Molecular dynamics simulations of HIV-1 protease complexed with saquinavir

Simon James Watson

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

Department of Infection & Immunity,
University College London,
University of London,
2009

I, Simon James Watson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Inhibition of the Human Immunodeficiency virus type-1 (HIV-1) protease enzyme blocks HIV-1 replication. Protease inhibitor drugs have successfully been used as a therapy for HIV-infected individuals to reduce their viral loads and slow the progression to Acquired Immune Deficiency Syndrome (AIDS). However, mutations readily and rapidly accrue in the protease gene resulting in a reduced sensitivity of the protein to the inhibitor. In this thesis, molecular dynamics simulations (MDS) were run on HIV proteases complexed with the protease inhibitor saquinavir, and the strength of affinity calculated through MMPBSA and normal mode analysis.

We show in this thesis that at least 13 residues can be computationally mutated in the proteases sequence without adversely affecting its structure or dynamics, and can still replicate the change in binding affinity to saquinavir caused by said mutations. Using 6 protease genotypes with an ordered decrease in saquinavir sensitivity we use MDS to calculate drug binding affinity. Our results show that single 10ns simulations of the systems resulted in good concurrence for the wild-type (WT) system, but an overall strong anti-correlation to biochemically derived results. Extension of the WT and multi-drug resistant (MDR) systems to 50ns yielded no improvement in the correlation to experimental. However, expansion of these systems to a 10-repetition ensemble MDS considerably improved the MDR binding affinity compared to the biochemical result.

Principle components analysis on the simulations revealed that a much greater configurational sampling was achieved through ensemble MD than simulation extension. These data suggest a possible mechanism for saquinavir resistance in the MDR system, where a transitioning to a lower binding-affinity configuration than WT occurs. Furthermore, we show that ensembles of 1ns in length sample a significant proportion of the configurations adopted over 10ns, and generate sufficiently similar binding affinities.

To my beloved wife, Rachel,

without whom,

none of this would have been possible

First and foremost I would like to thank my wife, Rachel, for guiding and supporting me through the many lows during writing, and for proof-reading the many drafts of this thesis. Without you, this thesis would never have been written.

I would also like to thank my friends and family: in particular, my parents for their continual love, support, and encouragement; my brother, Andrew, for keeping my spirits high; and John Davies, without whom I would not be in the position I am today. Thanks also go out to Jenny and Geoff Wilkinson, for their advice and support. I am forever indebted to you all.

I am grateful to Paul Kellam and Peter Coveney for advising me throughout my PhD course, and to my many colleagues, past and present, at the Windeyer Medical Institute and the Centre for Computational Science - Arshad Khan, Stephane Hue, Dan Frampton, Rob Gifford, Eve Coulter, Jane Rasaiyaah, Anne Palser, Lucy Dalton-Griffin, Imogen Lai, David Wright, Owain Kenway, Ileana Stoica, Stefan Zasada, Steven Manos.

A special note of thanks goes to Kashif Sadiq for his advice and expertise during our collaboration.

The work in this thesis was funded by the Medical Research Council.

Contents

List of Figures

List of Tables

Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
AMBER	Assisted Model Building with Energy Refinement
DNA	Deoxyribonucleic acid
Ele	Electrostatics energy
G, ΔG	Gibb's free energy, change in Gibb's free energy
H, ΔH	enthalpy, change in enthalpy
HIV	Human Immunodeficiency Virus
MD	Molecular Dynamics
MMGBSA	Molecular-Mechanics / Generalised-Born Surface Area
MMPBSA	Molecular-Mechanics / Poisson-Boltzmann Surface Area
NAMD	Nanoscale Molecular Dynamics
NMODE	normal mode analysis
PDB	Protein Data Bank
RNA	Ribonucleic acid
S, ΔS	entropy, change in entropy
SQL	Structured Query Language
TI	Thermodynamic Integration
UML	Unified Modeling Language
VdW	*van der Waals* energy

Chapter 1

Introduction

1.1 Biological Introduction

1.1.1 Proteins

Proteins are the most abundant and diverse biological macromolecule, making up over half of a cell's dry weight [85, 1]. Their importance is reflected by their remarkable range of functions, which are indispensable to life; including enzymatic catalysis, intra- and extra-cellular transport, control of cellular growth and differentiation, and defence against pathogens and antigens [107].

Proteins are composed of chains of amino acids, ranging in length from, for example, 51 monomers in ribonuclease A, to the largest protein Titin, containing 34,350 amino acids. These amino acid monomers are not all identical; although they all contain a hydrogen atom, a carboxyl group and an amino group surrounding a central carbon atom, they differ in the composition of their **R group** that is also attached to the central carbon atom, shown in Figure 1.1. As the other chemical groups are identical between different amino acids, the R group (also referred to as the side chain) alone determines the amino acid's biophysical properties such as its size, polarity and hydrophobicity [85]. There are 20 standard amino acids encoded by the genetic code, each with a distinctive side chain. These side chains are shown in Figure 1.2.

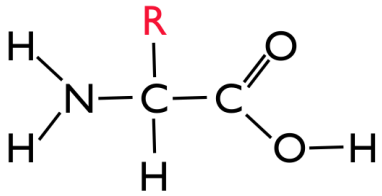


Figure 1.1: Chemical structure of the shared chemical groups of amino acids. The central carbon atom is denoted C_α, around which the functional groups are bonded. The **R** group, highlighted in red, is also called the side-chain. This differs between amino acids, giving them their different chemical properties [12].

These amino acids are joined together through a condensation reaction that covalently joins the amino group of one with the carboxyl group of a second, forming a peptide linkage (See Figure 1.3). For this reason, a protein chain is also referred to as a polypeptide chain [85]. The sequence of the amino acids in the polypeptide chain is referred to as the protein's primary sequence, and directly determines the biochemical properties of the protein, such as its size, shape and function by dictating the way that the protein folds into its secondary and tertiary structures, and, in the case of multi-chain proteins, its quaternary structure. When amino acids interact at a local level, they can form three different local secondary structures:

1. **α-helix**: These right-handed helices form rod-like structures where the CO atoms of an amino acid's carboxyl group are hydrogen-bonded to the NH atoms in the amino group of the amino acid four monomers ahead. In such structures, each Cα atom is related to the next by a rise of 1.5Å along the helical axis, and a rotation of 100°.

2. **β-sheet**: Much as with α-helices, β pleated sheets occur through hydrogen-bond interactions between CO and NH atoms of different amino acids. However, unlike in α-helices, these non-bonded interactions occur between residues that are located in adjacent chains lying in parallel to the plane of their chains. If the two adjacent chains run in the same direction then they are termed parallel β-sheets; if they run in opposite directions then they are termed antiparallel β-sheets. In contrast to α-helices, the axial distance of interacting residues in β-sheets is around 3.5Å.

3. **β-turn**: As with the other two secondary structures, a β-turn in a polypeptide chain is caused by hydrogen-bond interactions between the CO and NH atoms of adjacent residues. However, in the case of a β-turn, this interaction is between two amino acids four residues apart along the primary sequence. In bringing these two residues in close enough contact for the hydrogen-bonding to occur, the intervening amino acids curl around to form a tight turn. As β-turns abruptly change the direction of a polypeptide chain, they are often found in close conjunction with antiparallel β-sheets.

It is not just at the local level that amino acids can associate with one another. Once a protein has formed its secondary structures, further large-scale spatial rearrangement of the protein occurs that bring distant regions of the primary sequence into close proximity. This *tertiary structure* is stabilised through both non-bonded forces, such as *van der Waals* interactions between hydrophobic residues, and covalent interactions, such as disulphide bridges between two cysteine residues brought into close proximity [85]. Disulphide bridges are often found in secreted or cell-surface proteins, and are formed as shown in Equation 1.1.

$$-CH_2 - \mathbf{SH} + \mathbf{HS} - H_2C- \Longleftrightarrow -CH_2 - \mathbf{S\text{-}S} - H_2C- + 2H^+ + 2e^- \qquad (1.1)$$

In order for the disulphide bridge to be formed, the protein must be in the presence of an oxidative environment, which is why in eukaryotes this occurs in the rough endoplasmic reticulum, rather than the reducing environment of the cytosol.

Once the polypeptide chain has folded into its tertiary structure, it may associate with other polypeptide chains to form the functional protein. The arrangement to which these chains assemble into a complex is referred to as the protein's *quaternary structure*, and each polypeptide chain making up the complex is termed a subunit or monomer. Each subunit of a multi-subunit protein complex does not have to be identical; that is, two different polypeptide chains can interact to form a functional protein. In such cases, the protein is referred to as a heteromeric protein. In contrast, protein complexes composed of identical subunits are referred to as a homomeric proteins [1]. When describing hetero- or homo-meric protein complexes, it is common to include the

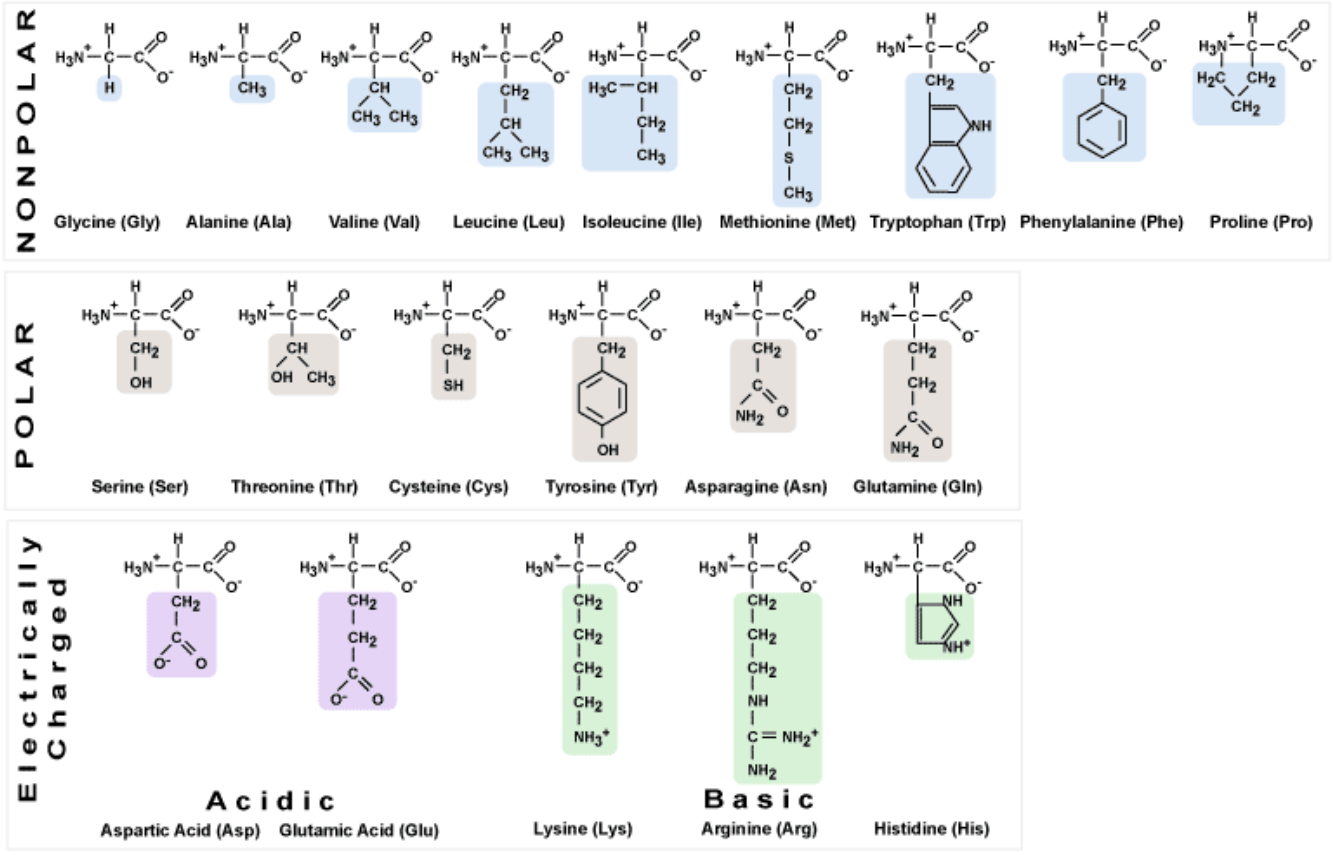


Figure 1.2: Chemical structure of the 20 amino acids' side-chains. The structures are grouped according to their side-chains' common properties. The invariable region of the amino acid (the atoms that are not coloured) is shown with an NH_3^+ amino group and a COO^- carboxyl group, which are the predominant forms at pH 7 [107]. Figure taken from <http://www.personal.psu.edu/staff/m/b/mbt102/bisci4online/chemistry/> [5].

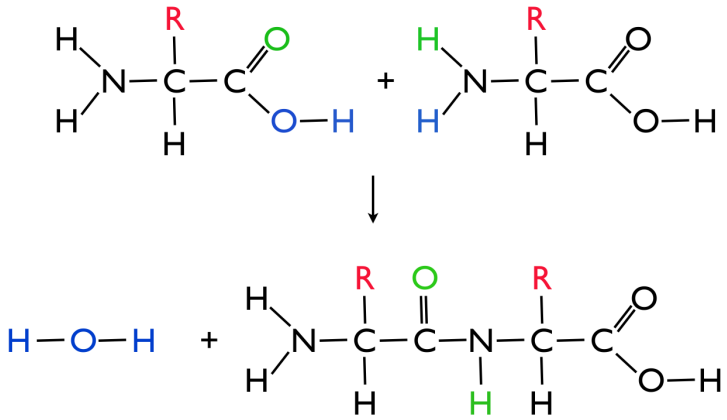


Figure 1.3: Chemical illustration of the formation of a peptide linkage between two amino acids. Highlighted in blue are the atoms that are lost as water in the reaction. In green are the carboxylic-oxygen and amino-hydrogen to aid visualisation. The peptide bond is formed between carboxyl-carbon and amino-nitrogen of adjacent amino acids. Figure adapted from Stryer (2000) [107].

number of subunits in the description. For example, HIV's aspartyl protease enzyme is composed of two identical polypeptide chains, and is therefore referred to as a homodimeric protein.

In theory, the quaternary structure of a protein should be deducible from its primary sequence. Given a particular order of amino acids, their interactions on a local scale can be determined, and subsequently the interactions of these secondary structures to form the protein or subunit. However, the actualisation of this has proven to be much more difficult, so alternative methods are employed to determine the tertiary or quaternary structure of a protein:

1. **X-ray crystallography**. X-ray crystallography is a technique for determining the atomic position of macromolecules. Unlike NMR, however, the macromolecules must be in crystalline form for the method to work. A narrow beam of x-rays with a wavelength of 1.5Å is directed through the protein crystal. While most of the x-rays pass straight through the protein sample, those that pass in close proximity of an atom will be deflected, forming a scattered diffraction pattern in which overlapping electromagnetic waves that are in phase amplify each other, and those out-of-phase interfere with each other and cancel each other

out. The position and intensity of the spots in the resultant x-ray diffraction pattern gives information about the position of the atoms in the crystal. The crystal is rotated step-by-step through 180° and the process repeated to attain the 3-dimensional structure of the protein, which is reverse-calculated from the diffraction patterns by determining the positions of all the atoms and combining it with knowledge of the amino acid sequence. The fidelity of this structure depends on the resolution of the diffraction patterns attained. At a resolution of 6Å the basic polypeptide chain can be distinguished, but little structural information can be seen. As the resolution increases, groups of atoms 2.8Å - 4.0Å apart can be distinguished, and at even higher resolutions, individual atoms 1.0Å - 1.5Å apart can be distinguished. In turn, the resolution of the diffraction pattern is determined by the perfection of the initial protein crystals, which is dependent on both the crystallisation protocol and protein itself; some proteins, particularly membrane proteins, are difficult to crystallise [1, 107].

2. **NMR spectroscopy**. As with x-ray crystallography, nuclear magnetic resonance spectroscopy (NMR) is a technique for determining the atomic positions of molecules. In contrast to X-ray crystallography, however, NMR data is determined from a solution of protein. This allows for protein-structure determination under a wider range of physiological conditions, such as pH, temperature, and salt concentration [130]. NMR exploits the fact that nuclei containing an uneven number of nucleons (*e.g.* ^{1}H, ^{14}N, ^{13}C, ^{31}P which are abundant in biological molecules) have a net spin and therefore generate a magnetic moment like a dipole. By placing the molecule in a constant magnetic field, those nuclei with a net spin will realign themselves to retain a low-energy configuration. Applying an electromagnetic pulse results in the nuclei transitioning to a high-energy spin state, however this pulse must be at the exact frequency that its photons contain the energy required to transition the nucleus to this high-energy state. This frequency, termed the resonance frequency, is unique for each atom, and more importantly differs for a particular atom depending on its surrounding chemical environment. Therefore the resonance frequency of a lone ^{1}H hydrogen nuclei will be different to the resonance frequency of the hydrogen nuclei in methane

(CH_4). The deviation in resonance frequency from a standard reference molecule is termed the 'chemical shift', and this gives information on the chemical structure; for example a ^{13}C chemical shift of 97ppm indicates that it is part of a CCl_4 group [88, 94, 107, 130].

While chemical shifts reveal the functional groups present in the molecule, they do not give information on the secondary or tertiary structure of the molecule. However, the phenomena of spin-spin coupling and nuclear Overhauser effect (NOE) allow the molecule's conformation to be elucidated. With spin-spin coupling, non-equivalent hydrogen atoms that are separated by up to 3 covalent bonds will couple their nuclear spins, appearing as split lines on an NMR spectrum. NOE is the transference of magnetisation from an excited nucleus to an unexcited one within 5Å as it relaxes back to its low-energy state. These appear as off-diagonal points on a 2-D nuclear Overhauser enhancement spectroscopy (NOESY) spectrum and can be interpreted to determine the 3-D relationship between the nuclei involved. ^1H is most commonly employed for these analyses due to the natural abundance of the isotope as well as its relative abundance in biological macromolecules [49, 94, 107].

3. **Homology modelling**. The principle behind homology modelling is that new proteins usually evolve through modification of duplicated genes. Therefore proteins with related functions often have similar sequences and basic structures. Members of the serine protease family, for example, share 40% primary sequence similarity, and have near identical tertiary structures. Therefore, by comparing a protein's primary sequence against those of other functionally-related proteins with known structures, its own structure can be inferred [1]. This is a computational method of structure prediction, and as more organisms' genomes are sequenced, and more depositories filled with the protein sequences determined, this method will become more powerful.

1.1.2 Enzymes

As mentioned in Section 1.1.1, the function of certain proteins is to catalyse the innumerable chemical reactions that occur in biological systems, increasing their rate of

reaction [107]. The reactions still occur in the absence of these catalysing proteins, but in some cases their rates are so low that a single transition from reactant to product may only occur once every year. The role of the enzyme, therefore, is to increase the rates of biochemical reactions to levels such that the products are harness-able by biological systems [1].

