–3631, 1986.

[42] S. A. Adelman and J. D. Doll. *J. Chem. Phys.*, 64:2375–2388, 1976.

[43] H. J. C. Berendsen, J. P. M. Postma, A. DiNola, and J. R. Haak. *J. Chem. Phys.*, 81:3684–3690, 1984.

[44] W. L. Jorgensen, J. Chandrasekhar, D. Madura, R. W. Impey, and M. L. Klein. *J. Chem. Phys.*, 79:925–935, 1983.

[45] H. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. *Intermolecular Forces.* 1981. 331-342.

[46] J. L. F. Abascal and C. Vega. *J. Chem. Phys.*, 123:234505, 2005.

[47] S. W. A. Rick. *J. Chem. Phys.*, 120:6085–6093, 2004.

[48] S. Alam, J. S. Vetter, P. K. Agarwal, and A. Geist. *ACM SIGPLAN Symposium on Principles and Practise of Parallel Programming*, 2006.

[49] K. A. Sharp and B. Honig. *J. Phys. Chem.*, 94:7684–7692, 1990.

[50] J. Tomasi and M. Persico. *Chem. Rev.*, 94:2027–2094, 1994.

[51] C. J. Cramer and D. G. Truhlar. *Chem. Rev.*, 99:2161–2200, 1999.

[52] W. Still, A. Tempczyk, R. Hawley, and T. Hendrickson. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[53] M. Born. *Z. Phys.*, 45, 1920.

[54] D. Bashford and D. Case. *Annu. Rev. Phys. Chem.*, 51:129–152, 2000.

[55] T. E. Cheatham and P. A. Kollman. *Ann. Rev. Phys. Chem.*, 51:435–471, 2000.

[56] V. Makarov, B. M. Pettit, and M. Feig. *Acc. of Chem. Res.*, 35:376–384, 2002.

[57] H. F. Kaiser. *Educational and Psychological Measurement*, 20:141–151, 1960.

[58] R. B. Cattell. *Multivariate Behavioral Research*, 1:629–637, 1966.

[59] A. Merlino, L. Vitagliano, M. A. Ceruso, and L. Mazzarella. *Proteins*, 53:101–110, 2003.

[60] M. A. Ceruso, A. Amadei, and A. Di Nola. *Protein Sci.*, 8:147–160, 1999.

[61] D. M. van Aalten, D. A. Conn, B. L. de Groot, H. J. Berendsen, J. B. Findlay, and A. Amadei. *Biophys. J.*, 73:2891–2896, 1997.

[62] B. L. de Groot, S. Hayward, D. M. van Aalten, A. Amadei, and H. J. Berendsen. *Proteins*, 31:116–127, 1998.

[63] E. Lindahl, B. Hess, and D. van der Spoel. *J. Mol. Mod.*, 7:306–317, 2001.

[64] W. Humpfrey, A. Dalke, and K. Schulten. *J. Mol. Graphics*, 14:33–38, 1996.

[65] M. K. Campell. *Biochemistry.* Harcourt Brace College Publishers, third edition, 1999.

[66] G. Rhodes. *Crystallography Made Crystal Clear.* Academic Press, second edition, 2006.

[67] J. K. M. Sanders and B. K. Hunter. *Modern NMR Spectroscopy.* Oxford University Press, second edition, 1993.

[68] J. J. Bozzola and L. D. Russell. *Electron Microscopy.* Southern Illinois University, second edition, 1999.

[69] R. S. Rowlett. *Protein X-ray Crystallography Methods.* Dept. of Chemistry, Colgate University, second edition, 2005.

[70] A. Bax, M. Ikura, L. E. Kay, D. A. Torchia, and R. J. Tschudin. *J. Magn. Reson.*, 40:6956–6963, 1990.

[71] T. J. Norwood, J. Boyd, J. E. Heritage, N. Soffe, and I. D. Campbell. *J. Magn. Reson.*, 87:488–501, 1990.

[72] P. Guntert. *Methods Mol. Biol.*, 278:353–378, 2004.

[73] K. Wuthrich. *J. Biol. Chem.*, 25:22059–22062, 1990.

[74] R. Ishima and D. A. Torchia. *Nat. Struct. Biol.*, 7:740–743, 2000.

[75] K. Wurthrich. Website, date accessed:Apr 2007. `http://nobelprize.org/nobel_prizes/chemistry/laureates/2002/wutrich-lecture.pdf`.

[76] A. Lewis, I. Rousso, E. Khachatryan, I. Brodsky, K. Lieberman, and Sheeves. M. *Biophys. J.*, 70:2380–2384, 1996.

[77] J. S. Salafsky. *J. Chem. Phys.*, 125:074701, 2006.

[78] S. Xie and J. D. Simon. *Proc. SPIE*, 1204:66–77, 1990.

[79] N. Go, T. Noguti, and T. Nishikawa. *Proc. Natl Acad. Sci. USA*, 80:3696–3700, 1983.

[80] B. Brooks and M. Karplus. *Proc. Natl. Acad. Sci. USA*, 80:6571–6575, 1983.

[81] K. Hinsen. Website, date accessed:Mar 07. `http://www.pasteur.fr/recherche/unites/Binfs/EMBO2004/coursenotes/hinsen_normal_modes.pdf`.

[82] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. *Science*, 220:671–680, 1983.

[83] L. I. Smith. Website, date accessed:Mar 07. `http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf`.

[84] B. L. de Groot, D. M. F. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. C. Berendsen. *Proteins*, 29:240–251, 1997.

[85] T. Schlick, E. Barth, and M. Mandziuk. *Annu. Rev. Biophys. Biomol. Struct.*, 26:181–222, 1997.

[86] T. Schlick. *J. Comput. Phys.*, 151:1–8, 1999.

[87] T. Schlick, R. D. Skeel, A. T. Brunger, L. V. Kalé, and J. *et al.* Board. *J. Comput. Phys.*, 151:9–48, 1999.

[88] H. J. Berendsen and S. Hayward. *Curr. Opin. Struct. Biol.*, 10:165–169, 2000.

[89] P. G. Debrunner and H. Frauenfelder. *Ann. Rev. Phys. Chem.*, 33:283–303, 1982.

[90] V. Hornak, A. Okur, R. Rizzo, and C. Simmerling. *Proc. Nat. Acad. Sci. USA*, 103:915–920, 2006.

[91] B. Isralewitz, M. Gao, and K. Schulten. *Curr. Opin. Struct. Biol.*, 11:224, 2001.

[92] B. Isralewitz, J. Baudry, J. Gullingsrud, D. Kosztin, and K. Schulten. *J. Mol. Graphics*, 19:13–25, 2001.

[93] B. Lu, C. F. Wong, and J. A. McCammon. *Protein. Sci.*, 14:159–168, 2005.