In enzyme-catalysed reactions, the reactant molecules are referred to as *substrates*, and in order for the enzyme to fulfil its role of accelerating the rate of transition to products, it must bind to the substrate (or substrates if the reaction involves more than one reactant species) to form an enzyme-substrate complex. After the reaction has taken place, the resultant product (or products) is released by the enzyme, which is then free to catalyse another reaction. As is described in more detail in Section 1.3.3, enzymes also increase the rate of the reverse reaction by the same factor as the forward reaction, so the products can be catalysed back into the reactant molecules by binding to the enzyme. This is shown diagrammatically in Equation 1.2 [1, 85, 107].

$$E + S \rightleftharpoons ES \rightleftharpoons EP \rightleftharpoons E + P \tag{1.2}$$

The region of the enzyme that the substrate binds to is called the *active site*. This is a three-dimensional cleft in the surface of the protein, formed by distal residues that come together as the protein folds into its tertiary or quaternary structure. The volume taken up by the active site comprises a relatively small percentage of the total volume of the protein [107]. Figure 1.4 shows an example of the surface area taken up by the active site in HIV-1's aspartyl protease enzyme. The first step in catalysis is the formation of an enzyme-substrate complex. In order for the complex to form, the enzyme must form multiple non-covalent bonds with the substrates. Depending on the chemical nature of the substrates and the active-site residues, the free-energy change associated with binding ranges from -3 to -12 kcal/mol; weakly-interacting substrates may only share hydrophobicity with the residues that make up the active site, whilst strongly-interacting substrates may share electrostatic bonds, hydrogen bonds, *van der Waals* interactions along with the shared hydrophobicity with the active site residues. The shape of the active site is an important factor in the strength of the

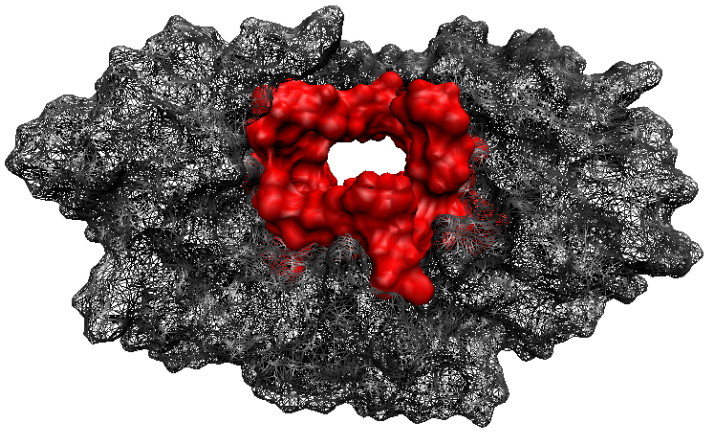


Figure 1.4: Cartoon representation of HIV-1's aspartyl protease showing the residues that form the surface area of the active site in red, and the remainder of the quaternary structure in grey wireframe. Note how the surface area of the enzyme's active site compares to the much larger surface area of the enzyme as a whole. The entire catalytic activity of this enzyme derives from just two aspartic-acid residues located at the bottom of the binding-groove running through the enzyme. This image was created from the 1FB7 PDB file using VMD.

interaction, because while a single *van der Waals* force is weak, the combined effect of multiple enzyme-substrate *van der Waals* bonds becomes a significant interaction. The strength of the *van der Waals* interaction increases as the distance between the relevant atoms decreases. However, as the atoms are brought closer together, strong repulsive forces arising from the atoms' electron clouds' electrostatic repulsion abruptly increases the pairwise potential energy. The interaction of these opposing forces results in an equilibrium interatomic distance, termed the *van der Waals* contact distance, where the pairwise potential energy is lowest (Figure 1.5). The complementarity between the active site and the substrate in the enzyme-substrate complex is therefore important as it ensures that the multiple *van der Waals* interactions are at their contact distance, thereby increasing their strength [7].

Once the substrate binds the enzyme to form a complex, as long as it has enough free energy to overcome the activation barrier, it will react to form the product. The biochemistry underlying this is described in Section 1.3.3, but the different biological mechanisms utilised by the enzyme to lower the activation barrier, and therefore

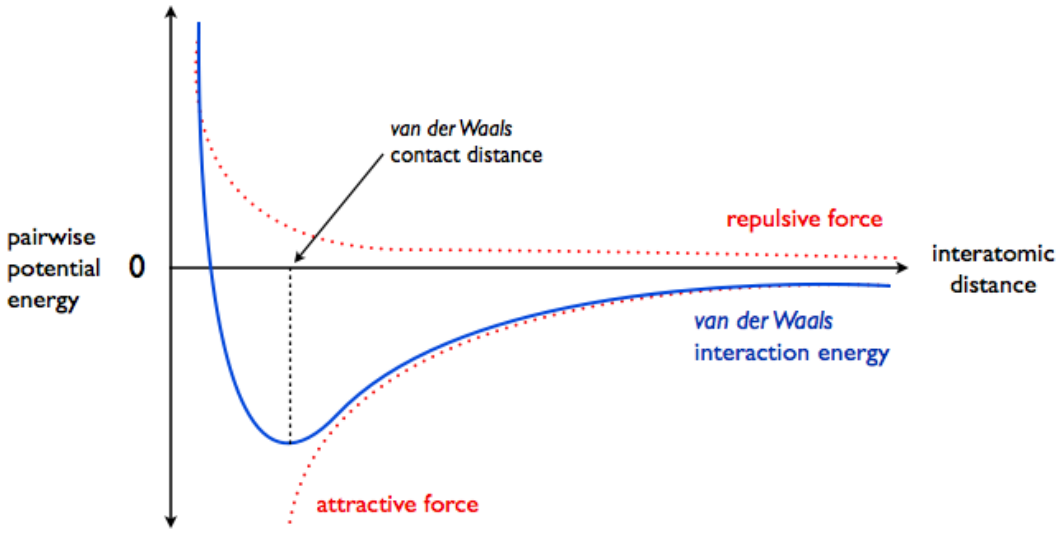


Figure 1.5: *van der Waals* interaction energy graph showing how the energy of the attractive and repulsive forces (in red) change as 2 atoms are brought closer together. The interaction between these opposing forces results in the *van der Waals* interaction energy (in blue), which is lowest at the 'contact distance' between the atoms. Figure adapted from Berg *et al.* (2002) [7].

catalyse the reaction, will be described here:

- **Substrate re-orientation**. As previously mentioned, the shape of the active site is complementary to that of its substrates. If a reaction involves multiple substrates, they would only be able to fit into the active site in the particular orientation that allows them to react. An example of this is the catalysis of peptide bond formation; the peptide bond can only be formed between the carboxyl group of one amino acid and the amino group of a second amino acid. So for this reaction to take place, the active site of the enzyme must accommodate these two amino acids in such orientation that the carboxyl group and amino group are facing each other and close enough for the peptide bond to form between them [85]. It used to be thought that the enzyme-substrate complementarity could be described through a *lock-and-key* metaphor, where the shape of the active site was exactly complementary to the substrate. However, Daniel E. Koshland Jr.'s *induced-fit* postulate is now thought to be more accurate. This hypothesises that the unbound active site is not necessarily complementary to the substrate, but

once bound, the additional non-bonded forces acting on the enzyme cause it to change shape such that it becomes complementary to the substrate [107]. HIV-1 protease provides an example of this; as seen in Figure 1.10, the active site runs through the middle of the enzyme, with two pairs of anti-parallel β-strand 'flaps' forming the roof of the active site. Hornak and Simmerling (2007) hypothesised that these flexible flaps are 'open' in the enzyme's unbound state to allow entry of substrate into the active site, and that upon substrate binding they undergo configurational changes to close over the substrate, locking it in the active site such that the catalytic residues are close enough to the substrate for catalysis to occur. These flap conformations are shown diagrammatically in Figure 1.11 [34].

- **Re-distribution of charge in the substrate**. Enzymes can also catalyse reactions by changing either the charge on the substrates, or by altering the electron distribution of the substrates such that the reaction is more likely to occur. This charge re-distribution can directly involve the side-chains in the enzyme's active site; in these cases the amino acids directly involved are termed the catalytic residues [85]. A detailed example of this mechanism is described for HIV-1 protease in Section 1.3.3. In brief, the enzyme's two catalytic residues cause a redistribution of electron charge both in the attacking water molecule and in the peptide bond, destabilising the peptide bond and making it more susceptible to attack. The catalytic residues then directly mediate the transfer of a proton from the water molecule to the substrate, breaking the peptide bond [71]. This catalytic mechanism can also be achieved by a metal ion co-factor, such as zinc, copper or iron, present in the quaternary structure of the enzyme [85].

- **Induced strain in the substrate**. Through the action of non-bonded forces on the substrate upon formation of the enzyme-substrate complex, bonds in the substrate can stretch, causing the substrate to adopt an unstable configuration. This can make the substrate more reactive to attacking molecules [85]. In fact, the unstable configuration that the enzyme forces the molecules to adopt is that of the substrate's transition structure for the reaction and the enzyme has its greatest affinity for this transition structure. For enzymes employing this mechanism

for a reversible reaction it is easy to see how it accelerates both the forward and reverse reactions by the same amount; when either the reactant or the product enters the active site, they are both conformationally strained into a structure more closely resembling the transition-state structure between the two, allowing the reaction to occur more readily in either direction. As the active site is most complementary to the transition structure, it will have mismatch with both the reactants and the products, so its catalytic activity will be determined by how easily it can strain the molecule into the transition-state structure [14].

In order for the enzyme to strain its substrates into the transition-state structure, it must undergo a certain degree of conformational change itself, brought about by the additional non-bonded forces from the substrate molecules in its active site. This, in combination with the fact that enzymes employing this mechanism have highest complementarity to the transition structure, gives credence to the induced-fit model of enzyme-substrate binding over the lock-and-key model. This mechanism has had a large impact on drug design, as it was realised that for enzymes that employ this method, for the inhibitor to be most effective it must be the most attractive to the enzyme. This requires that its structure resemble the transition-state structure, for which the enzyme has highest affinity, rather than the unbound substrate which is deformed into a more attractive structure by the enzyme. The biological effect of inhibitors will be described in more detail later in this section; the biochemistry of enzyme-inhibitor kinetics is described in detail in Section 1.3.3 [85, 107].

The action of enzymes can be both enhanced and inhibited through the action of specific atoms or molecules, termed *activators* and *inhibitors* respectively. These are not ubiquitous for all enzymes; not all enzymes have activators, nor do all enzymes have inhibitors [107].

- **Activators**. An enzyme activator is a species that reversibly binds to an enzyme to increase its activity, whilst itself not undergoing any net change in the reaction. It can therefore be considered the converse of reversible inhibition, which is described later in this section. The activator may be an inorganic metal ion,

such as Mg^{2+}, or an organic molecule such as fructose 2,6-bisphosphate [85].

- **Inhibitors**. As mentioned earlier, an inhibitor is a species that interferes with the normal action of an enzyme, preventing it from catalysing its reaction, either in part or completely. There are two types of inhibitors: reversible and irreversible. Irreversible inhibitors act by either reacting with the enzyme, resulting in an altered enzyme that is no longer enzymatically active [12] or by forming very strong covalent or non-covalent bonds with the enzyme, resulting in such a low dissociation constant that soon all enzyme molecules are saturated with inhibitor and no substrates can enter the active site [107]. An example of an irreversible inhibitor is the β-lactam class of antibiotics, which irreversibly bind to the bacterial-enzyme glycopeptide transpeptidase. This inhibition prevents the enzyme from catalysing the cross-linkage of peptidoglycan macromolecules in the bacterial cell wall, thus inhibiting its growth [107]. In contrast, reversible inhibitors bind to the enzyme to form an enzyme-inhibitor complex, but the association is weak and so the complex is only transient; the complex dissociates back into free inhibitor and enzyme which is then able to bind substrate and catalyse a reaction. Examples of reversible inhibitors include the viral protease inhibitors such as saquinavir, targeting HIV; or telaprevir, targeting hepatitis C [85]. There are two types of reversible inhibitors: competitive and non-competitive. Non-competitive inhibitors bind to binding-sites on the enzyme at a site other than the active site. This binding causes a change in the conformation of the enzyme, particularly around the active site, that results in either inability of the substrate to bind, or reduced catalytic activity of the enzyme due to its reduced effectiveness to lower the activation barrier in its new conformation [12, 107].

  Competitive reversible inhibitors have a different effect to non-competitive inhibitors because, instead of changing the enzyme's conformation, they bind directly to the active site in place of the substrate. By binding to the active site, it blocks the substrate's access, and as a result that enzyme is unable to catalyse its natural reaction [107]. The action of reversible inhibitors is therefore to reduce the rate of reaction by reducing the proportion of enzymes available to bind substrate

molecules at any point in time. This action can be overcome by increasing the concentration of enzyme's natural substrate so that is has a much greater chance of binding to a substrate molecule rather than an inhibitor molecule. In order for a competitive inhibitor to be effective, the enzyme must have as strong an affinity as possible for it, to lower the dissociation constant and therefore increase the time the inhibitor spends in the active site. This can be achieved by designing the inhibitor to mimic the enzyme's natural substrate; designing it to mimic the substrate's structure ensures that the enzyme's affinity for the inhibitor is similar in magnitude to its affinity for the substrate.

As mentioned earlier in this section, enzymes actually do not have the highest affinity for the substrate in its ground state. It was realised during the 1920s that as an enzyme distorts the substrate into the transition-state structure, its affinity for the substrate increases. The enzyme actually has its highest affinity with the transition-state structure [128]. It was therefore theorised by Linus Pauling in 1946 that unreactive compounds mimicking the enzyme's substrate's transition-state structure would make a more powerful antagonist than a compound resembling the ground state of the substrate [12]. Such inhibitors are termed *transition state analogs* and the principle has proven to be very effective in the design of novel inhibitors. An example of a transition state analog is the HIV protease inhibitor indinavir. This peptidomimetic inhibitor was developed by Merck & Co. and approved by the FDA as the third HIV protease inhibitor in 1996. The K_i between HIV-1 protease and indinavir was shown to be 0.34 nM [9] compared to a K_D of 0.27 nM between the protease and its natural substrate [46].

Due to the fallible nature of replicative enzymes such as DNA polymerase and reverse transcriptase, enzymes can evolve to escape the effect of inhibitors. Mutations that accumulate in the enzyme's genetic sequence can result in an alteration in the protein's primary sequence. These primary sequence mutations can cause a decrease in the binding affinity between the inhibitor and the enzyme, or they can cause an increase in affinity for the enzyme's natural substrate over the inhibitor. Alternatively, they can

cause a drop in affinity for both inhibitor and substrate, but with a larger decrease for the inhibitor such that the inhibitor's effect is sequestered. To this mutated primary sequence, further mutations can be added, resulting in an even greater resistance to the inhibitor's effect. Eventually the enzyme can accumulate enough mutations to circumvent the effect of the inhibitor so that it is able to fulfil its catalytic role. This may have resulted in a decreased affinity for the substrate compared to the original enzyme, or a lowered catalytic function of the enzyme, but in the presence of the inhibitor this mutant sequence restores the enzyme's phenotype as best as it can. A fuller description of enzyme mutations in the context of HIV-1 protease inhibition is given in Section 1.2.4.

1.2 Virological Introduction

When viruses were first discovered in 1899, they were said to represent the most basic form of life; being the smallest known life-form encoding the smallest number of proteins required to sustain itself . As obligate intracellular parasites, they are unable to replicate on their own, so require a host cell within which to propagate [47]. In the simplest of terms, their sole purpose is to transport their genome into a host cell so that the cell's transcriptional and translational machinery can be hijacked and redirected into producing the proteins encoded by its genome. These proteins' sole purpose is to either directly or indirectly produce new virions which can then infect other cells. However, the apparently simplicity of viruses has since been shown to be a misconception, and an entire field has been developed to study the complexities of viruses and the interactions with their vast array of host cells. An example of such complexities is the range of genetic material exhibited in viral genomes. Unlike eukaryotes and prokaryotes, which consistently use double-stranded DNA as their genetic material, viruses are known to utilise single-stranded DNA (*e.g. Parvoviridae*), double-stranded DNA (*e.g. Herpesviridae*), single-stranded RNA (*e.g. Coronaviridae*), and double-stranded RNA (*e.g. Reoviridae*) for their genomes. Furthermore, the DNA genomes can be either linear or circular, and RNA genomes can be either positive-sense or negative-sense. Viruses of the *Retroviridae* family store their genetic material as a positive-sense RNA molecule, which is converted to a linear double-stranded DNA intermediary prior

to transcription and translation [118]. This is especially interesting because this goes against the Central Dogma of molecular biology, which states that the flow of genetic material passes from DNA to protein through an RNA intermediary. It is testimony to the variability of viruses that no single classification scheme is able to comprehensively group viruses according to their basic properties. For example, the Baltimore classification separates viruses into seven groups based on their genetic material. However, virus families within a group do not share other properties such as whether the virions are surrounded by a lipid envelope prior to infection, cell tropism, or the nature of the disease they cause [47].

Nevertheless, despite their inability to replicate outside of a host cell, viruses are incredibly successful; they are able to infect members of all 5 biological kingdoms [47], and have caused several pandemics in human history. For example, the 1918-1919 influenza pandemic infected over 500 million people worldwide and resulted in more deaths than World War I. This pandemic was caused by an novel strain of influenza that evolved in late 1918. The emergence of the HIV pandemic in 1983 has since resulted in over 33 million people infected worldwide by 2007, and the death of over 2 million [115]. The emergence of these novel forms highlights the incredibly high evolutionary rate of viruses, which appears to be directed to aid in their transmissibility. For example, the influenza virus's genome is split into 8 genomic segments; co-infection of a host cell by 2 different influenza viruses can result in co-packaging of segments from each virus in the progeny virions, resulting in a virus with a novel genotype [118, 110]. Furthermore, viruses with an RNA genome require a virally-encoded **RNA-dependant RNA polymerase** to replicate its genome. These have no proof-reading $3'$ exonuclease activity and therefore the error rates in genome replication are approximately 1 in 10,000 nucleotides. As the average RNA genome is greater than 10kb, each replicated genome will have at least one mutation. Combined with the high replication rate of viruses, a large number of genetically-distinct progeny can be generated upon infection [47]. In addition to emergence of new viral genotypes that cause global pandemics, this has the effect of creating a low level of genetically-variable virions within an infected species. This has important implications when a strong selective pressure such as drug administration is applied to suppress the infection. In this situation any low-level ge-

netic variants which are less suppressed by the drug will become the fittest genotype and will therefore soon become the majority species and circumvent the drug. This is most important in chronic viral infections such as HIV, where anti-retroviral drugs are administered to suppress the virus and slow the progression to Acquired Immunodeficiency Syndrome (AIDS). Due to the presence of genetically-variable 'quasi-species' within an individual, the suppressive effect of a drug is only transient [86].

1.2.1 Global importance of HIV and AIDS

In 1981 the first case of a person presenting a collection of opportunistic infections and tumours that were normally suppressed by the immune system was identified. As more cases were identified, it was noticed that the opportunistic infections were associated with a marked decrease in the levels of circulating T-helper lymphocytes, and so the term Acquired Immunodeficiency Syndrome (AIDS) was coined to described the presentation of the opportunistic infections in the context of this immunosuppression. In 1983 a retrovirus, subsequently termed 'Human Immunodeficiency Virus type 1' (HIV-1), was isolated from the blood of patients with AIDS, whose cell tropism was mediated by the cellular receptor 'cluster of differentiation 4' (CD4). This receptor is present on the surface of T-helper cells, regulatory-T cells, monocytes, macrophages and dendritic cells, which explained how the retrovirus infection lead to immuno-compromisation and subsequent presentation of AIDS.

Since that first case back in 1981, HIV has become a global pandemic, with an estimated 33 million people worldwide living with HIV in 2007. In 2007 alone, there were 2.7 million new HIV infections and 2 million deaths due to AIDS. It is projected that by the year 2015, AIDS will have accounted for 115 million deaths in the 60 countries most affected. The additional macroeconomic impact of the pandemic is vast, with HIV predicted to reduce economic growth in high-prevalence countries by 0.5%-1.5% over the next 10-20 years. Furthermore, the annual global resource allocated for the prevention of new HIV infections, reduction of HIV-related illnesses, and mitigation of the epidemic's economic effects is 10 billion USD [115]. The global impact of the emergence of HIV cannot be underestimated, and therefore there is significant scientific interest in HIV treatment.

1.2.2 HIV Genome, Structure and Life-Cycle

Under the Baltimore classification system HIV is a type VI virus, meaning that it is encoded by a positive-sense, single-stranded RNA genome that replicates via a double-stranded DNA intermediate [118]. Each of its genome's two RNA strands is composed of roughly 9,800 nucleotides, encoding 3 major structural genes; *gag*, *pol* and *env*, and several non-structural accessory genes; *tat*, *rev*, *nef*, *vif*, *vpr*, *vpu* and, in some HIV isolates, *tev* (see Figure 1.6).