[94] P. Kruger, S. Verheyden, P. J. Declerck, and Y. Engelborghs. *Protein Sci.*, 10:798–808, 2001.

[95] J. Zhang, L. Chengjing, C. Kaixian, Z. Weiliang, X. Shen, and J. Hualiang. *Proc Natl. Acad. Sci. USA*, 103:13368–13373, 2006.

[96] H. Jónnson, G. Mills, K. W. Jacobson, B. J. Berne, G. Ciccotti, and D. F. Coker. *Classical and Quantum Dynamics in Condensed Phase Simulations.* World Scientific, Singapore, 1998.

[97] G. Henkelman, M. LaBute, C. Tung, P. W. Fenimore, and B. McMahon. *Proc. Natl. Acad. Sci. USA*, 102:15347–15351, 2005.

[98] D. Hamelberg, J. Mongan, and J. A. McCammon. *J. Chem. Phys.*, 120:11919–11929, 2004.

[99] H. Grubmuller. *Phys. Rev. E*, 52:2893, 1995.

[100] B. G. Schulze, H. Grabmuller, and J. D. Evanseck. *J. Am. Chem. Soc.*, 122:8700–8711, 2000.

[101] A. F. Voter. *Phys. Rev. Lett.*, 78:3908, 1997.

[102] A. F. Voter. *J. Chem. Phys.*, 106:4665–4677, 1997.

[103] J. A. Rahman and J. C. Tully. *Chem. Phys.*, 285:277–290, 2002.

[104] R. K. Hart, R. V. Pappu, and J. W. Ponder. *J. Comput. Chem.*, 20:531–545, 2000.

[105] R. E. Bruccoleri and M. Karplus. *Biopolymers*, 29:1847–1862, 1990.

[106] M. E. Tuckerman, B. J. Berne, and G. J.. Martyna. *J. Chem. Phys.*, 97:1990–2001, 1992.

[107] M. Watanabe and M. Karplus. *J. Phys. Chem*, 99:5680–5697, 1995.

[108] R. Czerminski and R. Elber. *Proteins*, 10:70–80, 1991.

[109] A. Ulitsky and R. Elber. *J. Phys. Chem.*, 98:1034–1043, 1994.

[110] C. Simmerling, J. L. Miller, and P. A. Kollman. *J. Am. Chem. Soc.*, 120:7149, 1998.

[111] M. C. Lee, J. Deng, J. M. Briggs, and Y. Duan. *Biophys. J.*, 88:3133–3146, 2005.

[112] X. W. Wu and S. M. Wang. *J. Phys. Chem. B*, 102:7238, 1998.

[113] X. Wu and B. Brooks. *Biophys. J.*, 86:19461958, 2004.

[114] I. Andricioaei and J. E. Straub. *J. Chem. Phys.*, 107:9117–9124, 1997.

[115] R. B. Sessions, P. Dauber-Osguthorpe, and D. J. Osguthorpe. *J. Mol. Biol.*, 210:617–633, 1989.

[116] V. Tozzini, J. Trylska, C. E. Chang, and J. A. McCammon. *J. Struct. Biol.*, 157:606–615, 2007.

[117] H. M. Chun, C. E. Padilla, D. N. Chin, M. Watanabe, V. I. Karlov, H. E. Alper, K. Soosaar, K. B. Blair, O. M. Becker, L. S. D. Caves, R. Nagle, D. N. Haney, and B. L. Farmer. *J. Comput. Chem.*, 21:159–184, 2000.

[118] U. H. E. Hansmann. *Annual Reports ZiF: 96/97, Center for Interdisciplinary Research (Zif) of the Universitt Bielefeld (Germany)*, page 46, 1998.

[119] A. M. Ferrenberg and R. H. Swendsen. *Phys. Rev. Lett.*, 61:2635–2638, 1998.

[120] B. A. Berg and T. Neuhaus. *Phys. Lett.*, 267:249–253, 1991.

[121] B. A. Berg and T. Neuhaus. *Phys. Rev. Lett.*, 68:9–12, 1992.

[122] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. *J. Chem. Phys.*, 96:1776–1783, 1992.

[123] K. Hukushima and K. Nemoto. *J. Phys. Soc. Jpn.*, 65:1604–1608, 1996.

[124] K. Hukushima, H. Takayama, and K. Nemoto. *Int. J. Mod. Phys. C*, 7:337–334, 1996.

[125] Y. Sugita and Y. Okamoto. *Chem. Phys. Lett.*, 329:261–270, 2000.

[126] A. Mitsutake, Y. Sugita, and Y. Okamoto. *J. Chem. Phys.*, 118:6676–6688, 2003.

[127] A. Mitsutake and Y. Okamoto. *J. Chem. Phys.*, 121:2491–2504, 2004.

[128] Y. M. Rhee and V. S. Pande. *Biophys. J.*, 84:775–786, 2003.

[129] N. E. Huang, S. R. Long, and Z. Shen. *Adv. Appl. Mech.*, 32:59–111, 1996.

[130] N. E. Huang *et al. Proc. R. Soc. London, Ser. A*, 454:903–995, 1998.

[131] N. E. Huang, S. R. Long, and Z. Shen. *Annu. Rev. Fluid Mech.*, 31:417–457, 1999.

[132] N. E. Huang, Z. Shen, S. R. Long, M. L. C. Wu, H. H. Shih, Q. N. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu. *Proc. R. Soc. London, Ser. A*, 454:903, 1998.

[133] A. P. Wiley, S. L. Williams, and J. W. Essex. Conformational motions of hiv-1 protease identified using reversible digitally filtered molecular dynamics. submitted to *J. Chem. Theory. Comput.*, 2007.

[134] *MATLAB 6.1.0.* The Mathworks Inc., Natick, MA, 2001.

[135] Y. Sugita and Y. Okamoto. *Chem. Phys. Lett.*, 314:141–151, 1999.

[136] D. A. Kofke. *J. Chem. Phys.*, 117:436–445, 2002.

[137] C. Predescu, M. Predescu, and C. V. Ciobanu. *J. Phys. Chem. B*, 109:4189, 2005.

[138] N. Rathore, M. Chopra, and J. J. de Pablo. *J. Chem. Phys.*, 122:024111, 2005.

[139] H. G. Katzgraber, S. Trebst, D. A. Hose, and M. Troyer. *J. Stat. Mech.*, 06:3018, 2006.

[140] D. A. Kofke. *J. Chem. Phys.*, 117:6911–6914, 2002.

[141] D. A. Kofke. *J. Chem. Phys.*, 120:10852, 2004.

[142] K. J. Sanbonmatsu and A. E. Garcia. *Proteins*, 46:225, 2002.

[143] A. Kone and D. A. Kofke. *J. Chem. Phys.*, 122:206101, 2005.

[144] U. H. E. Hansmann and Y. Okamoto. *J. Comput. Chem.*, 18:920, 1997.

[145] U. H. E. Hansmann. *Eur. Phys. J. B*, 12:607–611, 1999.

[146] C. Y. Lin, C. K. Hu, and U. H. E. Hansmann. *Proteins*, 53:436–445, 2003.

[147] U. Reimer, G. Scherer, M. Drewello, S. Kruber, M. Schutkowski, and G. Fischer. *J. Mol. Biol.*, 279:449–460, 1998.

[148] M. Schutkowski, A. Bernhardt, X. Z. Zhou, M. H. Shen, U. Reimer, J. U. Rahfeld, K. P. Lu, and G. Fischer. *Biochemistry*, 37:5566–5575, 1998.

[149] G. Fischer, R. L. Dunbrack, and M. Karplus. *J. Am. Chem. Soc.*, 116:11931–11937, 1994.

[150] J. M. Schmidt, R. Briischweiler, R. R. Emst, R. L. Dunbrack, D. Joseph, and M. Karplus. *J. Am. Chem. Soc.*, 115:8747–8756, 1993.

[151] I. Z. Sienion, M. Cebrat, and Z. Wieczorek. *Arch. Immunol. Ther. Exp. (Warsz)*, 47:143–153, 1999.

[152] H. Kessler, G. Gratias, M. Hessler, M. Gurrath, and M. Muller. *Pure and Appl. Chem.*, 68:1201–1205, 1996.

[153] G. V. Nikiforovich, K. E Kover, W. Zhang, and G. R. Marshall. *J. Am. Chem. Soc.*, 122:3262–3273, 2000.

[154] A. Loffet. *J. Pept. Sci.*, 8:1075–2617, 2002.

[155] G. Hummel, U. Reineke, and U. Reimer. *Mol. BioSyst.*, 2:499–508, 2006.

[156] D. Pahlke, C. Freund, D. Leitner, and D. Labudde. *BMC Structural Biology*, 5:8, 2005.

[157] P. E. Shaw. *EMBO Reports*, 3:521–526, 2002.

[158] F. X. Schmid. *Annu. Rev. Biophys. Struct.*, 22:123–143, 1993.

[159] K. A. Higgins, D. J. Craik, J. G. Hall, and P. R. Andrews. *Drug Des. Deliv.*, 3:159–170, 1988.

[160] H. Weisshoff, T. Wieprecht, P. Henklein, C. Frommel, C. Antz, and C. Mugge. *R Fed. Bioc. Soc. Lett.*, 387:201, 1996.

[161] R. L. Stein. *Adv. Quantum Chem.*, 11:1, 1993.

[162] F. X. Schmid, L. M. Mayr, M. Mueche, and E. R. Schoenbrunner. *Adv. Quantum Chem.*, 11:25, 1993.

[163] C. Grathwohl and K. Wuthrich. *Biopolymers*, 20:2623–2633, 1981.

[164] H. Dyson, M. Rance, R. Houghten, P. Wright, and R. Lerner. *J. Mol. Biol.*, 201:201–217, 1988.

[165] G. Muller, M. Gurrath, M. Kurz, and M. Kessler. *Proteins*, 15:235–251, 1993.

[166] G. Zanotti, M. Saviano, G. Saviano, T. Tancredi, C. Rossi, C. Pedone, and E. Benedetti. *J. Peptide Res.*, 51:460–466, 1998.

[167] D. Hamelberg, D. Shen, and A. McCammon. *J. Chem. Phys.*, 122:241103, 2005.

[168] G. M. Wu and M. W Deem. *J. Chem. Phys.*, 111:6625–6632, 1999.

[169] R. N. Riemann and M. Zacharias. *J. Peptide Res.*, 63:354–364, 2004.

[170] L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J.; Phillips, A. Shinozaki, K. Varadarajan, and K. schulten. *J. Comput. Phys.*, 151:283–312, 1999.

[171] G. Vriend. *J. Mol. Graphics*, 8:52–56, 1990.

[172] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, G Ferguson, D.and Seibel, and P. Kollman. *Comput. Phys. Commun*, 91:1–41, 1995.

[173] T. Fox and P. A. Kollman. *J. Phys. Chem. B*, 102:8070–8079, 1998.

[174] X. Grabuleda, C. Jaime, and P. A. Kollman. *J. Comput. Chem.*, 21:901, 2000.

[175] T. A. Darden, D. M. York, and L. G. Pedersen. *J. Chem. Phys.*, 98:10089, 1993.

[176] R. Faller, Q. Yan, and J. de Pablo. *J. Chem. Phys.*, 116:5419–5423, 2002.

[177] J. D. Reeves and R. W. Doms. *J. Gen. Virol.*, 83:1253–1265, 2002.

[178] Website, date accessed:Jan 2007. `http://www.wiley.com/legacy/college/boyer/0470003790/cutting_edge/aids_therapies/aids_therapies.htm`.

[179] E. A. Berger, P. M. Murphy, and J. M. Farber. *Annu. Rev. Immunol.*, 17:657–700, 1999.

[180] N. E. Kohl, E. A. Emini, W. A. Schleif, L. J. Davis, J. C. Heimbach, R. A. Dixon, E. M. Scholnick, and I. S. Sigal. *Proc. Natl. Acad. Sci. USA*, 85:4686–4690, 1988.

[181] C. Peng, B. K. Ho, T. W. Chang, and N. T. Chang. *J. Virol.*, 63:2550–2556, 1989.

[182] H. Gottlinger, J. Sodroski, and W. A. Haseltine. *Proc. Natl. Acad. Sci. USA*, 86:5781–5785, 1989.

[183] A. Wlodawer and J. W. Erickson. *Annu. Rev. Biochem.*, 62:543–585, 1993.

[184] P. Cigler, M. Koziesk, P. Rezacova, J. Brynda, Z. Otwinowski, J. Pokorna, J. Plesek, B. Gruner, L. Doleckova-Maresova, M. Masa, J. Sedlacek, J. Bodem, H. G. Kraeusslich, V. Kral, and J. Konvalinka. *Proc. Nat. Acad. Sci. USA*, 102:15394–15399, 2005.

[185] S. Spinelli, Q. Z. Liu, P.M. Alzari, P.H. Hirel, and R.J. Poljak. *Biochimie*, 73:1391–1396, 1991.

[186] B. Pillai, K.K. Kannan, and M. V. Hosur. *Proteins*, 43:57–64, 2001.

[187] R. Lapatto *et al.* and P. M. Hobart. *Science*, 245:616–621, 1989.

[188] M. A. Navia, P. M. D. Fitzgerald, B. M. McKeever, C. H. Leu, C. Heinbach, I. S. Sigal, P. L. Darke, and J. P. Springer. *Nature*, 337:615–620, 1989.