Each of the three structural genes encodes a polyprotein product; the Gag polyprotein product of the *gag* gene is cleaved into p24, p6, p7 and p17, which together make up the structure of the protein core that protects the RNA genome. The proteins obtained from cleaving the Pol polyprotein are functional enzymes essential for viral replication and maturation; reverse transcriptase converts single-stranded RNA into double-stranded DNA, and so converts the HIV ssRNA genome into a dsDNA intermediate. Integrase is then responsible for inserting the dsDNA intermediate into the host cell's genome; the reason for this is explained further on in this section. The final protein resulting from cleavage of the Pol polyprotein is protease. This enzyme is responsible for cleaving the translated polyproteins into the functional proteins. This occurs during, and after, the new HIV virion buds from the infected cell and ensures that the progeny virion is a fully-mature infectious particle. The final structural polyprotein is Env, which, when cleaved by host-cell proteases, forms two viral proteins that are post-processed into gp120 and gp41 respectively. These two glycoproteins are located in the envelope surrounding the virus particle when it buds out of the cell, and are responsible for interaction and fusion with target cells, allowing cell entry. The proteins attained from cleavage of the three major structural genes contain everything that the HIV particle needs to form mature, infectious progeny viruses. However, without the accessory proteins, the efficiency of viral transcription is low, and subsequently the viral fitness is lower.

All members of the *Retroviridae* family share the same basic genetic organisation of their major structural genes; immediately $3'$ of the $5'$ long terminal repeat (LTR) is

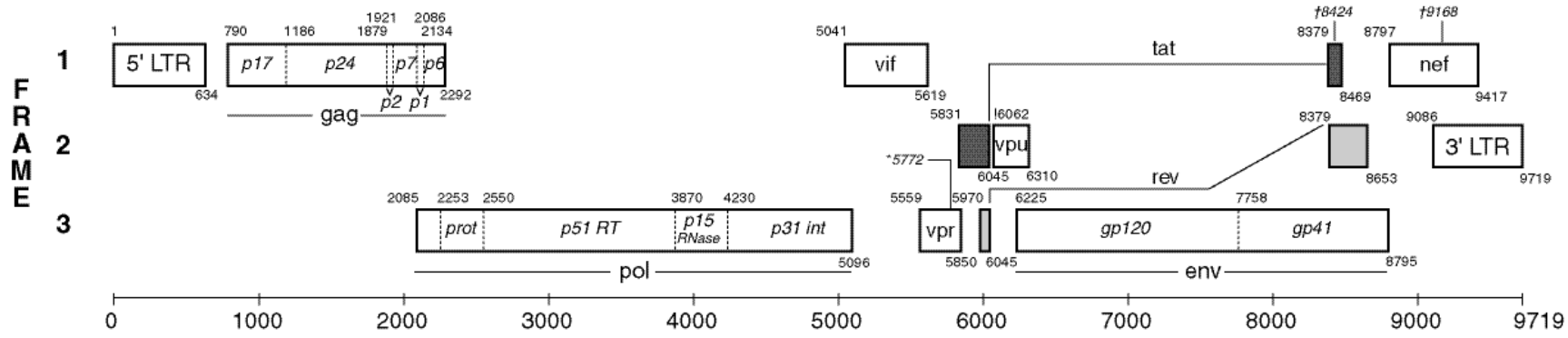


Figure 1.6: Schematic of the Hxb2 experimental consensus sequence HIV genome. Each row corresponds to the reading frame in which the genes lie, and along the bottom is the nucleotide number. Open reading frames (ORFs) are indicated as rectangles, and are split by dashed lines indicating the protein products encoded by that ORF. The starting nucleotide of each gene product is indicated in the top left of each sub-rectangle, and the number in the lower-right indicates the last nucleotide of the stop codon. Image taken from the HIV Sequence Database, <www.hiv.lanl.gov>.

gag, which leads directly onto *pol*, then further downstream is *env*, lying 5′ of the 3′ LTR. Where retroviruses differ is in the relative reading frames of their structural; *gag*, *pol* and *env* are not all found in the same reading frame, which has an important effect in regulating the relative levels of expression for each of these genes. In the case of HIV-1, *pol* is positioned at a reading frame of (-1) relative to *gag*; the start of the *pol* gene actually overlaps the end of the *gag* gene by a single nucleotide. *Gag* and *pol* are actually transcribed as a single polycistronic mRNA molecule, and as Pol lies in a different reading frame it requires the cellular ribosome to undergo a frameshifting event. This is achieved through the formation of an internal secondary structure in the viral mRNA strand; a 'hairpin' structure is formed at the end of the *gag* gene in conjunction with a 'slippery' sequence within the *gag* gene. Roughly 5% of the time the translating ribosome will pause at this hairpin structure and move back one nucleotide, resulting in it continuing to transcribe at a (-1) frameshift. Therefore, 95% of the time only Gag is transcribed, and the other 5% of the time a Gag-Pol polyprotein product is formed. This is important because the relative levels of viral protein products are carefully orchestrated to ensure optimal fitness [47].

The structure of the mature HIV virion is shown in Figure 1.7. The diploid ssRNA genome described above is complexed with viral proteins p6 and p7 resulting in a stable nucleocapsid surrounding the RNA that prevents the genome being digested by host-cell nucleases. Also packaged with the genome are the viral enzymes reverse-transcriptase and integrase, as these are required to convert the genome into its dsDNA intermediate in order for transcription to occur, so must be packed into the virion. Surrounding the genome-nucleocapsid complex is the capsid, composed of multiple copies of the viral p24 protein tessellated into a conical shape. Encompassing the capsid is a sphere of viral-encoded (p17) matrix protein. HIV protease is located in the space between the matrix and capsid because of its role in HIV maturation, which will be explained in more depth further on. The outer layer (envelope) of the virion is a lipid bilayer originating from the plasma membrane of the host cell from which it originated. Spanning the envelope are trimers of the viral glycoproteins gp120 and gp41. These are envelope glycoproteins that determine the virus's tropism; the virus can only infect those particular cells that contain receptors complementary to the gp120/gp41 trimers.

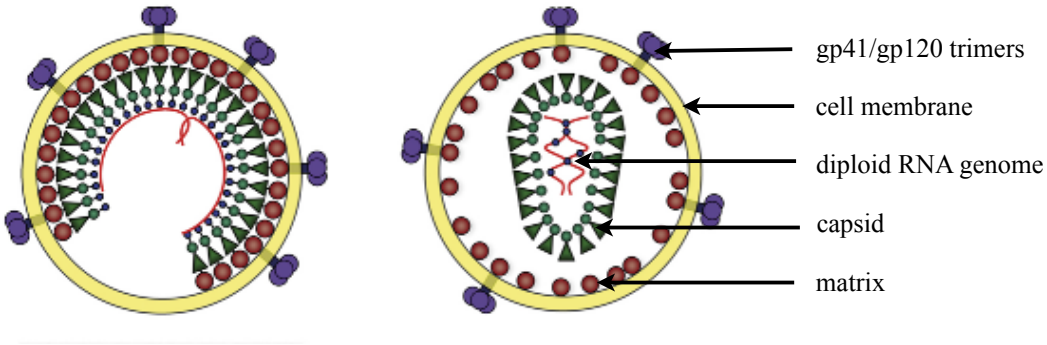


Figure 1.7: Cross-section of an immature HIV virion on the left, and a mature HIV virion on the right. In purple are the gp41/gp120 trimers; in yellow is the lipid membrane formed from budding out of the previous host cell; in maroon are the matrix proteins; in green are the capsid proteins that form a protective fullerene cone around the mature virus's diploid genome, shown in red; attached to the RNA genome are integrase, reverse transcriptase, nucleocapsid proteins, Vpr and p6. Image adapted from Ganser-Pornillos *et al.* (2008) [24].

The life-cycle of HIV is considered to start at the point of cell entry (Figure 1.8). The mature HIV particle, having budded out of the previous cell, is contained within a cell-derived envelope expressing the viral glycoproteins gp120 and gp41. The external gp120 recognises and binds to the cellular receptor CD4, which is present on the surface of immunological cells; primarily T-helper cells, monocytes, macrophages and dendritic cells. HIV's tropism is therefore limited to these particular cells. Once gp120 has interacted with CD4, it undergoes a conformational change allowing it to bind to chemokine co-receptors CCR5 or CXCR4, ultimately resulting in gp41-mediated fusion of the envelope surrounding the virion and the plasma membrane, allowing the virus to enter the cell [47]. Once in the cytoplasm, the reverse transcriptase enzyme present in the capsid begins to convert the RNA genome into a linear dsDNA intermediate through a complex series of steps. Although two copies of the genome are present in the particle, it is thought that this is only to increase the chance of successful DNA synthesis through a process termed 'copy choice'; if the reverse transcriptase encounters a break in the genome template it can switch to the second RNA template to continue synthesising the dsDNA intermediate [62]. If the two RNA genome strands are genetically distinct, then this copy-choice mechanism can result in genetic recombination of

the two genotypes into a new genotype. This can accelerate the appearance of new complex variants of the virus, which is important for generating resistant strains in response to anti-retroviral treatments [62, 66]. Once DNA synthesis is complete, the

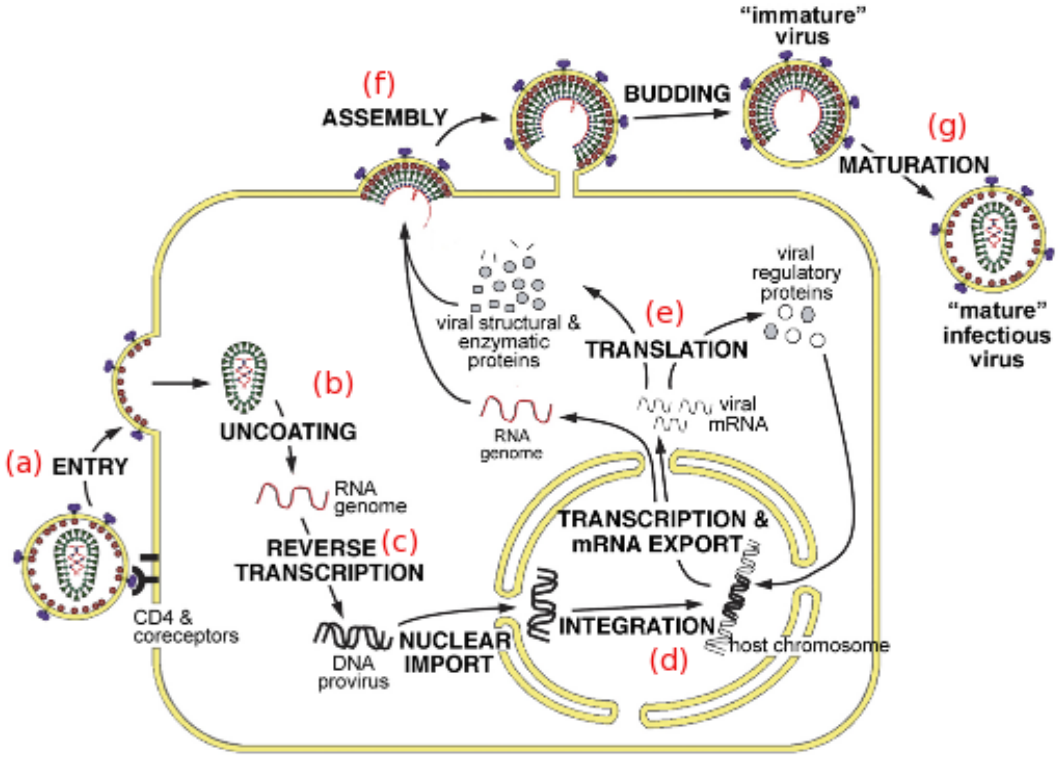


Figure 1.8: Cartoon showing the life-cycle of an HIV particle. The cycle starts at cell-membrane binding and entry (a). Following entry, the RNA genome is released into the cytoplasm (b) where it is reverse transcribed into its double-stranded DNA intermediary (c). This is then transported to the nucleus where it integrates with the host genome (d). This 'provirus' is then transcribed and translated by host cell machinery to produce polyprotein products and the RNA genome (e), which are transported to the cell membrane where they assemble and bud from the cell (f). Protease then cleaves the polyproteins to form a mature infectious virion (g). Figure adapted from Ganser-Pornillos *et al.* (2008) [24].

viral enzyme integrase cleaves the 3′ ends of the DNA genome, leaving exposed hydroxyl groups. The linear DNA, complexed with other viral proteins as a nucleoprotein complex is actively transported through the nuclear membrane, and encounters the host DNA in the nucleus. Catalysed by integrase, the 3′ hydroxyl groups attack the phosphodiester bonds on the target DNA and form new bonds between host and viral DNA, integrating it into the genome. Once the virus has integrated its dsDNA into the

host's genome, it is known as a provirus. Once integration has occurred, transcription of the proviral DNA can occur. HIV uses the cellular transcription factors to transcribe a small amount of full-length viral RNA, which is then spliced to form the mRNA for Tat and Rev. These accessory proteins play an important role in viral gene expression regulation; Tat is a gene-specific elongation factor which, at the start of transcription, binds the viral mRNA at a stable stem-loop region termed the 'transacting-responsive element' (TAR). The binding of Tat to this TAR increases the rate of transcriptional elongation by over 30,000 [18]. This results in the accumulation of Rev, which acts as a 'shuttling' protein, exporting the unspliced HIV transcripts into the cytoplasm. The various different viral transcripts are then translated in different organelles, depending on the product; the Env polyprotein (gp160) is synthesised in the endoplasmic reticulum, while the Gag and Gag-Pol polyproteins are synthesised on cytoplasmic ribosomes. Gag and Gag-Pol are then associated with two viral genomic RNA molecules and transported to the plasma membrane, where the Env polyprotein has localised. Budding occurs at the plasma membrane, generating an immature virion. Either during or immediately after budding, protease cleaves the Gag and Gag-Pol polyproteins to form a mature, infectious virion [47].

However, the genotype of the progeny virions are not necessarily identical to the genotype of the particle that infected the cell. The virally-encoded reverse transcriptase that HIV uses to convert its single-stranded RNA genome into a double-stranded DNA intermediary has no $3' - 5'$ exonuclease activity. Therefore, unlike the host cell's DNA polymerase, if a mismatched base is added to the nucleic acid chain during reverse transcription, reverse transcriptase is unable to excise the mismatched base and replace it with the correct one [28]. The fidelity of reverse-transcription has been calculated to be very poor; having an error rate of approximately 1 in 2000-7000 nucleotides [90, 40, 84]. As the HIV genome is approximately 9.8kb in size, this means that the dsDNA intermediary will have approximately 3 mutations on average. Therefore the progeny virions will have 3 mutations in their genomes compared to the virion that infected the cell. These mutations can be ubiquitously spread through out the viral genome; if a mutation results in non-infectious progeny then the mutation will be lost, but as the average viral life cycle lasts just 1-2 days and up to 10 billion virions produced daily [31], this

'shotgun' method will produce many genetically-distinct infectious virions.

As mentioned earlier, the rapid evolution of HIV within an individual is accelerated by a process termed 'copy-choice recombination'. This is a non-trivial method of producing novel genotypes, with at least 10% of circulating strains being generated by recombination between different subtypes. If a cell is co-infected with two genetically distinct virions, the progeny viruses may be heterozygous (containing an RNA strand from each distinct genotype). Once this heterozygous virion infects a new cell, when the reverse transcriptase swaps between genomes during dsDNA synthesis, a novel genotype can be generated that recombines the two [63].

These two evolutionary methods employed by HIV result in extremely variable HIV genotypes within an individual, within a population, and between populations. Within an individual these methods of novel genotype generation result in a circulating population of 'quasispecies'. In the constant selective pressure environment within an individual, the quasispecies that are fittest will outcompete the less fit, but due to the constant generation of new genetic variants there is always a low-level of these minority quasispecies. These quasispecies become important in the context of a change in selective pressure through administration of anti-retroviral drugs. These drugs may suppress the majority species to a low level of replication, but due to the high mutation and recombination rate, a quasispecies may be generated that is less sensitive to the drug and so has a higher replication rate than the other quasispecies. Eventually a genotype is generated whose replicative capacity is close to the pre-drug genotype in the presence of the drug. Viral loads increase, $CD4^+$ numbers decrease again, and so the drug's effect has been circumvented [16]. For this reason antiretroviral drugs are given as a cocktail called 'highly active anti-retroviral therapy' (HAART). HAART consists of 3 or 4 drugs inhibiting different stages of the viral life cycle taken in combination. For example, a protease inhibitor to prevent viral maturation is taken alongside a non-nucleoside reverse transcriptase inhibitor, which binds to RT preventing it from performing its function. In addition, a nucleoside or nucleotide reverse transcriptase inhibitor may be taken, which also prevents RT from performing its function.

At the population level, transmission of mutant genotypes has resulted in genotypes that are sufficiently distinct to be labelled as different 'subtypes'. There are 9 different subtypes, labelled A to K, excluding E and I which were thought to exist but were later found to be 'circulating recombinant forms' (CRFs). These subtypes are predominant in different regions of the world; for example, subtype B is prevalent in Europe and North and South America, while subtype C is predominant in Africa and Asia. However, despite these genetic differences, it has been shown that all subtypes show comparable viral load suppression upon administration of HAART [25]. Therefore despite the genetic differences between subtypes, which typically differ by 25% to 30% of their genes' nucleic acids [104], the mutations that confer resistance to HAART inhibitors are not found naturally.

1.2.3 HIV Pathogenesis

Upon primary infection with HIV, the virus infects $CD4^+$ cells in the blood, such as T_H lymphocytes, macrophages and dendritic cells. The infected cells then carry the virus to the lymph nodes and spleen where the residing activated-T_H lymphocytes are infected. The favourable conditions for HIV replication in these lymphoid tissues result in a considerable expansion of the HIV population; up to 10^{10} progeny virions are produced every day. During this 'acute' phase there is a transient depletion of peripheral $CD4^+$ cells and an associated high blood-plasma load of HIV (Figure 1.9). As a result, an 'acute-phase response' is launched by the host's immune system, where $CD8^+$ cytotoxic T lymphocytes (CTL) specific to HIV antigens are clonally-expanded, and neutralising antibodies are produced. This humoral response results in circulating viral loads dropping to undetectable levels, but the infected T_H lymphocytes contain the HIV provirus integrated into their genome and so act as a reservoir, producing virions. Eventually the T_H lymphocytes are destroyed by the virus or by CTLs that recognise them as infected. Slowly the levels of $CD4^+$ cells are depleted, which severely immuno-compromises the host. Once the levels of $CD4^+$ cells drop below a threshold level, symptoms due to opportunistic infections such as Kaposi's Sarcoma Herpesvirus are presented in the host. This is referred to as the onset of AIDS, and rapidly leads to death [85, 91].

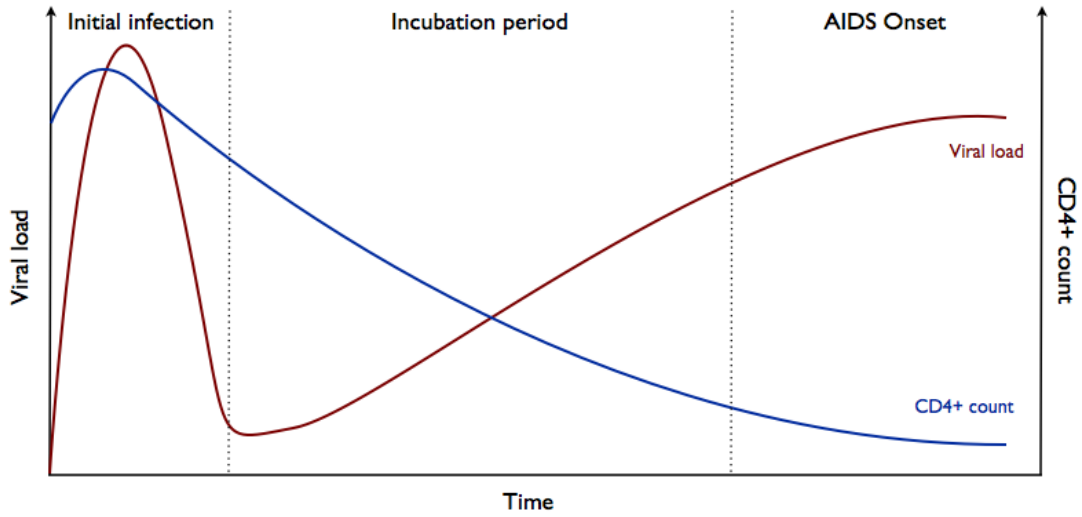


Figure 1.9: Course on an HIV infection described with respect to the circulating viral levels and numbers of $CD4^+$ cells. After initial infection by HIV and the subsequent acute-phase response by the host, there is a long incubation period that can last many years before $CD4^+$ cells drop below a threshold level and are no longer able to suppress opportunistic infections. The presentation of these infections marks the progression to AIDS. Figure adapted from Purves *et al.* (2001).