[189] A. Wlodawer, M. Miller, M. Jaskolski, B. Sathyanarayana, E. Baldwin, I. Weber, L. Selk, L. Clawsen, J. Schneider, and S. Kent. *Science*, 245:616–621, 1989.

[190] P. Martin, J. F. Vickrey, G. Proteasa, Y. L. Jimenez, Z. Wawrzak, M. A. Winters, T. C. Merigan, and L. C. Kovari. *Structure*, 13:1887–1895, 2005.

[191] S. W. Rick, J. W. Erickson, and S. K Burt. *Proteins*, 32:7–16, 1998.

[192] R. Ishima, D. I. Freedberg, Y. X. Wang, J. M. Louis, and D. A. Torchia. *Structure*, 7:1047–1055, 1999.

[193] J. E. Foulkes, M. Prabu-Jeyabalan, D. Cooper, G. J. Henderson, J. Harris, R. Swanstrom, and C. A. Schiffer. *J. Virol.*, 80:69066916, 2006.

[194] Website, Date acessed: Jan 2007. `http://xpdb.nist.gov/hivsdb/hivsdb.html`.

[195] L. J. Hyland, T. A. J. Tomaszek, and T. D. Meek. *Biochemistry*, 30:8454–8463, 1991.

[196] E. Rodriguez, C. Deckman, H. Abu-Soud, F Raushel, and F. Meek. *Biochemistry*, 32:7886–78911, 1993.

[197] D. C. Chatfields and B. R. Brooks. *J. Am. Chem. Soc.*, 117:5561–5572, 1995.

[198] H. Lee, T. A. Darden, and L. G. Pederson. *J. Am. Chem. Soc.*, 118:3946–3950, 1996.

[199] A. M. Silva, R. E. Cachau, H. L. Sham, and J. W. Erickson. *J. Mol. Biol.*, 255:321–346, 1996.

[200] B. R. Chatfield, D. C.and Eurenius K. P.and Brooks. *J. Mol. Struc. (THEOCHEM)*, 423:79–92, 1998.

[201] H. Park, J. Suh, and S. Lee. *J. Am. Chem. Soc.*, 122:3901–3908, 2000.

[202] K. Suguna, E. A. Padlan, C. W. Smith, W. D. Carlson, and D. R. Davies. *Proc. Natl. Acad. Sci. USA*, 84:6612, 1987.

[203] M. N. G. James and A. R. Sielecki. *Biochemistry*, 24:3701–3713, 1985.

[204] T. Hofmann, R. S. Hodges, and M. N. G. James. *Biochemistry*, 23:635–643, 1984.

[205] K. P. Eurenius, D. Chatfield, B. R. Brooks, and M. Hodoscek. *Int. J. Quant. Chem.*, 60:1189–1200, 1996.

[206] W. R. P. Scott and C. A. Schiffer. *Struct. Fold. Des.*, 8:1259–1265, 2000.

[207] E. Katoh, J. M. Louis, T. Yamazaki, A. M. Gronenborn, and D. A. Torchia. *Protein Sci.*, 12:1376–1385, 2003.

[208] E. Furfine, E. Dsouza, K. Ingold, J. Leban, T Spector, and D. Porter. *J. Am. Chem. Soc.*, 127:13778–13779, 1992.

[209] B. Maschera, G. Palfi, and L. *et al* Wright. *J. Biol. Chem.*, 271:33231–33235, 1996.

[210] S. Piana, P. Carloni, and M. Parinello. *J. Mol. Biol.*, 319:567–583, 2002.

[211] G. Toth and A. Borics. *J. Mol. Grap. Modell.*, 24:465–474, 2006.

[212] R. Bhattacharyya and P. Chakrabarti. *J. Mol. Biol.*, 331:925–940, 2003.

[213] A. Zarrinpar and W. Lim. *Nat. Struct. Biol.*, 331:611–613, 2000.

[214] W. Shao, L. Everitt, M. Manchester, D. D. Loeb, C. A. Hutchison, and R. Swanstrom. *Proc. Natl. Acad. Sci. USA*, 94:2243–2248, 1997.

[215] S. H. Gellman. *Biochemistry*, 30:6633–6636, 1991.

[216] D. I. Freedberg, R. Ishima, J. Jacob, Y. Wang, I. Kustanovich, J. M. Louis, and D. A. Torchia. *Protein Sci.*, 11:221–232, 2002.

[217] Harte and Beveridge. *Proc. Natl. Acad. Sci. USA*, 87:8864–8868, 1990.

[218] J. R. Collins, S. K. Burt, and J. W. Erickson. *Struct. Biol.*, 2:334–338, 1995.

[219] W. R. P. Scott *et al.* and van Gunsteren W. F. *J. Phys. Chem. A.*, 103:3596–3607, 1999.

[220] W. F. van Gunsteren *et al.* and I. G. Tironi. *The GROMOS96 manual and user guide.* 1996.

[221] A. L. Perryman, J. H. Lin, and J. A. Mccammon. *Protein Sci.*, 13:1108–1123, 2004.

[222] K. L. Meagher and H. A. Carlson. *Proteins*, 58:119–125, 2005.

[223] L. David, R. Luo, and M. K. Gilson. *J. Comput. Chem.*, 21:295–309, 1999.

[224] M. K. Gilson and B. Honig. *Proteins*, 4:7–18, 1988.

[225] R. J. Zauhar and R. S. Morgan. *J. Comput. Chem.*, 11:603–622, 1985.

[226] M. E. Davies, J. D. Madura, B. A. Luty, and J. A. McCammon. *Comput. Phys. Commun.*, 62:187, 1991.

[227] T. J. You and S. C. Harvey. *J. Comput. Chem.*, 14:484–501, 1993.

[228] B. J. Yoon and A. M. Lenhoff. *J. Comput. Chem*, 11:1080–1086, 1990.

[229] D. Hamelberg and A. McCammon. *J. Am. Chem. Soc.*, 127:13778–13779, 2005.

[230] W. Shao, Everitt, L., M. Manchester, D. Loeb, C. Hutchinson, and R. Swanstrom. *Proc. Natl. Acad. Sci. USA*, 123:6227–6231, 2001.

[231] A. P. Wiley, M. T. Swain, S. C. Phillips, and J. W. Edge, C. M.and Essex. *J. Chem. Theory Comput.*, 1:24–35, 2005.

[232] I. Luque, M. J. Todd, J. Gomez, N. Semo, and E. Freire. *Biochemistry*, 37:5791–5797, 1998.

[233] S. Vega, L. Kang, A. Valazquez-Campoy, Y. Kiso, L. M. Amzel, and E. Freire. *Proteins*, 55:594–602, 2004.

[234] A. L. Perryman, J. Lin, and J. A. McCammon. *Biopolymers*, 82:272–284, 2006.

[235] M. J. Bowmann, S. Bryne, and J. Chmielewski. *Chem. & Biol.*, 12:439–444, 2005.

[236] N. Boggetto and M. Reboud-Ravaux. *Biol. Chem.*, 383:1321–1324, 2002.

[237] J. H. Condra, W. A. Schlief, O. M. Blaky, L. J. Gadryelski, D. J. Graham, J. C. Quintero, A. Rhodes, H. L. Robbins, E. Roth, M. Shivaprakash, D. Titus, T. Yang, H. Teppler, K. E. Squires, P. J. Deutsch, and E. A. Emini. *Nature*, 374:569–571, 1995.

[238] R. N. Husson, T. Shirasaka, K. M. Butler, P. A. Pizzo, and H. Mitsuya. *J. Pediatr.*, 123:9–16, 1993.

[239] D. D. Rickman. *Textbook of AIDS Medicines.* 1994.

[240] Website, Date acessed: Apr 2007. `http://www.fda.gov`.

[241] K. Hertogs, S. Bloor, S. D. Kemp, C. Van den Eynde, T. M. Alcorn, R. Pauwels, M. Van Houtte, S. Staszewski, V. Miller, and B. A. Larder. *AIDS*, 14:1203–1210, 2000.

[242] T. D. Wu, C. A. Schiffer, M. J. Gonzales, J. Taylor, R. Kantor, S. Chou, D. Israelski, A. R. Zolopa, W. J. Fessel, and R. W. Shafer. *J. Virol.*, 77:4836–4847, 2003.

[243] H. Gatanaga, D. Das, Y. Suzuki, D. D. Yeh, K. A Hussain, A. K. Ghosh, and H. Mitsuya. *J. Biol. Chem.*, 281:1241–1250, 2005.

[244] L. Doyon, G. Croteau, D. Thibeault, F. Puolin, L. Pilote, and D. Lamarre. *J. Virol*, 70:3763–3769, 1996.

[245] H. Gatanaga, Y. Suzuki, H. Tsang, K. Yoshimura, M. F. Kavlick, K. Nagashima, R. J. Gorelick, S. Mardy, C. Tang, M. F. Summers, and H. Mitsuya. *J. Biol. Chem.*, 277:5952–5961, 2002.

[246] Y. M. Zhang, H. Imamichi, T. Imamichi, H. C. Lane, J. Falloon, M. B. Vasudevachari, and N. P. Salzman. *J. Virol.*, 71:6662–6670, 1997.

[247] S. Tamiya, S. Mardy, M. F. Kavlick, K. Yoshimura, and H. Mistuya. *J. Virol*, 78:12030–12040, 2004.

[248] S. Gulnik, J. W. Erickson, and D. Xie. *Vitam. Horm*, 58:213–256, 2000.

[249] A. Warshel. *Proc. Natl. Acad. Sci. USA*, 75:5250–5254, 1978.

[250] P. D. Yin, D. Das, and H. Mitsuya. *Cell. Mol. Life Sci.*, 63:1706–1724, 2006.

[251] N. Ozer, T. Haliloglu, and C. Schiffer. *Proteins*, 64:444–456, 2006.

[252] M Prabu-Jeyabalan, E Nalivaika, and C. Schiffer. *Structure*, 10:369–381, 2002.

[253] N. M. King, M. Prabu-Jeyabalan, E. A. Nalivaika, and C. A. Schiffer. *Chem. Biol.*, 11:1333–1338, 2004.

[254] S. Pazhanisamy, C. M. Stuver, A. B. Cullinan, N. Margolin, B. G. Rao, and D. J. Livingston. *J. Biol. Chem.*, 271:17979–17985, 1996.

[255] H. B. Schock, V. M. Garsky, and L. C. Kuo. *J. Biol. Chem.*, 271:31957–31963, 1996.

[256] Z. Chen, Y. Li, H. B. Schock, D. Hall, E. Chen, and L. C. Kuo. *J. Biol. Chem.*, 270:21433–21436, 1995.

[257] M. Markowitz, H. Mo, D. J. Kempf, D. W. Norbeck, T. N. Bhat, J. W. Erickson, and D. D. Ho. *J. Virol.*, 69:701–706, 1995.

[258] L. Schaffer and G. M. Verkhivker. *Proteins*, 33:295–310, 1998.

[259] G. J. Tawa, I. A. Topol, S. K. Burt, and J. W. Erickson. *J. Am. Chem. Soc.*, 120:8856–8863, 1998.

[260] B. Mahalingam, J. M. Louis, C. C. Reed, J. M. Adomat, J. Krouse, Y. F. Wang, R. W. Harrison, and I. T. Weber. *Eur. J. Biochem.*, 263:455–464, 1999.

[261] B. Mahalingam, J. M. Louis, Hung, R. W. J. Harrison, and I. T. Weber. *Proteins*, 43:455–464, 2001.

[262] C. D. Rosin, R. K. Belew, W. L. Walker, G. M. Morris, A. J. Olson, and D. S. Goodsell. *J. Mol. Biol.*, 287:77–92, 1999.

[263] L. Hong, X. C. Zhang, J. A. Hartsuck, and J. Tang. *Protein Sci.*, 9:1898–1904, 2000.

[264] W. Wang and P. A. Kollmann. *Proc. Natl. Acad. Sci. USA*, 98:14937–14942, 2001.

[265] M. Tisdale, R. E. Myers, B. Maschera, N. R. Parry, N. M. Oliver, and E. D. Blair. *Antimicrob. Agents Chemother.*, 39:1704–1710, 1995.

[266] T. Watkins, W. Resch, D. Irlbeck, and R. Swanstrom. *Antimicrob. Agents Chemother.*, 47:759–769, 2003.

[267] K. Hertogs, S. Bloor, S. D. Kemp, C. Van der Eynde, T. M. Alcorn, R. Pauwels, M. Van Houtte, S. Staszewski, V. Miller, and B. A. Larder. *AIDS*, 14:1203–1210, 2000.

[268] C. A. Kemper, M. D Witt, P. H. Keiser, M. P. Dube, D. N. Forthal, M. Leibowitz, D. S. Smith, A. Rigby, N. S. Hellmann, Y. S. Lie, J. Leedom, D. Richman, J. A. McCutchan, and R. Haubrich. *AIDS*, 15:609–615, 2001.

[269] M. J. Todd, I. Luque, A. Velazquez-Campoy, and E. Freire. *Biochemistry*, 39:11876–11883, 2000.

[270] J. T. Randolph and D. A DeGoey. *Curr. Top. Med. Chem.*, 4:1079–1095, 2004.