The length of the incubation period is variable, depending on factors such as host genetics, viral genetics, and environmental factors such as access to anti-retroviral drugs (ARDs). Administration of ARDs during the incubation period can reduce the viral loads to undetectable levels with an associated increase in $CD4^+$ levels. The inhibitory effect of the ARDs is only transient; due to high genetic variability of HIV, a population arises that has a reduced susceptibility to the drugs and consequentially viral loads increase and $CD4^+$ levels decrease. The length of time taken for the virus to escape the inhibitory effect is also variable, depending on factors such as host genetics, viral genetics and drug adherence. As previously mentioned, the probability of a viral population emerging with resistance to multiple drug that inhibit various stages of its life cycle is much smaller than if a single drug is administered. Therefore currently HAART prolongs the incubation period by co-administering 3 different inhibitors - one that inhibits the HIV protease and two that inhibit reverse-transciptase.

1.2.4 HIV-1 Protease Structure and Function

The protease enzyme of HIV-1 is a homodimeric aspartyl protease, composed of two identical 99-amino acid 'C'-shaped monomers that come together to form the functional protein (Figure 1.10) [13]. Stabilisation of the dimer is achieved almost entirely through

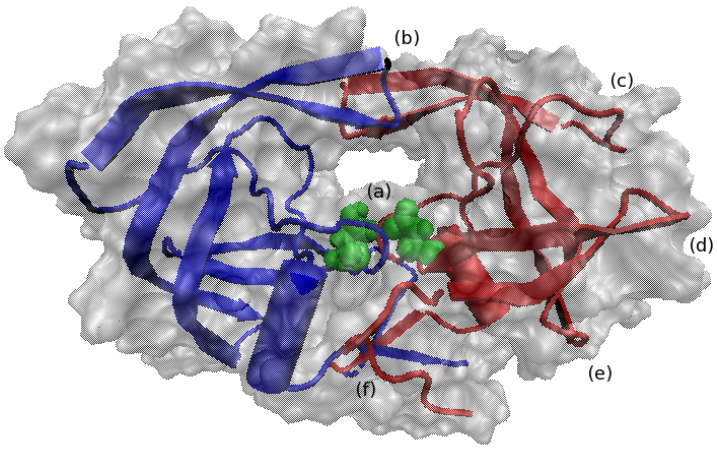


Figure 1.10: Cartoon of the aspartyl protease of HIV, with the backbones of two identical monomers coloured red and blue. The cartoon representation of the protein backbone, showing the secondary and tertiary structures, is outlined by the surface area of the protein, showing how the two subunits come together to form the active site in the middle of the protein (a). The two catalytic aspartic acid residues are shown as green *van der Waals* spheres. (b) denotes the flaps that cover the active site, (c) denotes the flap elbows, (d) denotes the fulcrum, (e) denotes the cantilever, and (f) denotes the dimerisation interface. Figure was created from 1HXB PDB file using VMD.

the interdigitation of the four N-terminal and four C-terminal residues' β-strands in each monomer. Due to the highly conserved nature of the sequence in this region, interface inhibitors have been designed to block formation of the functional homodimer by mimicking the C- or N-terminal sequence [23]. Dissociation of the dimer into component monomers results in complete loss of enzymatic activity [56]. The active site of the protease is formed at the dimer interface with each monomer providing a catalytic Asp-Thr-Gly triad at positions 25 to 27 respectively. Of this catalytic triad, the aspartic acid residues of each monomer interact directly with the substrates to catalyse

the reactions [56]. The roof of the active site is composed of a pair of flexible, overlapping 'flaps', each composed of two anti-parallel beta sheets connected by a beta turn. The remaining topological features of HIV-1 protease's structure are named according to their hypothesised role in the flap opening mechanism; the fulcrum, cantilever and flap elbows undergo a concerted downwards motion that results in the flaps opening upwards and the catalytic aspartic acid residues shifting up further into the active site (see Figure 1.11) [33, 113]. The flaps are therefore proposed to act as a gating mechanism for controlling access to the active site [113].

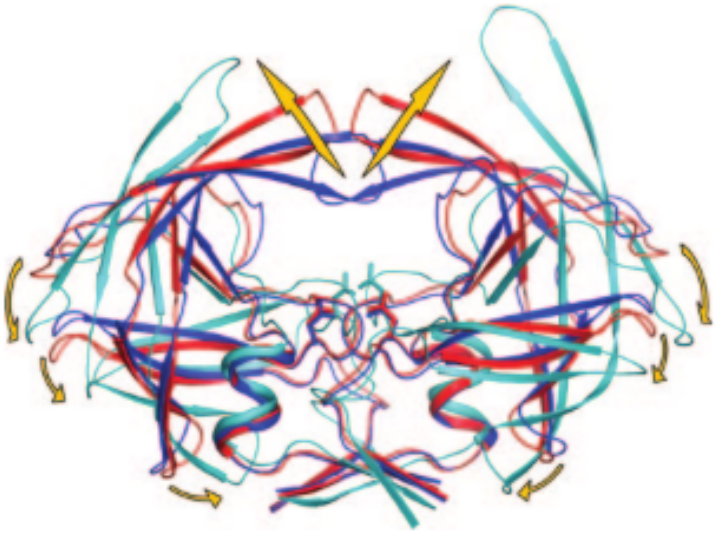


Figure 1.11: Proposed mechanism for flap opening by Hornak *et al.* (2006). The flap elbows, cantilevers, and fulcrums undergo a concerted downwards motion that results in an upwards motion of the flaps and the catalytic aspartic acid residues. This facilitates entry to the active site [33]. Image taken from Hornak *et al.* (2006).

Unlike the majority of the other HIV proteins, protease functions in the HIV particle itself, rather than in the cytoplasm or nucleoplasm of the host cell. As previously described, when all the components of the progeny particle accumulate at the cell membrane of the host cell, the particle buds out of the cell, taking some of the cell membrane with it as an envelope. However, once budding has occurred the particle is still not infectious, and is considered an immature virus. This is because the necessary viral proteins are still joined as Gag and Gag-Pol polyprotein precursors from the polycistronic translation, and require cleavage by HIV protease before they can re-

configure to produce a mature HIV virion [47]. There are 12 protease cleavage sites in the viral polyproteins: 5 in Gag (p17/p24, p24/p2, p2/NC, p7/p1 and p1/p6^{gag}), 6 in Gag-Pol (NC/TFP, TFP/p6^{pol}, p6^{pol}/PR, PR/RT, RT/p66, and p66/IN) and 1 in Nef (see Figure 1.12) [72]. An important note is that protease itself is translated as part of a polyprotein, so therefore to function as a dimer, it must cleave itself out. The mechanism by which this happens is by dimerisation of the polyprotein and subsequently the protease cleaves itself out by a *trans*-mechanism [51]. Following excision however, the protease does not cleave the polyproteins in a random fashion, nor is each site cleaved at the same rate. The binding cleft that runs through the protease and contains the catalytic residues is able to complement 7 amino acids. This coincides with the specificity of the protease, which is determined by the nature of the 4 amino acids upstream and 3 amino acids downstream of the scissile bond in the substrate. The amino acid immediately upstream of the bond must be hydrophobic and unbranched at the C_β atom. The surrounding substrate residues are variable between cleavage sites, which suggests why different cleavage sites have different rates [51]. Therefore, 3 related determinants are thought to control the order of the proteolytic processing: the sequence of the cleavage site, the structure around the cleavage site, and the accessibility of the site to the protease [79].

The mechanism by which HIV protease catalyses the hydrolysis of the peptide bonds at the cleavage sites is unknown, but is thought to occur via a general-acid/general-base (GA/GB) mechanism [80, 114, 71]. This mechanism requires one of the catalytic aspartic acids to be protonated and the other un-protonated; upon formation of the enzyme-substrate complex (ES), the un-protonated aspartic acid polarises a water molecule, making the oxygen atom electronegative (**TS1** in Figure 1.13). The protonated aspartic acid then donates its proton to the substrate's oxygen atom, making the substrate's C_α atom unstable and therefore open to nucleophilic attack by the hydroxyl group of the water molecule **(INT)**. The water's remaining proton is donated to the aspartic acid, temporarily swapping the protonation states of the two catalytic residues. The peptide bond of the substrate is then cleaved as the protonated aspartic acid donates its proton to the nitrogen atom, and the proton on the substrate's carboxylic acid group passes back to the other aspartic acid **(EP)**. The products are released and a

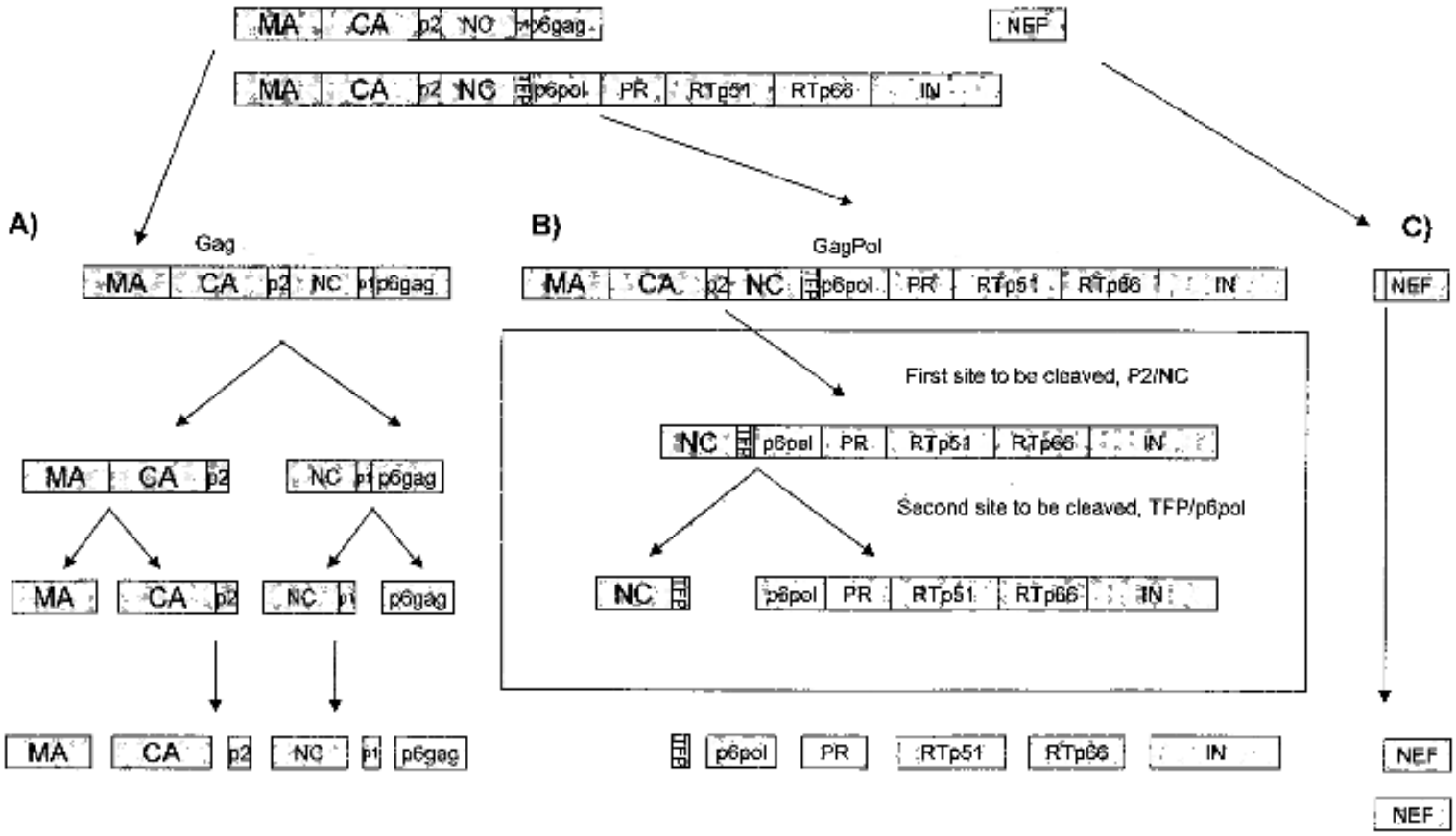


Figure 1.12: Location and order of the protease cleavage sites in the HIV polyproteins. **A)** shows the order of cleavage for the Gag polyprotein, **B)** shows the order of cleavage for the Gag-Pol polyprotein, and **C)** shows the cleavage for Nef. 'TFP' in the Gag-Pol polyprotein refers to a trans-frame peptide generated by ribosomal frameshifting required to synthesise Gag-Pol. Image adapted from de Oliveira *et al.* (2003).

new water and substrate molecule must enter the active site for catalysis to occur again **(E)** [80, 71].

1.2.5 HIV-1 Protease Inhibitors

HIV-1 protease's importance in the viral life cycle, combined with its unique recognition motif not shared by host cell proteases, make it an ideal target for inhibition to disrupt the viral life cycle. However, as HIV was a previously-unknown emergent pathogen, new pharmaceutical agents needed to be created. Protease was therefore identified as a prime-target for structure-assisted drug design. This approach utilises techniques such as protein crystallography, NMR, computational molecular design, and combinatorial chemistry to guide the synthesis of inhibitory compounds. The first drug to be developed using this rationale was saquinavir, which was developed on the concept that HIV protease uniquely cleaves Tyr-Pro or Phe-Pro dipeptide sequences, unlike its mammalian counterparts. Transition-state mimics were developed with the scissile bond replaced with an un-cleavable hydroxyethylmine moiety. The hydroxyl group of this moiety interacts with the two catalytic aspartic acid residues in the protease, significantly contributing to the inhibitor's potency [81]. The proline residue attached to the scissile bond was also replaced with (S,S,S)-decahydro-isoquinoline-3-carbonyl which was able to maintain important hydrogen bonds between water molecules that connected the inhibitor to the flaps residues. The resultant inhibitor was shown to be highly specific for the viral protease; causing only minor inhibition of human aspartyl proteases, with a K_i of over 10,000nM for renin and pepsin[129]. The success of saquinavir lead to the rapid development of new protease inhibitors (PIs) that improved upon the design of saquinavir [126, 127]. There are currently 10 different approved protease inhibitors for use in HAART: amprenavir (not currently administered as it was replaced by fosamprenavir), atazanavir, darunavir, fosamprenavir, indinavir, lopinavir, nelfinavir, ritonavir, saquinavir and tipranavir [116]. The structures of these inhibitors share many features (Figure 1.14) - they all contain the same un-cleavable hydroxyl group that makes the important contacts with the catalytic aspartic acid residues. This makes a large contribution to the binding affinity while its unreactivity ensures that it remains in the active site so that the protease is unable to fulfil its role. They also all contain many hydrophobic aliphatic and aromatic groups. These interact with the

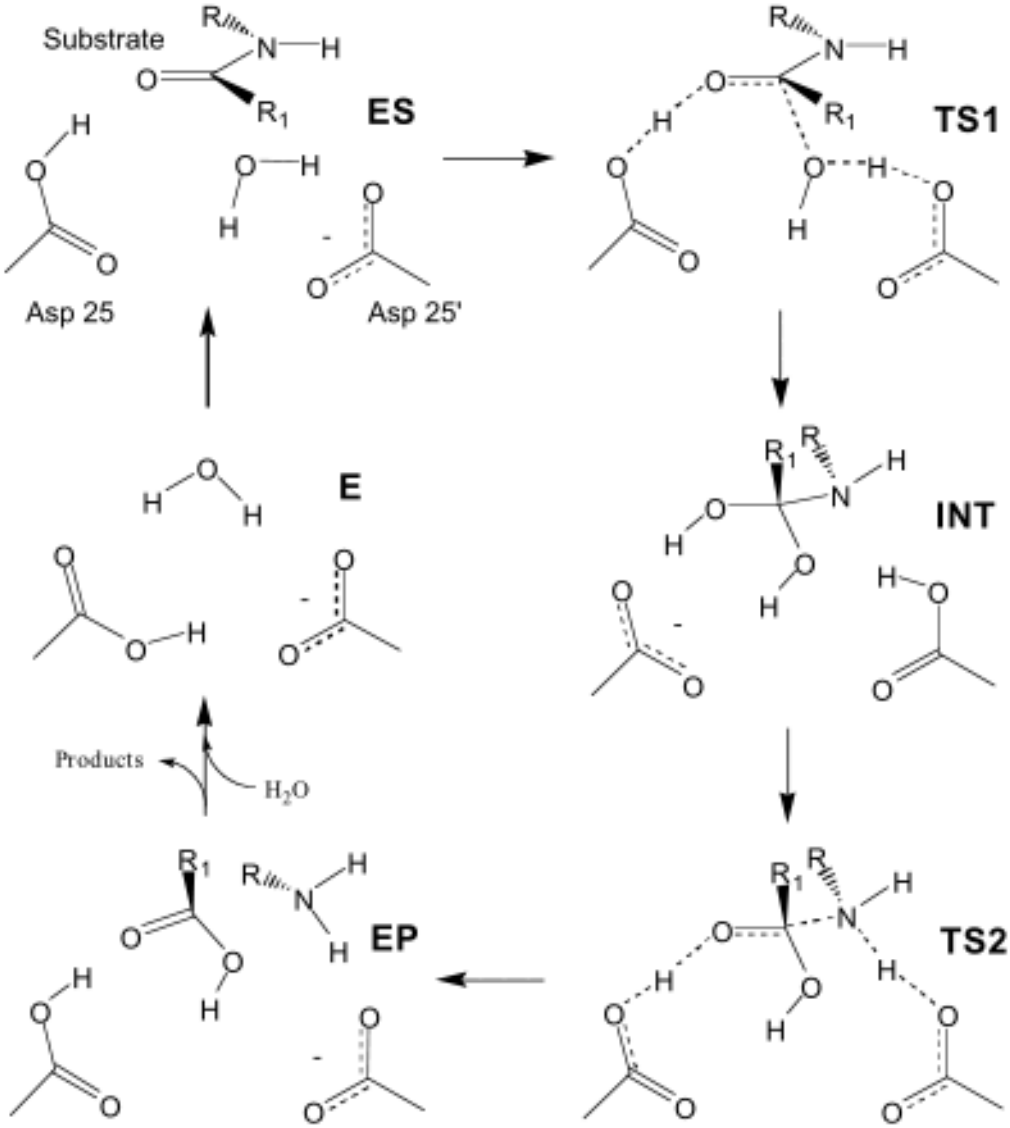


Figure 1.13: Proposed GA/GB mechanism by which HIV protease catalyses the hydrolysis of its substrate's peptide bond. (**ES**) Formation of the enzyme-substrate complex; (**TS1**) Nucleophilic attack of the substrate by a water molecule which has been ionised by the unprotonated aspartic acid residue (**TS2**) Covalent linkage of the hydroxy anion to the peptide carbon results in breakage of the peptide bond, donation of the aspartic acid's proton to the amino terminus of the product, and donation of the product's carboxylic terminus' proton back to originally-protonated aspartic acid (**EP**). The products are released and the cycle can restart (**E**). Figure taken from Piana *et al.* (2002).

mainly lipophilic residues in the protease's binding groove, increasing the change in entropy of complex formation [81].