[271] S. Muzammil, A. A. Armstrong, L. W. Kang, A. Jakalian, P. R. Bonneau, V. Schmelmer, L. M. Amzel, and E. Freire. *J. Virol.*, 2007. Epub ahead of print.

[272] R. A. Chrusciel and J. W. Strohbach. *Curr. Top. Med. Chem.*, 4:1097–1114, 2004.

[273] D. Schake. *AIDS*, 18:579–580, 2004.

[274] S. Muzammil, L. W. Kang, A. A. Armstrong, A. Jakalian, P. R. Bonneau, V. Schmelmer, L. M. Amzel, and E. Freire. *Antivir. Ther.*, 10:S70, Abstract 63, 2005.

[275] L. Doyon, S. Tremblay, L. Bourgon, E. Wardrop, and M. G. Cordingley. *Antiviral. Res.*, 68:27–35, 2005.

[276] R. B. Rose, C. S. Craik, and R. M. Stroud. *Biochemistry*, 37:2607–2621, 1998.

[277] P. J. Ala, E. E. Huston, R. M. Klabe, D. D. McCabe, J. L. Duke, C. J. Rizzo, B. D. Korant, R. J. DeLoskey, P. Y. S. Lam, and C. N. Hodge. *Biochemistry*, 36:1573–1580, 1997.

[278] R. M. Klabe, L. T. Bacheler, P. J. Ala, S. Erickson-Viitanen, and J. L. Meek. *Biochemistry*, 87:8735–8742, 1998.

[279] H. Ohtaka, A. Velazquez-Campoy, D. Xie, and E. Freire. *Protein Sci.*, 11:1908–1911, 2002.

[280] R. Ishima and D. A. Torchia. *Nat. Struct. Biol.*, 7:740–743, 2000.

[281] R. Ishima, J. Jacobi, Y. Wang, I. Kustanovich, J. M. Louis, and D. A. Torchia. *Protein Sci.*, 11:221–232, 2002.

[282] T. Hou and R. Yu. *J. Med. Chem.*, 50:1177–1188, 2007.

[283] C. A. Chang, T. Shen, J. Trylska, V. Tozzini, and J. A. McCammon. *Biophys. J.*, 90:3880–3885, 2006.

[284] S. Pazhanisamy, J. A. Partaledis, P. G. Rao, and D. J. Livingston. *Adv. Exp. Med. Biol.*, 436:75–83, 1998.

[285] T. W. Ridky, A. Kikonyogo, J. Leis, S. Gulnik, T. Copeland, J. W. Erickson, A. Wlodawer, I. Kurinov, R. W. Harrison, and I. T. Weber. *Biochemistry*, 37:13835–13845, 1998.

[286] S. Piana, P. Carloni, and U. Rothlisberger. *Protein Sci.*, 11:2393–2402, 2002.

[287] C. A. Chang, J. Trylska, V. Tozzini, and J. A. McCammon. *Chem. Biol. Drug Des.*, 69:5–13, 2007.

[288] G. Toth and A. Borics. *Biochemistry*, 45:6606–6614, 2006.

[289] V. Hornak, A. Okur, R. Rizzo, and C. Simmerling. *J. Am. Chem. Soc.*, 128:2812–2813, 2006.

[290] M. Jaskolski, A. G. Tomasselli, T. K. Sawyer, R. L. Staples, R. L. Heinrikson, J. Schneider, S. B. H. Kent, and A. Wlodawer. *Biochemistry*, 30:1600–1609, 1991.

[291] J. Trylska, P. Grochowski, and J. A. McCammon. *Protein Sci.*, 13:513–528, 2004.

[292] J. Trylska, P. Bala, M. Geller, and P. Grochowski. *Biophys. J.*, 83:794–807, 2002.

[293] R. Smith, I. M. Brereton, R. Y. Chai, and S. B. H. Kent. *Nat. Struc. Biol.*, 3:946–950, 1996.

[294] S. Piana and P. Carloni. *Proteins*, 39:26 – 36, 2000.

[295] W. Wang and P. A Kollman. *J. Mol. Biol.*, 303:567–582, 2000.

[296] D. Kovalskyy, V. Dubyna, A. E. Mark, and A. Kornelyuk. *Proteins*, 58:450–458, 2005.

[297] W. E. Harte and D. L. Beveridge. *J. Am. Chem. Soc.*, 115:3883–3886, 1993.

[298] S. Piana, D. Sebastiani, P. Carloni, and M. Parrinello. *J. Am. Chem. Soc.*, 123:8730–8737, 2001.

[299] T. Yamazaki, L. K. Nicholson, D. A. Torchia, P. Wingfield, S. J. Stahl, J. D. Kaufman, C. J. Eyermann, C. N. Hodge, P. Y. S. Lam, Y. Ru, P. K. Jadhav, C. H. Chang, and P. C. Weber. *J. Am. Chem. Soc.*, 116:10791–10792, 1994.

[300] J. Brynda, P. Rezacova, M. Fabry, M. Horejsi, R. Stouracova, J. Sedlacek, M. Soucek, M. Hradilek, M. Lepsik, and J. Konvalinka. *J. Med. Chem.*, 47:2030–2036, 2004.

[301] M. Lepsik, Z. Kriz, and Z. Havlas. *Proteins*, 57:279–293, 2004.

[302] K. Wittayanarakul, O. Aruksakunwong, S. Saen-oon, W. Chantratita, V. Parasuk, P. Sompornpisut, and S. Hannongbua. *Biophys. J.*, 88:867–879, 2005.

[303] K. Wittayanarakul, O. Aruksakunwong, P. Sompornpisut, V. Sanghiran-Lee, V. Parasuk, S. Pinitglang, and S. Hannongbua. *J. Chem. Inf. Model.*, 45:300–308, 2005.

[304] A. Engelman and R. Craigie. *J. Virol.*, 66:6361–6369, 1992.

[305] A. Engelman, F. D. Bushman, and R. Craigie. *EMBO J.*, 12:3269–3275, 1993.

[306] D. C. van Gent, C. Vink, A. A. Groeneger, and R. H. Plasterk. *EMBO J.*, 12:3261–3267, 1993.

[307] S. P. Lee, J. Xiao, J. R. Knutson, M. S. Lewis, and M. K. Han. *Biochemistry*, 36:173–180, 1997.

[308] E. Deprez, P. Tauc, H. Leh, J. F. Moucadet, C. Auclair, and J. C. Brochon. *Biochemistry*, 39:9275–9284, 2000.

[309] R. Zheng, T. M. Jenkins, and R. Craigie. *Proc. Natl. Acad. Sci. USA*, 93:13659–13664, 1996.

[310] E. Kahn, J. P. G. Mack, R. A. Katz, J. Kulkosky, and A. M. Skalka. *Nucleic Acids Res.*, 19:851–860, 1991.

[311] C. Vink, A. M. Oude, Groeneger, and R. H. Plasterk. *Nucleic Acids Res.*, 21:1419–1425, 1993.

[312] A. M. Woerner and C. J. Marchis-Sekura. *Nucleic Acids Res.*, 21:3507–3511, 1993.

[313] A. Engelman, A. B. Hickman, and R. Craigie. *J. Virol.*, 68:5911–5917, 1994.

[314] R. A. Puras-Lutzke, C. Vink, and R. H. Plasterk. *Nucleic Acids Res.*, 22:4125–4131, 1994.

[315] P. O. Brown. *Retroviruses.* Cold Spring Harbour Laboratory Press, 1997. pages 161-203.

[316] J. C. H. Chen, J. Krucinski, L. J. W. Miercke, J. S. Finer-Moore, A. H. Tang, A. D. Leavitt, and R. M. Stroud. *Proc. Natl. Acad. Sci. USA*, 97:8233–8238, 2000.

[317] S. A. Chow, K. A. Vincent, V. Eliison, and P. O. Brown. *Science*, 255:723–726, 1992.

[318] F. D. Bushman, A. Engelman, I. Palmer, P. Wingfield, and R. Cragie. *Proc. Natl. Acad. Sci. USA*, 90:3428–3432, 1993.

[319] P. O. Brown, B. Bowerman, H. E. Varmus, and J. M. Bishop. *Cell.*, 49:437–356, 1987.

[320] L. I. Lobel, J. E. Murphy, and S. P. Goff. *J. Virol.*, 63:2629–2637, 1989.

[321] P. Gallay, S. Swingler, C. Aiken, and D. Trono. *Cell.*, 80:379–388, 1995.

[322] A. S. Espeseth, P. Felock, A. Wolfe, M. Witmer, J. Grobler, N. Anthony, M. Egbertson, J. Y. Melamed, and S. *et al.* Young. *Proc. Natl. Acad. Sci. USA*, 97:11244–112449, 2000.

[323] F. Dyda, A. B. Hickman, T. M. Jenkins, A. Engelman, R. Craigie, and D. R. Davies. *Science*, 266:1981–1986, 1994.

[324] G. Bujacz, M. Jaskolski, J. Alexandratos, A. Wlodawer, G. Merkel, R. A. Katz, and A. M. Skalka. *J. Mol. Biol.*, 253:333–346, 1995.

[325] G. Bujacz, M. Jaskolski, J. Alexandratos, A. Wlodawer, G. Merkel, R. A. Katz, and A. M. Skalka. *Structure*, 4:89–96, 1996.

[326] Y. Goldgur, F. Dyda, A. B. Hickman, T. M. Jenkins, R. Craigie, and D. R. Davies. *Proc. Natl. Acad. Sci. USA*, 95:9150–9154, 1998.

[327] A. P. Eijkelenboom, F. M. van den Ent, A. Vos, J. F. Doreleijers, K. Hard, T. D. Tullius, R. H. Plasterk, R. Kaptein, and R. Boelens. *Curr. Biol.*, 7:739–746, 1997.

[328] A. P. Eijkelenboom, R. A. Lutzke, R. Boelens, R. H. Plasterk, R. Kaptein, and K. Hard. *Nature Struct. Biol.*, 2:807–810, 1995.

[329] P. J. Lodi, J. A. Ernst, J. Kuszewski, A. B. Hickman, A. Engelman, R. Craigie, G. M. Clore, and A. M. Groenenborn. *Biochemistry*, 34:9826–9833, 1995.

[330] M. Cai, R. Zheng, M. Caffrey, R. Craigie, G. M. Clore, and A. M. Gronenborn. *Nat. Struct. Biol.*, 4:567–577, 1997.

[331] Z. G. Chen. *J. Mol. Biol.*, 296:521–533, 2000.

[332] J. C. H. Chen, J. Krucinski, L. J. W. Miercke, J. S. Finer-Moore, A. H. Tang, A. D. Leavitt, and R. M. Stroud. *Proc. Natl. Acad. Sci. USA*, 97:8233–8238, 2000.

[333] Z. N. Yang, Muserm T. C., F. D. Bushman, and C. C. Hyde. *J. Mol. Biol.*, 296:535–548, 2000.

[334] T. M. Jenkins, A. B. Hickman, F. Dyda, R. Ghirlando, D. R. Davies, and R. Craigie. *Proc. Natl. Acad. Sci. USA*, 92:6057–6061, 1995.

[335] P. Cherepanov, G. Maertens, P. Proost, B. Devreese, J. van Beeumen, Y. Engelborghs, De Clercq Y., and Z. Debyser. *J. Biol. Chem.*, 278:372–381, 2003.

[336] F. M. I. van den Ent, C. Vink, and R. H. Plasterk. *J. Virol.*, 68:7825–7832, 1994.

[337] G. Kukolj and A. M. Skalka. *Genes and Dev.*, 9:2256–2567, 1995.

[338] J. Wang, H. Ling, W. Yang, and R. Craigie. *EMBO J.*, 20:7333–7343, 2001.

[339] V. Ellison and P. O. Brown. *Proc. Natl. Acad. Sci. USA*, 91:7316–7320, 1994.

[340] A. L. Wolfe, P. J. Felock, J. C. Hastings, C. Uncapher Blau, and D. J. Hazuda. *J. Virol*, 70:1424–1432, 1996.

[341] C. Vink, R. A. Lutzke, and R. H. Plasterk. *Nucleic Acids Res.*, 22:4103–4110, 1994.

[342] D. J. Hazuda, P. J. Felock, J. C. Hastings, B. Pramanik, and A. L. Wolfe. *J. Virol.*, 71:7005–7011, 1997.

[343] S. Maignan, J. P. Guilloteau, Q. Zhou-Liu, C. Clement-Mella, and V. Mikol. *J. Mol. Biol.*, 282:359–368, 1998.

[344] C. Laboulais, E. Deprez, H. Leh, J. Mouscadet, J. Brochon, and M. Le Bret. *Biophys. J.*, 81:473–489, 2001.

[345] R. D. Lins, J. M. Briggs, T. P. Straatsma, H. A. Carlson, J. Greenwald, S. Choe, and J. A. McCammon. *Biophys. J.*, 76:2999–3011, 1999.

[346] A. Wijitkosoom, S. Tonmunphean, T. N Truong, and S. Hannongbua. *J. Biomol. Struc. and Dyn.*, 3:613–624, 2006.

[347] R. D. Lins, A. Adesokan, T. A. Soares, and J. M. Briggs. *Pharmacol. Ther.*, 85:123–131, 2000.