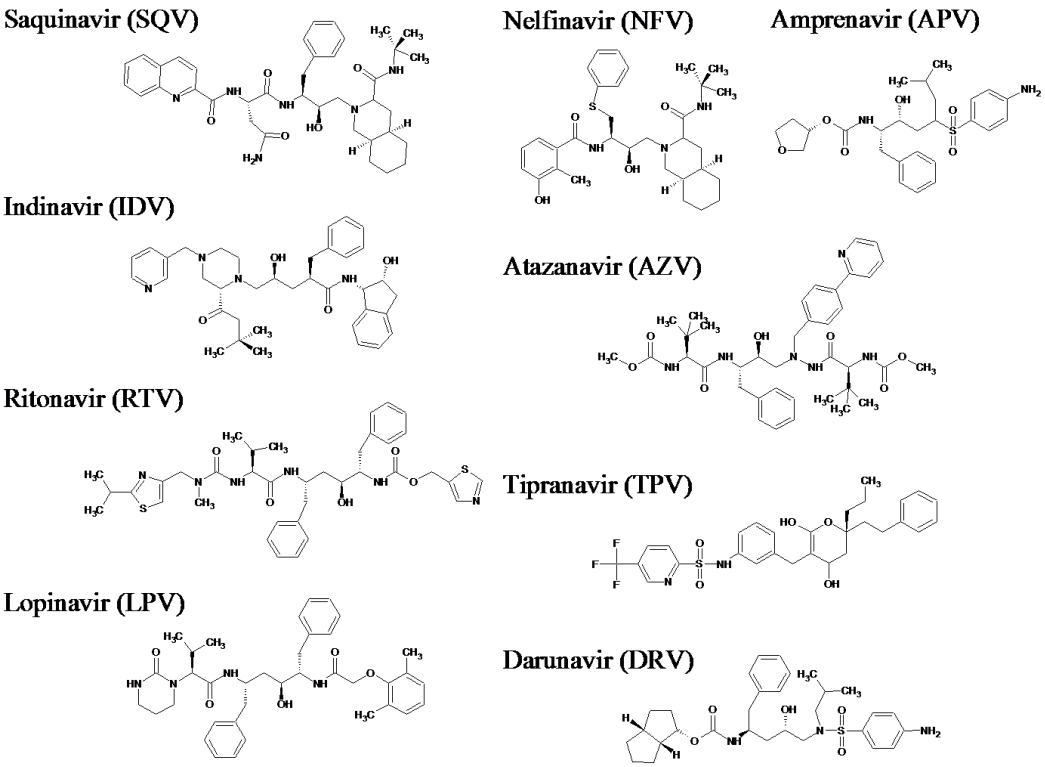


Figure 1.14: Chemical structures of the 9 currently-administered protease inhibitors. A common feature of all inhibitors is the hydroxyl group near the center of each molecule. This makes an important interaction with the catalytic aspartic acids while remaining un-cleavable. Also of note are the many aliphatic and aromatic groups which interact with the lipophilic residues of protease's binding groove, helping stabilise the complex [81]. Figure taken from Sadiq (2009).

By reversibly binding to the active site, protease inhibitors sequester the action of HIV protease by preventing the polyproteins from entering. This means that they are not hydrolysed into the function proteins, and therefore the progeny virions remain 'immature' and non-infectious. The inhibitors are not 100% effective though, and so a low level of viral replication still occurs under the drug presence. This low-level replication combined with the error-prone nature of its reverse transcriptase means that genetically-diverse progeny are produced. While many of the mutations will be poly-

morphisms common to the HIV genome and have no effect on the viral fitness, some will confer a benefit to the virus - making it less susceptible to the inhibitory effect of drug. As the drug places a strong selective pressure on the virus, any sub-populations in an individual with an increased fitness in the presence of the drug will be selected for and become the majority population. As it has a reduced susceptibility to the drug, it will produce a greater number of mature progeny, resulting in a greater chance of an even fitter population emerging that completely circumvents the action of the inhibitor. Under protease-inhibitor monotherapy, this evolution of a drug-resistant population is common. The response used to be to change the administered inhibitor to one with a greater inhibitory effect on that population, but quickly a population would emerge that was resistant to the new drug. Furthermore, multi-drug resistant (MDR) populations would arise with mutations that conferred a measure of resistance to more than one drug. For this reason HAART was developed, as the chance of a mutant population arising that circumvents the effects of all the drugs is much lower, and therefore the time taken for a resistant population to emerge is much longer.

Mutations of HIV-1 protease are grouped into two classifications: primary and secondary (or accessory) mutations (Figure 1.15). Primary mutations confer reduced susceptibility to one or more protease inhibitors by itself. They are commonly situated around the substrate-cleft and are not polymorphisms seen naturally in viral populations of drug-naïve individuals, due to selective disadvantage compared to wild-type without the drug present [44, 89, 105]. Examples of primary mutations are I50L, V32I, I47V, G48V, V82F, I84V and L90M. These are not all necessarily located in the substrate binding-cleft; for example, L90M is situated at the dimerization interface at the base of the protein. The mutations that do occur at the active site have been shown to directly alter the interactions between the protease and inhibitor. For example, the G48V mutation alone is sufficient to cause reduced susceptibility to saquinavir due to a steric conflict between the substituted valine and the P2 subsite of saquinavir and a disruption of the hydrogen bonding between the residue and the P2 subsite [124]. However, the mechanism of action of the mutations located outside of the active site is unknown [9, 67]. Therefore there is considerable interest in elucidating the action of these mutations, and computational methodologies such as molecular dynamics sim-

ulations are providing hitherto unforeseen insights. For example, molecular dynamics simulations of the V82F/I84V MDR mutant suggest that these mutations change the dynamics of the protease so that the flaps are more often in a semi-open configuration rather than a closed-configuration (see Figure 1.11). Therefore the inhibitor has a greater enthalpic penalty in closing the flaps upon complex formation [77]. This also helps to explain how these mutations cause resistance to more than one inhibitor, because a change in protease dynamics is likely to have the same dissociative effect for multiple inhibitors. Another example is that of the L90M mutation, which confers resistance to nelfinavir and saquinavir on its own. Simulations of a protease containing this mutation bound to saquinavir, lopinavir and nelfinavir suggested that the mutant side-chain changes an interaction with the backbone of residue 25, which subsequently becomes slightly dislocated, resulting in a rotation of reside 84's side-chain. This causes the displacement of the inhibitor from its binding position [67].

Secondary mutations are regularly natural polymorphisms present in the various quasi-species of a viral population. Therefore, unlike primary mutations, these mutations are not deleterious to viral replication in the absence of the inhibitor. The appearance of primary mutations in a drug-suppressed HIV population is commonly followed by the acquisition of one or more secondary mutations, which either act to compensate any deleterious effects of the primary mutation on the protease, or result in an additive reduction in sensitivity to the protease inhibitor, or create a favourable context for the emergence of a specific resistance mutation at a different position [32]. For example, the L63P is a natural polymorphism in HIV, occurring in $> 50\%$ of drug-naïve viral populations [108]. However, in the context of V82F and I84V, this mutation improves the replicative capacity of the virus. Furthermore, in the context of M46L, this polymorphism improves the protease catalytic efficacy without affecting the inhibitor. The mechanism by which these secondary mutations perform their actions is unknown [58]. Examples of secondary mutations are M36I, I54V, L10I and K20M.

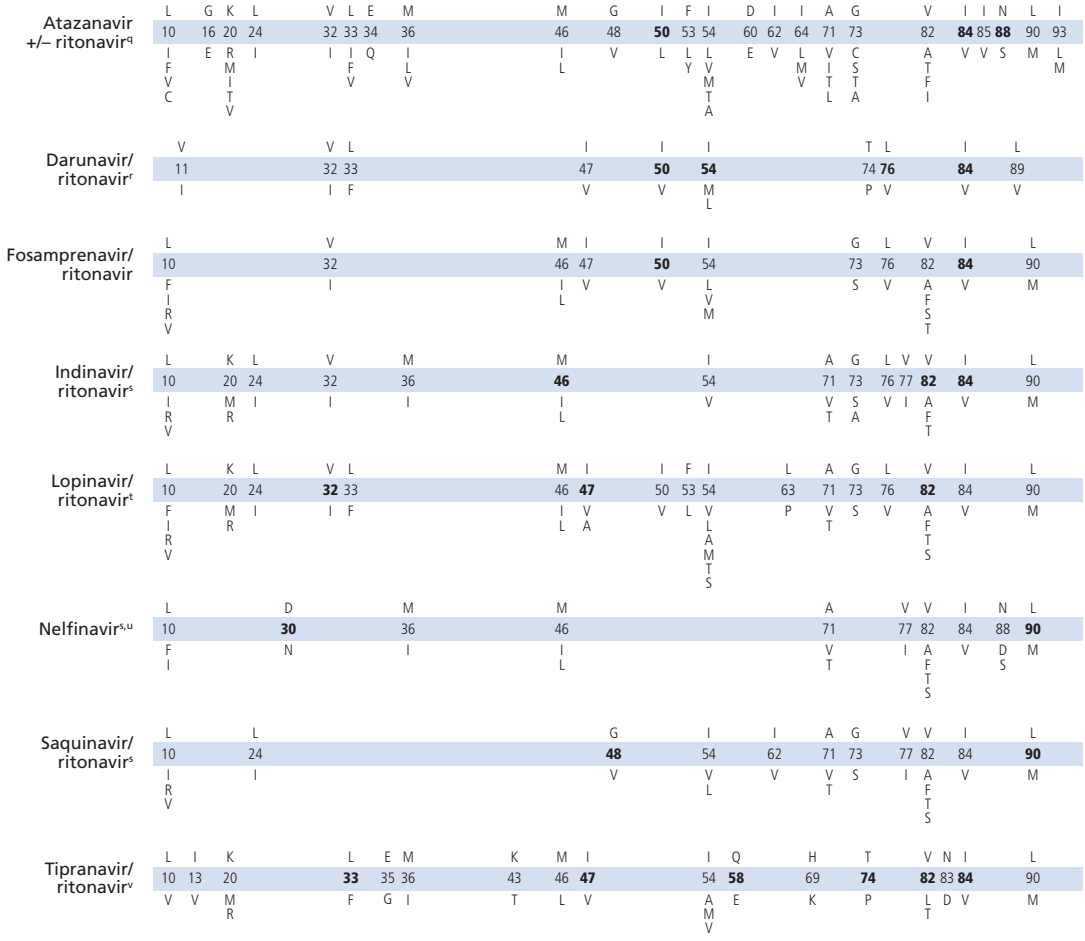


Figure 1.15: Common drug resistance mutations in the protease gene for the associated inhibitors. In bold are primary resistance positions for that inhibitor. Figure taken from Bennett *et al.* (2009).

1.3 Biochemical Introduction

1.3.1 Thermodynamics

Everything in the universe has to obey the laws of physics and chemistry, and biological organisms and systems are no different. Therefore all enzyme-catalysed reactions must adhere to the three laws of thermodynamics:

1. **The total energy of a system and its surroundings is always constant**. In thermodynamics, the *system* is the matter contained within a defined region-of-interest, and the *surroundings* is the rest of the matter in the universe outside of this system. This first law states that regardless of the energy conversions which occur during a reaction, the sum of the various types of energies stays the same. Equation 1.3 shows this mathematically:

$$\begin{aligned} \Delta E &= E_B - E_A \\ &= Q - W \end{aligned} \tag{1.3}$$

where E is energy; E_B is the energy of a system at the end of the reaction; E_A is the energy of a system at the start of the reaction; Q is the heat energy absorbed by, or given to, the system during the reaction; and W is the work done by, or done to, the system during the reaction. Due to the fact that energy is always conserved, the change in energy of a system is equal to the sum of the heat gained or lost by the system and the work done by or done to the system. It is important to note that the change in energy of a reaction is only dependent on the energies of the initial and final states, and not of any of the intermediate states taken to reach this final state.

When energy is released from a chemical bond during a reaction, it is passed to surrounding molecules in the system as an increase in translational, vibrational and rotational energy; they have an increased thermal motion, which is analogous to raising their temperature. As these molecules have increased thermal motion, they will collide with their surrounding molecules, transferring thermal energy to these molecules. Eventually this thermal energy is passed out of the system and

into its surroundings. In the end, all the thermal energy passes out of the system, which returns to its original temperature. Therefore all the energy stored in the chemical bond was converted into thermal energy and then transferred out of the system and into its surroundings. According to the first law of thermodynamics, the change in energy of the system must be equal and opposite to the amount of thermal energy transferred. This is shown in Equation 1.4:

$$\Delta E = -h \tag{1.4}$$

Where h is the thermal energy transferred. The negative sign indicates that the energy is transferred out of the system into its surroundings. However, the increase in thermal energy during the course of a reaction may cause an increase in the volume of the system. In this situation, in order to expand, the system must do work to push against the pressure of its surrounding matter. The energy used to do this work is equal to the pressure of the surroundings multiplied by the change in volume of system ($P\Delta V$). According to Equation 1.3, the energy used to do this work must decrease the energy of the system. Therefore an amended function must be used that describes energy changes that occur when both temperature and pressure or volume are altered. This amended function is called enthalpy, and its mathematical equation is shown as Equation 1.5.

$$H = E + PV \tag{1.5}$$

where H is the enthalpy of a system; E is the thermal energy of the system; P is the pressure of the surroundings; and V is the volume of the system.

It is therefore the change in enthalpy of a system, rather than the change in energy of a system, that is equal to thermal energy transferred to its surroundings. During a reaction, if ΔH decreases then there is a net release of thermal energy to the surroundings, and the reaction is said to be exothermic. If ΔH increases then the system had to absorb thermal energy from its surroundings and is said to be endothermic. To be exact, Equation 1.4 should be amended to

Equation 1.6:

$$\Delta H = -h \tag{1.6}$$

In the context of biological processes, however, the volume change caused by a reaction is so negligible that the PΔV term can be ignored. In such cases, the enthalpic change in a reaction is equal to the energy released from the chemical bond [1]:

$$-h = \Delta H \cong \Delta E \tag{1.7}$$

2. **A reaction can only occur spontaneously if it results in a net increase of entropy**. Entropy is a measure of the disorder of a system; when a system becomes more disordered, its entropy increases. The change in entropy of a reaction where one mole of compound A is converted into one mole of compound B is given by:

$$\Delta \mathrm{S} = \mathrm{R} \ln \frac{\mathrm{p_B}}{\mathrm{p_A}} \tag{1.8}$$

where ΔS is the change in entropy of the reaction; R is the gas constant; p_A is the probability of state A; and p_B is the probability of state B [1]. As can be seen, ΔS is directly proportional to the natural logarithm of $\frac{\mathrm{p_B}}{\mathrm{p_A}}$, so a positive change in entropy results from the probability of state B occurring being higher than the probability of state A occurring ($\frac{\mathrm{P_B}}{\mathrm{P_A}} > 1$).

The entropy of a system can decrease during a spontaneous reaction (where the probability of the products occurring is lower than the probability of the reactants) as long as the subsequent increase in entropy of its surroundings results in a net increase. Therefore the formation of highly ordered structures seen in biological processes, which have a negative entropy, can occur spontaneously due to a larger increase in the entropy of its surroundings caused by the release of heat energy from the reaction. This is written mathematically as:

$$\sum(\Delta \mathrm{S_{system}} + \Delta \mathrm{S_{surroundings}}) > 0 \tag{1.9}$$

3. **As the temperature of a system reaches absolute zero, all reactions**

cease and its entropy reaches a minimum state. This states that as the temperature of a system approaches -273.15°, both ΔS and S itself reach a constant, which in the case of perfect crystalline substances is 0. This law is not of any concern for biochemical reactions of living organisms, as they never naturally approach absolute zero [29].

Of these three laws, the second law is of considerable importance in biochemistry, as it states whether a process will occur spontaneously. However, the problem with using entropy as an indicator of whether a biochemical reaction will occur spontaneously is that it is not easily measurable, especially as it requires knowledge of $\Delta S_{surroundings}$, which, in the case of a cellular reaction, is everything in the universe outside of the cell [107]. For this reason a composite thermodynamic function called *Gibb's free energy* was defined which combines equations from the first and second laws of thermodynamics without requiring consideration of any thermodynamic properties of the surroundings. Gibb's free energy equation is shown in Equation 1.10.

$$G = H - TS \tag{1.10}$$

$$\Rightarrow \Delta G = \Delta H - T\Delta S \tag{1.11}$$

where G is Gibb's free energy; H is the enthaplic function described in Equation 1.5; T is the temperature; and S is the entropic function described in Equation 1.8.

Equation 1.11 shows that the change in Gibb's free energy (ΔG) during a reaction in a system of volume (V) at a constant temperature (T) and pressure (P) is equal to the change in enthalpy of the system (ΔH) during the reaction minus the product of the temperature and the change in entropy of the system (ΔS) during the reaction. The change in the Gibb's free energy value is an important criterion as to whether a reaction will occur spontaneously, and as can be seen from Equation 1.11, requires no knowledge of any thermodynamic functions of the system's surroundings. This is because the surrounding matter's entropic information is contained within the Gibb's

free energy equation, as shown in Equation 1.13.

$$\begin{aligned} -\Delta G &= -\Delta H + T\Delta S \\ -\frac{\Delta G}{T} &= \frac{\Delta H}{T} + \Delta S_{\text{system}} \qquad (1.12) \\ \Rightarrow -\frac{\Delta G}{T} &= \Delta S_{\text{surroundings}} + \Delta S_{\text{system}} \qquad (1.13) \end{aligned}$$

By taking a rearrangement of Equation 1.11 and dividing through by the temperature, Equation 1.12 is attained. Of particular note is the term $\frac{\Delta \text{H}}{\text{T}}$, which is the change in enthalpy of the system during a reaction divided by the temperature. As shown in Equation 1.6, in biological situations the change in enthalpy during a reaction (ΔH) is equal to the amount of thermal energy transferred to the surroundings (-h). This addition of thermal energy in the system's surroundings increases the number of different arrangements that the surrounding molecules can adopt, and therefore increases its entropy. This transfer of thermal energy has a greater disordering effect at lower temperatures than at high temperatures, and thus the change in entropy of the surrounding matter during a reaction is equal to the amount of heat transferred from the system divided by the temperature ($\frac{\text{h}}{\text{T}}$). Therefore the change in Gibb's free energy during a reaction is a direct measure of the entropy change of the universe [1], as shown in Equation 1.13. As mentioned previously in this chapter, a reaction will only occur spontaneously if it causes a net increase in entropy in the universe. Applying this to Equation 1.13 shows that for the change in entropy of the universe to increase, the Gibb's free energy change must become more negative.

- A positive ΔG means that the reaction cannot occur spontaneously as it causes a net decrease in the universe's entropy, and therefore requires an input of free energy in order for the reaction to occur.
- A ΔG value of zero means that the system is at equilibrium as the reaction proceeds at equal rates in the forward and reverse directions.
- A negative ΔG means that the reaction causes a net positive entropy change in the universe and so the reaction occurs spontaneously.

As with the change in enthalpy, the change in Gibb's free energy of a reaction is only dependent on the free energy of the products and the free energy of the reactants. The free energy of the transitional products synthesised during this transformation has no bearing on whether the reaction will occur spontaneously. Also, the ΔG of a reaction gives no indication of its rate. A highly-negative ΔG simply indicates that the reaction results in a greater entropic disorder in the universe, but does not therefore mean that it occurs at a faster rate. The rate is governed by other thermodynamic properties, which are discussed in more detail in Section 1.3.3 [1, 107, 85].

1.3.2 Reaction Kinetics

As discussed in Section 1.3.1, a reaction will proceed in the direction that has an associated negative change in ΔG. However, when considering a reversible reaction, such as that shown in Figure 1.14, even if the forward reaction ($\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} + \mathrm{D}$) has an associated negative ΔG, the reverse reaction ($\mathrm{C} + \mathrm{D} \rightarrow \mathrm{A} + \mathrm{B}$) will still occur, though at a marked lower rate.

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C} + \mathrm{D} \tag{1.14}$$

If the concentrations of C and D greatly exceed those of A and B, it is possible for the overall reaction to proceed in the reverse direction. This is because in such situations there are more molecules undergoing the reverse reaction than there are the forward one, even though it occurs at a lower rate. This highlights that the ΔG of the reaction depends not only on chemical potential energy of the molecules, but also on their concentrations. Equation 1.15 reflects this amendment to the definition of Gibb's free energy.

$$\Delta G = \Delta G^{\circ} + RT\ log_e \frac{[C][D]}{[A][B]} \tag{1.15}$$

where ΔG° is the *standard Gibb's free energy change*, and represents the ΔG when the concentrations of the products and reactants are equal at 1.0M. This is therefore the thermodynamic function calculated in Equation 1.11, and encompasses the intrinsic characteristics of the reacting molecules; R is the gas constant; T is the temperature (in Kelvin); and *[A]*, *[B]*, *[C]*, and *[D]* are the concentrations of the various molecular species involved in the reaction. This half of the equation reflects ΔG's dependance on the molecules' concentrations [1, 107, 12].

As mentioned in Section 1.1.2, enzymes catalyse reactions that occur naturally at a particular rate. In the absence of an enzyme, the two molecular reactants come into close proximity of each other due to random thermal motion, and make weak non-covalent bonds with each other that persist until either the molecules react to form the products, or their thermal motions pull the molecules apart again. The stronger the transient bonds between the reactants, the slower their rate of dissociation. In the situation where molecules A and B bind to form an AB complex, a proportion of the reactants will be forming a complex whilst a proportion of the complexes will be dissociating back into reactants. Each will have a rate of reaction, dependent on the intrinsic properties of the molecules, and on their concentrations. These are shown in Equations 1.16 and 1.17.

$$\text{dissociation rate} = k_{off}[AB] \tag{1.16}$$

$$\text{association rate} = k_{on}[A][B] \tag{1.17}$$

$$K_{eq} = \frac{k_{on}}{k_{off}} = \frac{[AB]}{[A][B]} \tag{1.18}$$

$$K_d = \frac{1}{K_{eq}} = \frac{k_{off}}{k_{on}} \tag{1.19}$$

where k_{on} is the association rate constant; k_{off} is the dissociation rate constant; K_{eq} is the equilibrium constant, also referred to as the association constant; and K_d is the dissociation constant.

A reaction will proceed until it reaches an equilibrium point, at which the rates of association and dissociation are equal. At this equilibrium point, the concentrations of the products and reactants can be used to calculate the equilibrium constant (K_{eq}) of the reaction, which is a reflection of the 'binding affinity' between the two reactant molecules (Equation 1.18). If two molecules have a greater affinity for each other, there will be a greater concentration of the AB complex at equilibrium, and so K_{eq} will subsequently have a higher value; for this reason, K_{eq} is also referred to as the association

constant. The reciprocal of K_{eq} is the dissociation constant, and is occasionally used in place of the association constant (See Equation 1.19). A larger K_d value indicates a greater concentration of reactant molecules A and B at equilibrium, so therefore a lower binding affinity [1].

In Section 1.3.1 it was shown that a reaction is at equilibrium when ΔG = 0. Substituting this into Equation 1.15, the equation for calculating the reaction's equilibrium constant from its free-energy change can be determined.

$$0 = \Delta G^{\circ'} + RTlog_e\frac{[C][D]}{[A][B]} \tag{1.20}$$

$$\Rightarrow \Delta G^{\circ'} = -RTlog_e\frac{[C][D]}{[A][B]} \tag{1.21}$$

$$\Rightarrow \Delta G^{\circ'} = -RTlog_eK'_{eq} \tag{1.22}$$

$$\Rightarrow K'_{eq} = e^{-\Delta G^{\circ'}/RT} \tag{1.23}$$

where $\Delta G^{\circ'}$ is the free-energy change for a biochemical reaction under standard biochemical conditions; the standard state has a pH of 7 with the concentrations of H^+ and H_20 being 1.0M; R and T are the same as in Equation 1.15.

1.3.3 Enzyme Kinetics

As mentioned in Section 1.1.2, enzymes are catalysts that increase the speed at which reactants will react to form products. By ensuring that, for example, two reactants are brought together into a favourable orientation in the enzyme-substrate complex, the enzyme is catalysing the reaction. However, this does not mean that the enzyme alters the equilibrium of the chemical reaction; in a reversible reaction the enzyme also increases the speed at which the reverse reaction occurs. In the previous example, the enzyme is also able to bind to the products, placing them in a favourable orientation for the reverse reaction to occur, reforming the reactant molecules; the enzyme increases the k_{on} and k_{off} rates in Equation 1.18 by the same amount, keeping the equilibrium at the same point. Therefore, enzymes have no effect on the free-energy change of a reaction, nor do they affect the reaction's equilibrium constant [107].

When a molecule reacts to form a product, it does not alchemically mutate into the product - it undergoes a series of chemical changes that step-by-step transform the reactant into the product. These changes may be the addition of a functional group, or simply the rotation of a bond or redistribution of electrons. At each intermediate step the molecule has different thermodynamic properties, and so has an altered Gibb's free energy. In Section 1.3.1 it was shown that a reaction has a corresponding change in Gibb's free energy that dictates whether the reaction will occur spontaneously, and that this is dependent only on the free energy states of the initial and resultant species. However, the intermediate states become important when considering the rate of the reaction, because if during a reaction the reactant molecule must adopt a geometry that has a prohibitively high free energy, it is less probable that it will do so, and as such the rate at which it transitions to product will be lower [107]. This is referred to as the transition-state theory, and is shown diagrammatically in Figure 1.16 [15]. The configuration that the reactant molecule adopts with the highest free energy is termed the transition state, and the change in free energy resulting from this conversion is termed the *activation barrier* ($\Delta G^{\ddagger}$). Figure 1.16 shows the relationship between ΔG and $\Delta G^{\ddagger}$.

Enzymes act to increase the rate of reaction by decreasing $\Delta G^{\ddagger}$. By binding the substrates to form an enzyme-substrate complex, a different reaction pathway is followed by the substrates which has a reduced activation barrier. This increases the probability that the substrates will be able to reach this state, and so the rate of reaction is accelerated. It is important to note that enzymes are not simply chaperones whose sole responsibility is to align substrates in the correct orientation to reach a lower activation barrier; they are regularly directly involved in the making and breaking of covalent bonds in their substrates, often forming covalent bonds with the substrates themselves [1]. This helps to explain why enzymes are so specific in both the reaction that they catalyse and in their choice of substrates. For example, the GA/GB mechanism for catalysis of peptide-bond hydrolysis requires an acid and a base to be appropriately aligned for effective electron movement and transfer. This requires that the enzyme's quaternary structure can not only accommodate the peptide substrate and a water molecule in its active site, but also that it can align them with respect to an acid

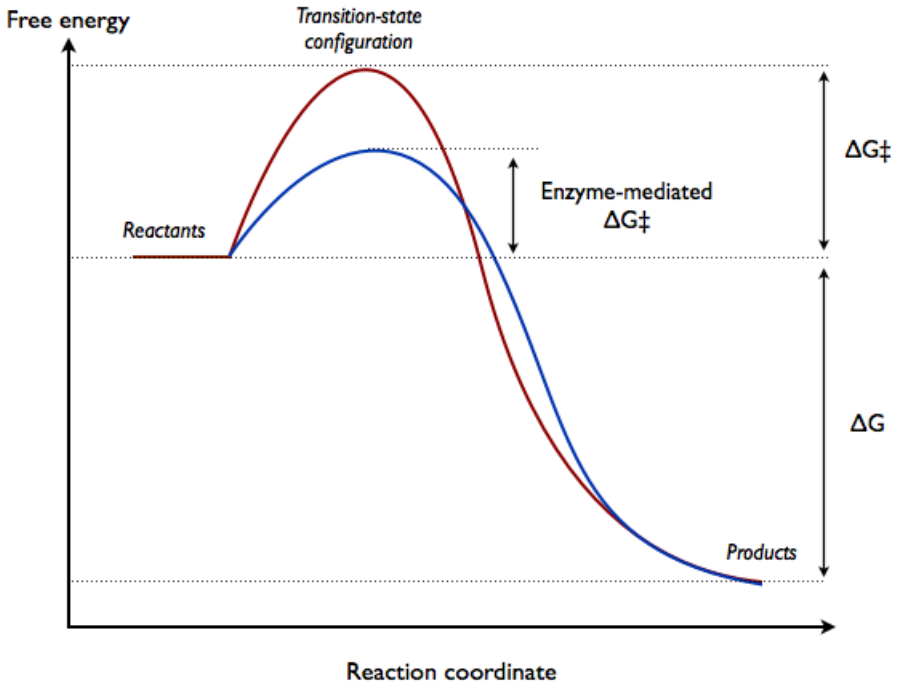


Figure 1.16: Illustration of the transition-state theory for a spontaneous reaction. The reactants have a certain amount of free energy (G). Energy is required to make the reactants adopt the transition-state configuration, which has a higher G and is therefore unfavourable. The free energy required to adopt this structure is the activation barrier ($\Delta G^{\ddagger}$). As the reaction proceeds beyond this state, the ΔG gets more negative and therefore proceeds spontaneously until the products are produced. The difference in free energy between the reactants and the products is ΔG. In this example, $\Delta G < 0$ so the reaction occurs spontaneously. The enzyme-mediated reaction pathway (in blue) has a reduced ($\Delta G^{\ddagger}$), making it easier to overcome the activation barrier, but the ΔG of the reaction remains the same.

side-chain and a base side-chain for the catalysis to occur. Therefore the quaternary structure of the enzyme, and in particular of the active site cleft, is as important as its catalytic residues. This reinforces the notion that mutations in residues spatially far away from the catalytic residues can still have a pronounced effect on enzyme-substrate binding; returning to the previous example, the mutation causes a mal-alignment of the acid and base residues, this can result in the water not being held in the optimal position for the water's oxygen to attack the peptide's carboxyl carbon, thus increasing the reaction's activation energy. This is so sensitive that replacing a catalytic glutamic acid residue with an aspartic acid, which shifts the position of the catalytic carboxylate ion by 1 Å, reduces the enzyme's proteolytic activity in the order of a thousand-fold.

1.3.4 Inhibitor Kinetics

The biochemistry of competitive-reversible-inhibitor reaction kinetics is much the same as that of enzyme-substrate kinetics, except that once bound by an inhibitor, no reaction occurs (See Equation 1.24).

$$\begin{array}{rcl} E, S, I & \Longleftrightarrow ES \longrightarrow E + P & \\ \Updownarrow & & (1.24) \\ EI & & \end{array}$$

$$\Rightarrow \quad K_i = \frac{[E][I]}{[EI]} \tag{1.25}$$

$$\Rightarrow \quad \Delta G_i = RT log_e K_i \tag{1.26}$$

$$\equiv \quad K_i = e^{\Delta G_i / RT} \tag{1.27}$$

Therefore, the same kinetics that were applied to enzyme-substrate reactions can be applied to competitive inhibitors. Equation 1.25 is the same as 1.19 but for competitive inhibitors, and shows that the dissociation constant between the enzyme and inhibitor (K_i) reflects the strength of binding between an enzyme and an inhibitor; the greater the strength of attraction between the two, the lower the value of K_i. This dissociation constant can then be used to directly calculate the change in free-energy upon inhibitor binding (ΔG_i), as shown in Equation 1.26. It can be seen from Equation 1.26 that the

change in free-energy upon binding is directly proportional to the natural logarithm of the dissociation constant, so an inhibitor that binds more strongly (reflected by a smaller K_i) would have a more negative change in free energy. The rearrangement of Equation 1.26 shown in Equation 1.27 is used more often, as biochemical techniques such as Isothermal Titration Calorimetry are able to determine the change in free energy associated with inhibitor binding, which can then be used to calculate the dissociation constant [107, 103].

Isothermal Titration Calorimetry (ITC) is a method for measuring the temperature change of a chemical reaction through the stepwise addition of one of the reactants to another. Upon addition, the reaction causes a change in temperature of the system depending on whether the reaction is endothermic or exothermic. ITC measures the energy required to maintain the system at a constant temperature, so if the reaction is exothermic, less energy is required by the 'heating apparatus' to keep the system at a constant temperature. Conversely, an endothermic reaction requires an input of energy to counteract the thermal energy removed from the system for the reaction. From the energy required to maintain the temperature, the change in enthalpy upon binding can be directly determined, and the change in entropy and free energy indirectly calculated. Information such as the dissociation constant (K_d) and the stoichiometry of binding can also be ascertained [101]. This principle can be extended to protein-inhibitor binding to calculate the inhibitor's kinetic parameters, though standard ITC experiments are unable to provide accurate estimates of the binding affinity for strong inhibitors. Therefore an adapted method called 'displacement ITC' (DITC) is used, where the inhibitor is titrated into a solution containing a protein pre-bound to a weak inhibitor. By displacing the weaker inhibitor, comparable values to standard ITC can be attained. DITC has been successfully used to calculate thermodynamic data on inhibitors to HIV-1 protease, revealing information on how mutant proteases lower their sensitivity to the inhibitors [70, 117].

While biochemical assays such as ITC are the standard for calculating inhibitor binding affinities, recent advances in computing have meant that computational method for calculation are feasible. Thermodynamic analysis of computer simulation methods

such as molecular dynamics or Monte Carlo simulations can be implemented to calculate free energy, enthalpy and entropy changes upon inhibitor binding. These values can then be compared against the biochemical assays for validation. The benefit of the computational approach is that it can be used to generate binding affinities for inhibitors that have not yet been synthesised, and so can be used to screen a range of potential drug structures without having to synthesise them.

1.4 Computational Introduction

1.4.1 Molecular Modelling

Molecular modelling is the application of theoretical techniques to model the behaviour of atoms and molecules, most commonly through time. For example, molecular mechanics can be used to model the Brownian motion of a gas molecule, or model the movement of an ion through an ion channel, or model the binding of a ligand to its receptor. Originally, the calculations required to describe the atomic or molecular behaviour were performed by hand, which meant that only the simplest molecules could be described in this way. However, the advent of the computer, with its ability to repetitively perform the necessary calculations, meant that the molecular descriptions could become more detailed, and the models theorised over longer periods of time [49]. As the processor power of computers, and their ability to be networked together to allow multiple computers to share workloads, has increased, the scope of molecular modelling has increased to where it is now possible to describe the behaviour of an entire viral capsid containing over 1 million atoms over 13 nanoseconds [22].

In order to model the behaviour of molecules, a description of the atomic interactions is necessary. There are two main methods of describing their interactions: quantum mechanics and molecular mechanics. Quantum mechanics (QM) explicitly considers the electrons of atoms, modelling the interactions between atoms as a function of their electronic distributions. Molecular mechanics (MM), conversely, ignores the electrons and models the interaction between atoms as a function of their bond lengths, angles, dihedral angles and non-bonded forces. As QM explicitly includes electron distributions, it gives a more accurate representation of the atomic interactions, but is subsequently

much more computationally-demanding than molecular mechanics. Consequentially it is intractable to perform QM calculations on biological systems, which typically contain thousands of atoms [49]. However, a more recent alternative approach is the quantum mechanics/molecular mechanics (QM/MM) hybrid which models the majority of the system with a Newtonian (MM) representation, but describes a constrained region around certain important atoms (such as catalytic residues in the active site) according to a QM model. The benefit of this hybrid approach is that it allows for macromolecular systems to be modelled with explicit electrons at biologically-important atoms in a feasible timescale. However, this approach requires a coupling of the QM and MM regions; those atoms interacting within both the QM region and the MM region require special treatment which can introduce errors into the simulation [52].

1.4.2 Molecular Mechanics

At the heart of a molecular mechanical description of atomic interactions is the Born-Oppenheimer approximation, which states that the motions of the nuclei can be decoupled from the motions of the electrons, and so because the mass of the electron is so much less than the mass of the nucleus, any movements of the nucleus will have an instantaneous movement of its surrounding electrons accordingly. Therefore the energy of an atom can be written as a function of its nuclear co-ordinates alone, allowing the electrons to be disregarded. Subsequently, molecular mechanics calculates the potential energy of an atom through 5 relatively simple atomic interactions:

$$\mathcal{V}(\mathbf{r}^N) = \sum_{\text{bonds}} \frac{k_i}{2}(l_i - l_{i,0})^2 \tag{1.28}$$

$$+ \sum_{\text{angles}} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 \tag{1.29}$$

$$+ \sum_{\text{torsions}} \frac{V_n}{2}(1 + \cos(n\omega - \gamma)) \tag{1.30}$$

$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \tag{1.31}$$

where $\mathcal{V}(\mathbf{r}^N)$ is the potential energy, which is a function the positions ($\mathbf{r}$) of N atoms,

$\mathbf{l_i}$ is the bond length between two bonded atoms; $\mathbf{l_{i,0}}$ is the 'optimal' bond length for the equivalent bond; $\theta_\mathbf{i}$ is the angle between two bonded atoms; $\theta_\mathbf{i,0}$ is the 'optimal' angle for the bond between the equivalent atoms; ω is the dihedral angle between the two planes set by four bonded atoms; γ is the 'optimal' dihedral angle for the equivalent bonded atoms; $\mathbf{r_{ij}}$ is the distance between two non-bonded atoms; and $\mathbf{q_i}$ and $\mathbf{q_j}$ are the charges on two non-bonded atoms i and j. All the other terms (k_i, V_n, σ_{ij}, ϵ_{ij}, ϵ_0) are empirically-derived constants designed to reflect the differing nature of the various atom constituents. So, for example, a bond between 2 carbon atoms will have empirically-derived optimal $l_{i,0}$, $\theta_{i,0}$, γ, k_i, V_n, σ_{ij}, ϵ_{ij}, ϵ_0 values, and these will be different to a carbon-hydrogen bond, which will have its own empirically-derived values [49, 17, 65].

This equation states that the potential energy of N atoms ($\mathcal{V}(\mathbf{r}^N)$) is equal to the sum of the contributions due to bonds deviating from an equilibrium value (Equation 1.28); the contributions due to bond angles deviating away from an equilibrium value (Equation 1.29); the contributions due to rotations around a bond (Equation 1.30); and finally the contributions due to the two types of non-bonded interactions - non-polar van der Waals and polar electrostatic (Equation 1.31) [49]. This equation (also called a *force field*) is primarily designed to reproduce structural properties of a system, and as such the equilibrium values will be empirically parameterised to computationally reproduce the structure. There are many different forcefields that differ in their equilibrium values; there is no one correct forcefield as some perform better under certain circumstances and with particular systems than others [49].

1.4.3 Molecular Dynamics

Molecular dynamics (MD) is a deterministic method of molecular modelling, which means that subsequent states of the system can be predicted from its current state. The subsequent states are generated by calculating the potential energy of the system through molecular mechanics, and from this calculating the forces on each atom. These forces can then be converted into an acceleration of each atom through Newton's second Law of Motion, which, when combined with the current atomic positions and velocities, can calculate the future position of the atom based on the forces currently acting on it.

By repeating this multiple times over very small timsteps, the motions of all the atoms are generated with respect to time [49].

Molecular dynamics utilises a molecular mechanics forcefield to calculate the potential energy of all the atoms in a system. However, even using this heuristic method, calculation of the potential energy of large biomolecular systems is computationally-demanding. Therefore MD software packages employ techniques to increase the computational efficiency with minimal effect on the potential energy calculation:

- **Periodic Boundary Conditions**. In order to simulate a biomolecular system under *in vivo* conditions, it must be simulated in solvent. However, to effectively simulate the effect of the solvent, the number of solvent molecules required would be too large to efficiently compute. Therefore the system is placed in a cubic, hexagonal, octahedral, rhombic dodecahedral or elongated dodecahedral cell containing a relatively small number of solvent particles. The reason for choosing these cell shape geometries is because they all allow tessellation in 3-dimensions. Tessellating the cell in 3-dimensions enables the particles in a cell to experience forces as though they were in a much larger bulk fluid. When an atom moves out of the cell, it is replaced by an image particle that enters from the opposite side [49].

- **Hydrogen Constraints**. MD simulations determine the future position of each atom by determining its position, velocity and acceleration, and using these to calculate its position at a future point in time. However, the length of time over which the equations of motion are integrated must be relatively small because as the atomic positions change, the forces experienced by the atoms summarily change. If small enough timesteps are not used then the change in forces experienced by each atom is not correctly captured and subsequently the simulation will not be realistic. The length of time over which the equations of motion are integrated is therefore governed by the fastest oscillating atoms, as these will experience change in forces acting on them most rapidly. In biomolecular systems these are the hydrogen atoms, which oscillate at a frequency that requires timesteps of 1 femtosecond. However, the flexibility of covalent hydrogen bonds is

often less important than the low-frequency motions of the biomolecular system over much longer time periods. Therefore covalent bonds to hydrogen atoms are kept rigid to allow the timestep to increase to 2 femtoseconds, which allows observation of large-scale motions over longer time periods for the same computational cost [49].