[348] E. Asante-Appiah, S. H. Seeholzer, and A. M. Skalka. *J. Biol. Chem.*, 271:35078–35087, 1998.

[349] V. Molteni, J. Greenwald, D. Rhodes, Y. Hwang, W. Kwiatkowski, F. D. Bushman, J. S. Siegel, and S. Choe. *Biol. Crystallogr.*, 57:536–544, 2001.

[350] G. Bujacz, J. Alexandratos, Z. L. Qing, C. Clement-Mella, and A. Wlodawer. *FEBS lett.*, 398:175–178, 1996.

[351] J. Greenwald, V. Le, S. L. Butler, F. D. Bushman, and S. Choe. *Biochemistry*, 38:8892–8898, 1999.

[352] L. S. Beese and T. A. Steitz. *EMBO J.*, 10:25–33, 1991.

[353] D. C. van Gent, A. A. Groeneger, and R. H. Plasterk. *Proc. Natl. Acad. Sci. USA*, 89:9598–9602, 1992.

[354] D. C. van Gent, A. A. Groeneger, and R. H. Plasterk. *Nucleic Acids Res.*, 21:3373–3377, 1993.

[355] L. De Luca, G. Vistoli, A. Pedretti, M. Barreca, and A. Chimirri. *Biochem. Biophys. Res. Commun.*, 336:1010–1016, 2005.

[356] A. S. Fanci. *Science*, 262:1011–1018, 1993.

[357] H. W. Sheppard and M. S. Ascher. *Annu. Rev. Microbiol.*, 46:533–564, 1992.

[358] S. J. Little, S. Holte, J. P. Routy, E. S. Daar, M. Markowitz, A. C. Collier, R. A. Koup, J. W. Mellors, E. Connick, B. Conway, M. Kilby, L. Wang, J. M. Whitcomb, N. S. Hellmann, and D. D. Richman. *New Engl. J. Med.*, 387:385–394, 2002.

[359] N. Neamati. *Expert Opin. Ther. Pat.*, 12:709–724, 2002.

[360] R. Dayam and N. Neamati. *Curr. Pharm. Des.*, 9:1789–1802, 2003.

[361] S. P. Gupta and A. N. Nagappa. *Curr. Med. Chem.*, 10:1779–1794, 2003.

[362] C. Maurin, F. Bailly, and P. Cotelle. *Curr. Med. Chem.*, 10:1795–1810, 2003.

[363] A. A. Johnson, C. Marchand, and Y. Pommier. *Curr. Top. Med. Chem.*, 4:671–686, 2004.

[364] R. Dayam, J. Deng, and N. Neamati. *Med. Res. Rev.*, 26:271–309, 2006.

[365] R. Elber and M. Karplus. *J. Am. Chem. Soc.*, 112:9161–9175, 1990.

[366] A. Roitberg and R. Elber. *J. Chem. Phys.*, 95:9277–9287, 1991.

[367] Y. Goldgur, R. Craigie, G. H Cohen, T. Fujiwara, T. Yoshinaga, T. Fujishita, H. Sugimoto, T. Endo, H. Murai, and D. R. Davies. *Proc. Natl. Acad. Sci. USA*, 96:13040–13043, 1999.

[368] N. Guex and M. C. Peitsch. *Electrophoresis*, 18:2714–2723, 1997.

[369] M. Barreca, K. W. Lee, A. Chimirri, and J. M. Briggs. *Biophys. J.*, 84:1450–1463, 2003.

[370] A. Brigo, K. W. Lee, G. I. Mustata, and J. M. Briggs. *Biophys. J.*, 88:3072–3082, 2005.

[371] W. Weber, H. Demirdjian, R. D. Lins, J. M. Briggs, R. Ferreira, and J. A. McCammon. *J. Biomol. Struct. Dyn.*, 16:733–745, 1998.

[372] A. Brigo, K. W. Lee, F. Fogolari, G. I. Mustata, and J. M. Briggs. *Proteins*, 59:723–741, 2005.

[373] L. De Luca, G. Vistoli, A. Pedretti, M. L. Barreca, and A. Chimirri. *Biochem. Biophys. Res. Commun.*, 336:1010–1016, 2005.

[374] J. C. Chen, J. Krucinski, L. J. Miercke, J. S. Finer-Moore, A. H. Tang, A. D. Leavitt, and R. M. Stroud. *Proc. Natl. Acad. Sci. USA*, 97:8233–8238, 2000.

[375] J. Y. Wang, H. Ling, W. Yang, and R. Craigie. *EMBO J.*, 20:7333–7343, 2001.

[376] M. Cai, R. Zheng, M. Caffrey, R. Craigie, G. M. Clore, and A. M. Gronenborn. *Nat. Struct. Biol.*, 4:2669–2674, 1997.

[377] K. Podtelezhnikov, K. F. D. Gao, Bushman, and J. A. McCammon. *Biopolymers*, 68:110–120, 2003.

[378] Z. Xiang and B. Honig. *J. Mol. Biol.*, 311:421–430, 2001.

[379] M. P. Jacobson, R. A. Friesner, Z. Xiang, and B. Honig. *J. Mol. Biol.*, 320:597–608, 2002.

[380] A. Amadei, A. B. A. Linssen, and H. J. C. Berendsen. *Proteins*, 17:412–425, 1993.

[381] D. M. F. vab Aalten, J. B. C. Findlay, A. Amadei, and H. J. C. Berendsen. *Prot. Eng.*, 8:1129–1135, 1996.

[382] D. M. F. van Aalten, A. Amadei, J. B. C. Bywater, R. Findlay, H. J. C Berendsen, C. Sander, and P. F. W. Stouten. *Biophys. J.*, 70:684–692, 1996.

[383] A. Krohn, S. Redshaw, J. C. Ritchie, B. J. Graves, and M. H. Hatada. *J. Med. Chem.*, 34:3340–3342, 1991.

[384] Z. Chen, Y. Li, E. Chen, D. L. Hall, P. L. Darke, C. Culberson, J. A. Shafer, and L. C. Kuo. *J. Biol. Chem.*, 269:26344–26348, 1994.

[385] D. A. Case, D. A. Pearlman, J. W. Caldwell, III Cheatham, T. E., J. Wang, Simmerling C. L. Ross, W. S., T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, Crowley M. Tsui V. Vincent, J. J., H. Gohlke, Duan Y. Radmer, R. J., J. Pitera, I. Massova, Seibel G. L., U. C. Singh, P. K. Weiner, and Kollman P. A. Amber, 2002.

[386] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and Klein. M. L. *J. Chem. Phys.*, 79:926–935, 1983.

[387] T. M. Jenkins, A. B. Hickman, F. Dyda, R. Ghirlando, D. R. Davies, and R. Craigie. *Proc. Natl. Sci. USA*, 92:6057–6061, 1995.