- **Non-bonded *van der Waals* Force Calculation Optimisation**. The calculation of the non-bonded contributions to the potential energy of each atom is the most time-consuming part of the MD simulation because it must sum all of the non-bonded potentials for every combination of atom pairs in the system. Therefore, while the bonded terms in the MM equation are proportional to the number of atoms in the system (N), the non-bonded terms increase by N^2. However, calculating the interactions between every pair of atoms in the system is unnecessary because the non-polar *van der Waals* effect is proportional to r^{-6}. The contribution of distant atoms to the *van der Waals* energy term is therefore insignificant enough to ignore. A cut-off distance is applied such that all atoms beyond the cut-off are ignored in the *van der Waals* calculation. When used in conjunction with the period boundary conditions, the maximum size of the cut-off is limited to half the length of the cell to ensure that an atom does not experience its own *van der Waals* force. A cut-off value of 10Å usually gives relatively small errors [49]. However, determining whether each atom lies within the cut-off distance for each timestep is almost as time-consuming as calculating the energy itself. Therefore a 'non-bonded neighbour list' is utilised to identify which atoms to include in the non-bonded calculation without having to recalculate its distance. This method stores all the atoms within the cut-off distance, along with all atoms slightly further away, in an array. Under the assumption that the atoms within the 'extended' cut-off distance do not significantly change position over 10 to 20 timesteps, only the contribution from those atoms are considered for the force calculation over the 10-20 timesteps. The atom list is therefore only updated every 10-20 timesteps, which reduces the required number of calculations [49].

- **Non-bonded Electrostatics Force Calculation**. Unlike the *van der Waals*

term, the electrostatics term only decreases by r^{-1}, so therefore has significant contributions at much longer ranges. The Ewald summation method is employed in this thesis to deal with these electrostatic contributions. In this method, a particle interacts with all the other particles in the simulation cell, and with all of their images in the infinite array of periodic cells. The electrostatic component is split into two parts: a short-ranged potential with a cut-off, and a periodic long-ranged potential which can be represented by a finite Fourier series. To screen the atomic charges at each atomic position, a Gaussian distribution of neutralising charge is placed around each atom. These screening charges are constructed to make the electrostatic potential due to the atom rapidly decay to nearly 0. The sum of the atomic charges and the neutralising distributions are then calculated, and a second charge distribution added to the system which exactly counteracts the first neutralising distribution. As this second charge distribution is not efficiently summed in 'real space', it is performed in 'reciprocal space' and then the result converted back into real space.

The Ewald summation is computationally expensive; depending on the width of the Gaussian distribution for the neutralising charge it can scale from $N^{\frac{3}{2}}$ up to N^2. This can be improved upon by using fast Fourier transform (FFT) to compute the reciprocal space summation for the second set of neutralising distributions, which scales as $Nln(N)$. However, the FFT method requires that the data are discrete values. Therefore the particle-mesh Ewald method is employed to place a grid across the simulation box. The charges of the atoms are then distributed among its 27 surrounding grid points so as to reproduce the potential of the charge at the original location. FFT is then used to calculate the potential due to the Gaussian distributions at the grid points, which corresponds to the desired potential at each of the particles [49].

Having calculated the potential energy of each atom in the system through molecular mechanics, the force acting on each atom as a result of its potential energy can be calculated by:

$$\vec{F_i} = -\nabla_i V \tag{1.32}$$

where F is the force on atom i and $\nabla_i V$ is the directional gradient of the potential energy on atom i. From this vector force acting on each atom, it is possible to determine the acceleration experienced by each atom as a result through Newton's second Law of Motion:

$$\vec{F_i} = m \cdot \vec{a_i} \tag{1.33}$$

where a is the acceleration experienced by atom i. This force-derived acceleration is then combined with the atom's position and velocity to calculate its position at a time $t + \delta t$. However, at the first timestep there are no velocities carried over from the previous timestep, so the initial distribution of velocities are chosen randomly from a Maxwell-Boltzmann distribution such that the probability that atom i with mass m_i has a velocity v_{ix} in the x direction at temperature T is given by [17, 49]:

$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T}\right)^{\frac{1}{2}} exp\left[-\frac{m_i v_{ix}^2}{2k_B T}\right] \tag{1.34}$$

To ensure that the total momentum of the system is zero, the sum of the components of the atomic momenta along each axis is divided by the total mass of the system. This value is then subtracted from the atomic velocities, resulting in an overall momentum of zero for the system.

With the positions of the atoms attained from a static structure, the velocities randomly assigned to each atom, and the accelerations calculated from the potential energy, the **finite differences method** can be employed to integrate the equations of motion and determine the future positions of the atoms. This method breaks down the integration into many small stages separated of length δt. The equations of motion are integrated between times t and $t + \delta t$ over which the forces are assumed to be constant on each atom. This is then repeated between times $t + \delta t$ and $t + 2\delta t$ using the positions and velocities outputted from the previous timestep and recalculating the forces on each atom at $t + \delta t$ to recalculate the acceleration. The algorithms that implement the finite differences method assume that the positions, velocities and accelerations can be

approximated as Taylor series expansions:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) + \frac{1}{6}\delta t^3\mathbf{b}(t) + \frac{1}{24}\delta t^4\mathbf{c}(t) + \cdots \quad (1.35)$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \delta t\mathbf{a}(t) + \frac{1}{2}\delta t^2\mathbf{b}(t) + \frac{1}{6}\delta t^3\mathbf{c}(t) + \cdots \quad (1.36)$$

$$\mathbf{a}(t+\delta t) = \mathbf{a}(t) + \delta t\mathbf{b}(t) + \frac{1}{2}\delta t^2\mathbf{c}(t) + \cdots \quad (1.37)$$

$$\mathbf{b}(t+\delta t) = \mathbf{b}(t) + \delta t\mathbf{c}(t) + \cdots \quad (1.38)$$

where $\mathbf{v}$ is the vector velocity, $\mathbf{a}$ is the vector acceleration, $\mathbf{c}$ is the third derivative of the vector positions with respect to time $\mathbf{r}$(t) *et cetera*. There are many algorithms for integrating these equations of motion in an MD simulation, the most common of which is the Verlet algorithm.

1.5 Project Aims

Prior research involving molecular dynamics simulations of HIV-1 protease inhibitor resistance can been divided into two main areas of interest:

1. Underlying structural basis behind the effects of various primary and secondary mutations on protease inhibitor efficacy. Examples of this research include the structural analysis of the I47A lopinavir-resistant mutation [41], and structural comparisons between WT and V82F/I84V simulations [77]. This area is retrospective in nature, as it tries to explain the underlying causes of identified resistance mutations. It also has an important role in drug design, for example in identifiying novel drug targets based on protease dynamics [78].

2. Identifying inhibitor resistance-level as a reflection of changes in the strength of binding between protease and inhibitor. An example of this research is the work done by Maschera *et al.* (1996) into dissociation rate constants for drug-resistant protease mutants [59]. In contrast to structural understanding of resistance, this is both prospective as well as retrospective in nature, because every genotype can theoretically be simulated with every inhibitor to generate a strength of binding between the two which can be directly correlated into a level of resistance. It is not feasible to extend this to structural understanding because without knowledge

of the resistance-phenotype of a genotype, it is difficult to know how to interpret changes in dynamics.

Conception of this project began with the research undertaken by Wang *et al.* (2001). They developed a value termed the 'free energy/variability' (FV) value that combined the strength of binding between HIV-1 protease and an inhibitor, calculated using the heuristic MM/PBSA method, with the variability of each position across the protease [120]. This FV value represented the fold-change in drug resistance for a particular mutation. Their results showed that they could accurately predict the phenotypic effect a protease mutation had on inhibitors, with an average of 76% concurrence to experimental results. However, they only collected data from their simulations over 120 picoseconds, yet work undertaken by Ishima *et al.* (1999) showed that the flaps move on a timescale in the microsecond range [36]. Therefore, as the prediction accuracy shown by Wang *et al.*'s results was very encouraging, it was decided to investigate whether utilising this method over longer timescales, encompassing more of the protease's dynamics, would yield an even greater prediction accuracy. Successful application of nanosecond-timescale molecular dynamics to the calculation of a drug binding affinity value would enable the methodology to be used as a diagnostic tool, where clinician submits the genotype of the majority HIV population to an automated pipeline which simulates the genotype complexed to the FDA-approved drugs and calculates a strength of binding with respect to wild-type. With sufficient computational resource, a result could be generated in days, rather than the two weeks for current diagnostics.

Chapter 2

Methodology

2.1 NAMD

As discussed in Section 1.4.3, there are multiple software packages written to implement the theory of molecular dynamics, including CHARMM [10], AMBER [76], LAMMPS [82] and NAMD [42]. During this research, the Nanoscale Molecular Dynamics (NAMD) package was chosen to perform the simulations. The reason for this was primarily because NAMD was specifically designed to efficiently utilise parallel computers, thus distributing the high computational requirement of simulation over multiple computers. Furthermore, the NAMD package incorporates various algorithms to reduce the computational demand at each timestep, including the Particle Mesh Ewald algorithm, and the multiple time-step velocity Verlet integration method [64]. These allow for a faster performance on networked processors, such as those in Grid networks or in supercomputers.

2.1.1 Input Files

In order to start a molecular dynamics simulation, NAMD requires 4 input files:

1. **Forcefield file**. This contains the empirically-derived force constants described in Equations 1.28, 1.29, 1.30 and 1.31. These constants are specific for each atom-pairing in the amino acids. For example, the equilibrium bond length and angle for an N-H bond in arginine will be listed, but these will be different to an N-H bond in aspartic acid, and will also be different from another N-H bond in

arginine's side-chain. Every force constant for every atom-pairing in all 20 amino acids is given in the forcefield file. There are many forcefield files available, each containing slightly different empirically-derived constants. The simulations run during this research used the AMBER forcefield for description of proteins (ff03). The standard forcefield, however, only provides force constants for the atom-pairings found in amino acids. Therefore, as saquinavir contains atom-pairings not found in amino acids, its own forcefield file had to be generated using the General AMBER Force Field (GAFF).

2. **PDB file**. This contains the atomic coordinates of all the atoms in the system. However, the file provides no information as to how the atoms are structurally linked. Atomic coordinates in this file may be manipulated prior to simulation using Visual Molecular Dynamics (VMD), which is a molecular visualisation program designed for displaying, animating, manipulating and analysing biomolecular systems. This will be described in more detail below.

3. **Topology file**. This contains the complete description of all the interactions in the system. This includes a list of all the atoms in the system along with their connectivity to other atoms. In conjunction with the PDB file, this contain all the structural information of the system.

4. **Configuration file**. This contains all the information NAMD requires to run the simulation. Information contained includes location of forcefield file, PDB file and topology file; number of timesteps over which to run; output filenames; and size of system.

These four files together contain all the information necessary to run a molecular dynamics simulation. The generation of these files for the HIV-1 protease systems simulated in this study will now be discussed. The initial atomic coordinates are attained from the PDB files downloaded from the Protein Data Bank. This is loaded into VMD, and the monomers (and ligand if present) are split into separate PDB files, leaving behind all ions and water molecules that were resolved in the original file. If it is necessary to mutate any residues in the protease, then each monomer in turn is loaded back into VMD, and the 'mutate' command invoked on the residues, which alchemically mutates

them into the specified mutation. The protonation state of each monomer is assigned in this way; in the case of complex formation with saquinavir, the aspartic acid at residue 25 on monomer A is designated as protonated, and that on monomer B designated as un-protonated. This α-protonated state was chosen after calculating the free energy of binding between Hxb2-genotype HIV protease and saquinavir for the 4 possible protonation states (α-, β-, un- and di-protonated). The 4 systems were simulated for 4ns and then analysed using MMPBSA and NMODE. The results showed that the α-protonated state had the most negative ΔG, which was in agreement with literature, so this protonation state was used. Hydrogen atoms, which are generally too small to be resolved in the crystal structure, are also added at this point. Histidine residues were protonated on their ϵ-nitrogen and un-protonated on their δ-nitrogen, as this is their dominant form at physiological pH values. The caveat of the alchemical mutation is that surrounding atoms are not moved to make way for the mutated atoms, so if a small internally-located residue is mutated into a larger residue, there will be energetic ramifications. This is resolved by minimisation of the protein prior to simulation, and subsequently carefully allowing the protein freedom of movement around the mutated residues to allow optimal reorientation of surrounding atoms.

Following mutation and rehydrogenation of the protease, the monomer files are concatenated with the inhibitor file to regenerate the protease complexed with the ligand. Using the Leap module of the AMBER software package, the complex is placed in a TIP3P water box whose sides are 14.0Å away from the surface of the protease. This generates an approximately 40,000 atom box of water of the specified dimensions, then removes all the water molecules in the center to form a cavity just large enough for the protein to fit. This protocol forms a vacuum immediately surrounding the protein which must be considered during the equilibration phase of the simulation. The reason for placing the protein in such a large box of water is because during the course of a simulation the protein rotates and moves around, and so quite often will move to the edge of the box. NAMD employs 'periodic boundary conditions' to the system, which means that for force calculations it acts as though there are multiple boxes tessellating around the water box, therefore simulating the water box extending to infinity. However, if the water box is not sufficiently large, the protein may experience non-bonded forces from

the protein in tessellated systems. Therefore it is prudent to place the protein in a large water box. Leap is then used to calculate the total charge of the system; it creates a grid with a 1Å spacing and calculates the Coulombic potential at each grid point and then places counterions (Na^+ or Cl^-) at the points of highest or lowest electrostatic potential in the solvent. As HIV protease systems are positively-charged, sufficient Cl^- ions were added to neutralise the system. Once the system is neutralised, Leap outputs a .pdb coordinate file of the whole system and a .top topology file which contains both the topological and parametric information of all the atoms in the system. Aside from the configuration file containing parameters for NAMD, these files contain all the information required to run the simulation.

2.1.2 Simulation Protocol

Molecular dynamics codes, such as NAMD, have multiple parameters that allow the user to specify exactly how they would like the simulation to run, allowing for either more rigorous force calculations and therefore slower performance, or reduced cut-off distances for non-bonded forces and therefore faster turnaround of results. The configuration file contains the user's exact specifications for their simulation and is a required input for NAMD. All simulations were performed with an integrative timestep of 2 femtoseconds. To allow this, all simulations were performed with the SHAKE algorithm imposed on atoms covalently bound to hydrogen to constrain their high vibrational frequencies [93, 49]. A non-bonded cut-off distance of 12Å was employed for the particle mesh Ewald (PME) calculation of non-bonded electrostatic forces for all simulations. Detailed below are the parameters specific for the minimisation and equilibration phases of the simulation protocol.

- **Minimisation**. 'Conjugate gradients' and 'steepest descents line search' minimisation algorithms were applied to the protease for 2000 timesteps (4 picoseconds). Essentially these two methods attempt to minimise the energy of the protease by altering its configuration towards the point on its potential energy landscape where:

$$\frac{\delta f}{\delta x_i} = 0; \qquad \frac{\delta^2 f}{\delta x_i^2} > 0 \tag{2.1}$$

  where f is the molecular mechanics energy of the Cartesian coordinates of the

atoms (x_i). The line search algorithm achieves this by locating 3 points along a specified direction through the landscape where the energy of the middle point is lower than the two outer points. Through an iterative procedure, the configuration of the minimum energy point along this line is determined [49]. This method only locates the minimum energy along the chosen direction, which in all probability will not intersect the minimum energy configuration of the local landscape. Therefore, subsequent line searches must be performed starting at the result of the previous line search. The direction in which to perform the next line search is determined by the conjugate gradients method. After 2000 iterations the gradient of the landscape at the protein's configuration should be low enough that proximity to a minima is assured.

- **Equilibration**. The temperature of the system was set at 50 Kelvin and a restraining force of 4kcal/mol/Å^2 applied to all non-hydrogen atoms in the complex. This prevents the complex from gaining too much energy too rapidly and subsequently adopting erroneous configurations, and also allows the unrestrained water to fill the vacuum left around the protein by Leap's hydration protocol, which also helps prevent undesired configurations [60]. The system was then heated to 300 Kelvin over 25,000 timesteps (50ps) while still restraining the non-hydrogen atoms. Once the desired temperature was reached, it was maintained for all subsequent simulation steps using a Langevin thermostat with a damping coefficient of $\lambda = 5/\text{ps}$ applied to all non-hydrogen atoms. The simulations were performed under constant pressure through coupling the system to a Berendsen pressure bath, with a target pressure of 1.01325 bar and a pressure coupling constant of 100 femtoseconds. Simulations were therefore run in a constant NPT ensemble.

  The water molecules were allowed to equilibrate for 100,000 steps (200ps) before the restraints were slowly lifted off the protein. If VMD had been used to mutate any residues,the restraints were removed for all atoms in a 5Å sphere around each mutated residue in turn for 25,000 steps (50ps) before re-applying to the non-hydrogen atoms. This was to allow the alchemically-mutated residues the freedom to reorientate to a more energetically-relaxed configuration. As this

may require the residue to deform the surrounding structure to 'free' itself, the restraints on the surrounding atoms were lifted as well. The residues' constraints in each monomer were lifted simultaneously, so that a 5Å sphere around residue 90 in both monomers was released for 50ps before re-restraining both. Once all mutated residues had been given time to re-orientate, the restraint forces were gradually lifted off the ligand in 1kcal/mol/Å^2 steps every 50ps to prevent it from suddenly being given the freedom to dissociate from the protease. The restraints were then gradually lifted off the protease in the same 1kcal/mol/Å^2 step every 50ps. To give the protease time to configurationally and energetically relax before data collection, the simulation was further run until the total equilibration time lasted 2 nanoseconds. This length of time was therefore dependant on the number of mutated residues in the monomer.

Once equilibration had been completed, the simulation was able to be run as an all-atom unrestrained molecular dynamics simulation for the desired length of time. Simulations were run in 500,000-step sections (1ns), over which time, the coordinate positions of all the atoms in the system were outputted to a binary .dcd file every 500 timesteps (1ps). The coordinates and velocities outputted at the end of each 1ns section were used in conjunction with the .top file to start the next simulation. This phase of the protocol is referred to as either the 'production-phase' or the 'data-collection phase'. The coordinate file outputted at the end of a 1ns simulated section could be loaded into VMD along with the .dcd file to visualise the simulation over that time period. VMD could then be employed to structurally analyse the protein through RMSD or RMSF. This will be described in more detail in Section 2.2.

2.2 Structural Analyses

2.2.1 Root Mean Square Analysis

Root Mean Square Deviation (RMSD) is a statistical measure of the similarity between two sets of values. In the field of biology these values are commonly the atomic coordinates of homologous proteins. The proteins first need to be structurally aligned through the method of least-squares, which rotates one protein around its geometric

centre to match its orientation to the other protein. Then the protein is translated such that the sum of the distances between homologous pairs of atoms between the proteins is minimised. Once aligned, the RMSD is performed as described in Equation 2.2:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {d_i}^2} \tag{2.2}$$

where N is the number of superimposed atoms in each set of values, and d_i is the Euclidean distance between the i^{th} pair of superimposed values [123].

Most commonly only the C_α atoms from the backbones of each protein are considered. This is because proteins that do not share exact sequence similarity do not necessarily share the same number of side-chain atoms. However, all amino acids share the same core backbone atoms (NH-CH-CO_2), so regardless of the protein's genotype, as long as the RMSD is performed over homologous subsets of the proteins' structures, it will calculate a measure of structural similarity. Furthermore, as the C_α atoms are located in the structural backbone of the protein, they represent a good indication of the tertiary- or quaternary-structure of the proteins.

There are different ways of applying RMSD in bioinformatics:

- **Comparing the structures of two proteins**. This is the most common use for RMSD in bioinformatics; determining the overall structural similarity of two proteins, whether unrelated or closely-related. In Equation 2.2, N is the number of residues compared between the proteins, and d_i is the Euclidean distance between identical atoms in the two proteins. This is referred to as 'global RMSD' in this thesis as the sum of the differences for each residue is divided by the number of residues, giving a mean residue value across the whole protein.

- **Comparing the structures of a single protein across a simulation**. As a simulation proceeds, the protein will naturally flex and distort, resulting in the adoption of different conformations. Making N in Equation 2.2 the number of residues under comparison, and d_i the Euclidean distance between identical atoms in the two structures, the global RMSD between the initial structure and

the structure at snapshot X is calculated. Plotting of the global RMSD between the starting structure and snapshot X against snapshot number reveals how the global structure of the protein changes through the simulation. This is referred to as 'evolution of global RMSD' in this thesis.

- **Comparing the structures of a protein against its average structure**. This is termed Root Mean Square Fluctuation (RMSF) and is a specialised case of the previous RMSD applications. An average structure is generated through determination of the mean Cartesian coordinate for each atom. This average structure is then used instead of the initial starting structure for the RMSD calculation, such that d_i is the Euclidean distance between structure X and the average structure $\bar{X}$. If $\bar{X}$ is averaged from the snapshot structures across a simulation, then the plotting of RMSF against snapshot number reveals how the global structure of the protein fluctuates across a simulation.

- **Comparing proteins on a per-residue level**. Instead of averaging the RMSD for each residue across the protein, the root-squared Euclidean distance is calculated between two structures for each residue and plotted against residue number, an indication of the regions of high and low structural similarity can be determined. This is termed 'profile RMSD' or 'profile RMSF' in this thesis as it provides information on the similarity across the proteins' residue profiles. Furthermore, by making N in Equation 2.2 the number of snapshots, and d_i the Euclidean distance between an average structure's residue and snapshot N's residue, the average flexibility of that residue across a simulation can be plotted. d_i can also be the Euclidean distance between the initial structure's residue and snapshot N's residue, which indicates the motion of that residue away from the starting structure. This can then be repeated for each residue to show the motions of different regions of the protein through the simulation. This can be used in conjunction with global RMSD to reveal more information about structural differences; while global RMSD may show that two structures are different, it contains no information about *how* they are different. Profile RMSD can then show which regions cause the observed differences.

RMSD and RMSF analyses in this research were performed by VMD through execution of 'tcl' scripts that were written for this purpose. Different scripts were written to implement either global RMSD/RMSF, or profile RMSD/RMSF. These scripts load the proteins, and trajectories if necessary, into VMD. They then invoke VMD's ability to superimpose the proteins through the least-squares method, and then calculate either the profile or global RMSD depending on the script. VMD also has a command to calculate the average structure of a simulation, so the 'tcl' scripts written to implement RMSF calculations invoke this command before calculating the RMSD to this average structure. The scripts output the results to a plain text file which can then be loaded into a graphing program such as Microsoft Excel for analysis.

RMSD is typically used in molecular dynamics simulations as an indicator of the system having reached equilibrium. As the simulation proceeds, the system relaxes into a more natural configuration by undergoing conformational changes. These changes will be reflected by a change in RMSD; a plot of RMSD against time will show a gradual increase as the system relaxes, then plateaus out once relaxed. The system will then fluctuate around this relaxed configuration due to thermal motion of atoms and natural fluctuations in the system. If the rate of change of RMSD against time is approximately 0, then the system is considered equilibrated from a structural perspective.

2.2.2 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique used to reduce the complexity of multivariate data to identify the combinations of variables which explain the largest amount of variation within the multivariate data set. This is achieved by creating new variables, called principal components, which are linear combinations of the original variables that explain as much of the information in the data as possible. The first principal component describes the largest possible amount of information in the data; the second principal component describes the second largest amount of information not contained within the first principal component; the third principal component describes the next largest amount of information not captured by the first two; and so on. Although the total number of principal components is equal to the total number of variables in the data set, the first few principal components are usually sufficient

to jointly describe the majority of the variation in the original multivariate data set [20].

Molecular dynamics trajectories are examples of multivariate data; each system typically has thousands of degrees of freedom [131]. Applying PCA to molecular dynamics trajectories reduces the number of degrees of freedom, extracting the important concerted motions of groups of atoms from the random thermal atomic fluctuations. In simulations involving proteins, these concerted motions reveal information about the important, slow-timescale dynamics of the protein that are integral to the protein's thermodynamics, but would otherwise require long simulation times to observe [4].

The mathematical theory underpinning PCA on MD trajectories is as follows:
For a simulation with M snapshots, let $\vec{x}(t) = (x_1(t), x_2(t), ...x_{3N}(t))$ be the coordinate vector of N atoms in a system at time t. From these vector coordinates, a covariance matrix $\mathbf{C}$ is generated:

$$\mathbf{C} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,3N} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,3N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{3N,1} & c_{3N,2} & \cdots & c_{3N,3N} \end{pmatrix} \tag{2.3}$$

where the element $c_{i,j}$ is given by:

$$c_{i,j} = M^{-1} \sum_{t=1}^{M} (x_i(t) - \langle x_i \rangle)(x_j(t) - \langle x_j \rangle) \tag{2.4}$$

where $\mathbf{M}$ is the number of snapshots, $\langle x_i \rangle$ is the mean value of x_i across the trajectory, and $\langle x_j \rangle$ is the mean value of x_j across the trajectory. The eigenvectors and their associative eigenvalues are calculated by diagonalising the covariance matrix $\mathbf{C}$ such that:

$$\boldsymbol{\Lambda} = \boldsymbol{V}^{\boldsymbol{T}} \boldsymbol{C} \boldsymbol{V} \tag{2.5}$$

where Λ is the diagonal matrix containing the 3N eigenvalues as its diagonal elements and $\mathbf{V}$ is a square matrix containing the 3N eigenvectors:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{3N} \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} v_1(1) & v_2(1) & \cdots & v_{3N}(1) \\ v_1(2) & v_2(2) & \cdots & v_{3N}(2) \\ \vdots & \vdots & \ddots & \vdots \\ v_1(3N) & v_2(3N) & \cdots & v_{3N}(3N) \end{pmatrix} \tag{2.6}$$

The eigenvectors are numerically sorted according to their associated eigenvalue. The eigenvectors with the largest eigenvalues are termed the principal components, and represent the low-frequency correlated motions that contribute the largest fraction of the total variance. These values, however, do not indicate *when* the eigenvectors move in the simulation, so to determine this the eigenvector's **projection** is calculated:

$$p_i(t) = \mathbf{v}_i(\ \mathbf{x}(t) - \langle \mathbf{x} \rangle) \tag{2.7}$$

where p_i is the projection of the i^{th} principal component; v_i is the i^{th} eigenvector; $\langle x \rangle$ is a vector of the average Cartesian position of each of the 3N coordinate vectors in $\mathbf{q}$(t) across the trajectory. To ensure that differences observed between configurations across a simulation are due to structural fluctuations, the translational and rotational motions of the protein are removed through the least-squares method (Section 2.2.1 for more details) prior to analysis.

PCA was performed through two different software packages in this thesis. For the majority of the thesis, the program PCAZIP [61] was used. The MD simulation trajectory files were first converted from the .dcd files outputted by NAMD into .traj files through the PTRAJ module in the AMBER software suite. On conversion to the .traj format, all waters and ions were stripped from the system as they were unnecessary for the PCA analysis. This also drastically reduced the file size. A 'mask' .pdb file containing the coordinates and internal indices of the protein's C_α atoms was created through execution of a Perl script that was written specifically for this purpose. The script loads the PDB file containing the system's initial structure into VMD and then uses VMD's commands to select only the C_α atoms. The coordinates and atom numbers associated

with this selection is then outputted to the new PDB file. This 'mask' file contains the atoms over which to perform the PCA; due to the time required to diagonalise the covariance matrix, PCA was performed over just the backbone C_α atoms of the protease. The program was executed serially on a local computer, and took approximately 5 minutes to analyse the backbone atoms of a 10ns simulation. However, the maximum number of snapshots the program could analyse was 10,000. Therefore, once the ensembles' collective snapshots exceeded this number, PCAZIP could no longer be used. Instead, the PCA was implemented through the PTRAJ module of AMBER. A script was written that allows the user to specify the NAMD trajectory files over which the PCA is to be performed and how many eigenvectors to calculate. The trajectories are then automatically converted into .traj files containing just the complex, and then the PTRAJ commands are invoked to perform the PCA. As with PCAZIP, PTRAJ is not parallelised, and so the script took approximately 3 hours to implement the required calculations on the extended WT and HM ensembles (Section 6.5).

As can be seen in Equation 2.4, in order to analyse a simulation through PCA, an average structure over the trajectory must be created. For all non-ensemble simulations in this thesis, the average structure was generated automatically by PCAZIP, utilising every outputted snapshot in the simulation to determine each atom's time-averaged position. In Sections 6.4 and 6.5, PCA was used to analyse ensembles of simulations. In these cases, the time-averaged structure was calculated by concatenating all replicate .dcd files into a single long file and then using PTRAJ to generate the average structure from all snapshots. In this way, the principal components for each replicate in an ensemble are comparable against the other replicates.

2.3 Free Energy Calculations

Free energy is considered to be the most important quantity in thermodynamics [49]. Through statistical thermodynamics, it can be shown that if the number of molecules in a system is large, the behaviour of the system will be the same as the behaviour of the individual molecules that make up the system. Therefore, a molecular dynamics simulation of a single molecule can allow calculation of macroscopic thermodynamic properties

if it is run for sufficiently long. This underpins the justification for using molecular dynamics simulations of a single molecule to calculate macroscopic thermodynamic quantities such as the change in free energy upon ligand binding; a Boltzmann-distributed *ensemble* of micro-states sampled across a simulation will represent the macroscopic thermodynamic properties of the system at equilibrium. However, the Helmholtz free energy of a system (A), which is equivalent to the Gibb's free energy for systems with constant NVT rather than NPT, can be written as [29]:

$$A = -\frac{1}{\beta} ln Q_{NVT}(\mathbf{q}, \mathbf{p}) \tag{2.8}$$

where $\beta = 1/k_B T$; k_B is the Boltzmann constant; and Q_{NVT} is the 'partition function', which is the sum of all the Boltzmann factors. Through substitution, this equation can be rewritten as:

$$A = -\frac{1}{\beta} ln \langle e^{\beta H(\mathbf{q},\mathbf{p})} \rangle \tag{2.9}$$

This equation states that the free energy is directly proportional to the *ensemble average* of the exponential of the Hamiltonian. Therefore, configurations with a high energy make a significant contribution to the calculation of the free energy value of the system [29]. Molecular dynamics simulations, however, preferentially sample the lower energy configurations, and will never adequately sample the important high-energy regions, and so lead to poorly converged and inaccurate free energies [49]. Nevertheless, computational techniques for calculating the difference in free energy between two systems have been developed by exploiting the **thermodynamic cycle** principle. The techniques range from computationally-demanding, yet highly accurate, methods such as Free Energy Perturbation (FEP) and Thermodynamic Integration (TI) through to more heuristic techniques such as Molecular Mechanics Poisson-Boltzmann Solvation Area (MMPBSA) and Linear Response (LR) methodologies.

2.3.1 Thermodynamic Cycles

Equation 2.8 showed that in order to calculate an accurate free energy value, high-energy states which have a low probability of being sampled are important. Unfortunately, molecular dynamics and Monte Carlo simulations preferentially sample low-

energy states, so will not adequately sample these high-energy states in a reasonable timeframe. Accurate free energy differences between two states cannot therefore be directly calculated. For example, if the difference in binding between 2 ligands to a receptor were to be calculated:

$$R + L_1 \rightarrow R \cdot L_1 \qquad (\Delta G_1) \tag{2.10}$$

$$R + L_2 \rightarrow R \cdot L_2 \qquad (\Delta G_2) \tag{2.11}$$

Two simulations would need to be run where the ligand and receptor were gradually brought from a large distance apart to formation of the complex. This would require such a major reorganisation of the molecules involved (including the solvent) that adequate sampling of the whole phase space would be difficult. However, Equations 2.10 and 2.11 above can be linked into the cycle shown in Figure 2.1.

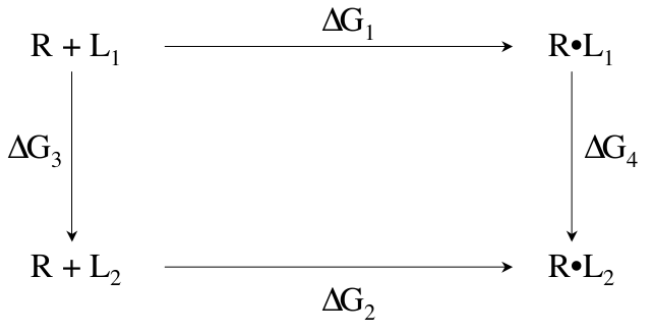


Figure 2.1: Conversion of Equations 2.10 and 2.11 into a thermodynamic cycle.

Thermodynamic cycles work on the principle that free energy is a state function, so the sum of the ΔG values around the cycle must equal 0 [49]. Therefore:

$$\Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3 \tag{2.12}$$

ΔG_3 represents the free energy difference of the unbound components in solution, and though this is not a value that can be determined experimentally, it can be calculated computationally. Equally, ΔG_4, which represents the free energy difference of the complexes in solution, cannot be determined experimentally, but can be computed. Therefore by 'alchemically mutating' L_1 into L_2 in both solution and in the receptor,

ΔG_3 and ΔG_4 can be calculated, and subsequently the $\Delta\Delta G_{3,4}$ between them, which is exactly equal to $\Delta\Delta G_{1,2}$. This *thermodynamic cycle perturbation approach* therefore calculates the difference in free energy of binding between the two ligands [49, 121]. This technique is employed by the computationally-demanding methods such as TI and FEP. However, it can also be adapted to calculate the absolute difference in free energy of binding in solution (Figure 2.2).

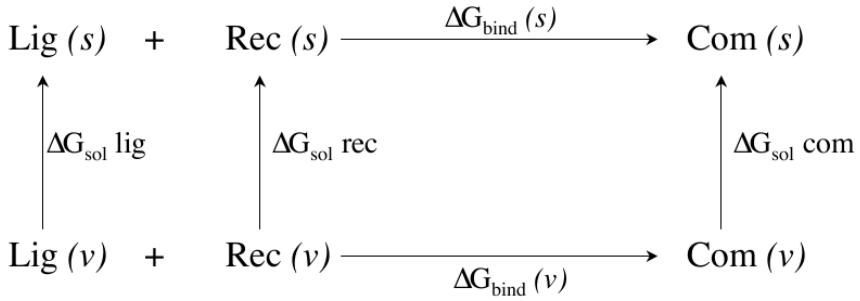


Figure 2.2: Thermodynamic cycle to calculate the change in free energy upon ligand binding ($\Delta G_{bind}(s)$).

Here, *(s)* denotes 'solvated'; *(v)* denotes 'in vacuum'; ΔG_{sol} denotes the change in free energy upon solvation for ligand *(lig)*, receptor *(rec)*, and complex *(com)*. The principles underlying thermodynamic cycles allow the $\Delta G_{bind}(s)$, which cannot be calculated through standard molecular dynamics, to be indirectly determined through Equation 2.13:

$$\Delta G_{bind}(s) = \Delta G_{bind}(v) + \left[\Delta G^{sol}_{com} - \left(\Delta G^{sol}_{lig} + \Delta G^{sol}_{rec}\right)\right] \tag{2.13}$$

Therefore, to calculate the the free energy change upon inhibitor binding, computational methods, such as MMPBSA, calculate these 4 free energy changes.

2.3.2 MMPBSA

The Molecular Mechanics/Poisson-Boltzmann Solvation Area method (MMPBSA) uses the thermodynamic cycle approach (see Figure 2.2) to approximately calculate the absolute change in free energy upon inhibitor binding in solution (Equation 2.13) [49]. To calculate the ΔG_{bind} *(v)* term, the *in vacuo* free energy of the ligand and the receptor

is subtracted from the *in vacuo* free energy of the complex:

$$\Delta G_{bind}(v) = G_{com}(v) - [G_{rec}(v) + G_{lig}(v)] \tag{2.14}$$

where $\Delta G_{bind}(v)$ is the change in Gibb's free energy upon ligand binding in a vacuum; $G_{com}(v)$ is the Gibb's free energy of the receptor-ligand complex in a vacuum; $G_{rec/lig}(v)$ is the Gibb's free energy of the separate protein and ligand components respectively in a vacuum. To determine the free energy of each of the components, the electrostatic, *van der Waals*, and internal molecular mechanics interaction energies are calculated for each component, and then the energies of the receptor and ligand are subtracted from the complex, such that:

$$G_X(v) \approx U_X(v) = U_{ele} + U_{vdw} + U_{int} \tag{2.15}$$

where X denotes either complex, ligand or receptor; U denotes energy; *ele* denotes non-bonded electrostatic forces; *vdw* denotes non-bonded *van der Waals* forces; and *int* denotes internal forces. The internal energy (U_{int}) is caused by strain from the deviation of bonds, angles and torsional angles away from their equilibrium values. Therefore the $\Delta \mathrm{G_{bind}(v)}$ is calculated as:

$$\begin{aligned} \Delta G_{bind}(v) &\approx \Delta U_{bind}(v) \\ &= \Delta U_{ele} + \Delta U_{vdw} + \Delta U_{int} \end{aligned} \tag{2.16}$$

where $\Delta G_{bind}(v)$ refers to the difference shown in Equation 2.14.

The solvation free energy $\Delta \mathrm{G_{sol}}$ components of $\Delta \mathrm{G_{bind}(s)}$ are the free energy changes to transfer the molecule from vacuum to solvent. This is considered to have a polar component and a non-polar component:

$$\Delta G_{sol} = \Delta G_{pol} + \Delta G_{np} \tag{2.17}$$

The non-polar contribution to the solvation free energy ($\Delta \mathrm{G_{np}^{sol}}$) is further divided into 2 components: the *van der Waals* interaction between the molecule and the solvent,

and the free energy change of forming a cavity in the solvent in which the molecule is placed. Due to the short-range distances over which the *van der Waals* forces act, and the major component of the free energy change associated with cavity formation arising in the first layer of solvent molecules, the non-polar contribution can be calculated from the solvent-accessible surface area (SASA) of the molecule:

$$\begin{aligned} \Delta G_{np}^{sol} &= \Delta G^{vdw} + \Delta G^{cav} \\ &= \gamma SASA + b \end{aligned} \tag{2.18}$$

where γ and b are empirically-derived constants that reflect the surface-tension of the solvent and an off-set value respectively.

The polar contribution to the solvation free energy (ΔG_{pol}^{sol}) is representative of the electrostatic forces that occur between the charged atoms in the molecule and the solvent upon solvation. This is more complicated than the other components of the $\Delta G_{bind}(s)$ value, and different methods have been developed to calculate this. Of particular note are the **Poisson-Boltzmann** (PB) implementation and the **generalised-Born** (GB) implementation of the electrostatic component. Both of these methods consider the solvent as an implicit continuum rather than considering every solvent molecule. This assumption relies on the fact that individual molecules are unimportant in the association reaction, but rather the solvent as a whole provides an environment which strongly affects the behaviour of the solute molecule.

The PB method treats the solute molecule as a body of constant low dielectric and the solvent as a continuum of high dielectric. The PB equation relates the electrostatic potential at a point $\phi(\mathbf{r})$ to the charge density $\rho(\mathbf{r})$:

$$\nabla \cdot \epsilon(\mathbf{r}) \nabla \phi(\mathbf{r}) - \kappa' \phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{2.19}$$

where κ' is a constant related to the Debye-Hückel inverse length (κ) and represents the ionic strength; and $\epsilon(\mathbf{r})$ is the dielectric constant.

This equation is solved by superimposing a cubic lattice onto the solute molecule and surrounding solvent. The values of the electrostatic potential, charge density, dielectric constant and ionic strength are then assigned to each grid point. As the atomic charges do not necessarily coincide with the grid point, the atom's charge is allocated to the surrounding grid points such that the closer the atomic charge is to the grid point, the greater the proportion of the total charge that grid point receives. The boundary between the solvent and the solute is defined by the SASA; all grid points within this boundary are assigned the dielectric constant of between 2 and 4, and all outside are assigned a dielectric constant of 80. The electrostatic potential is then determined by a finite difference formula which sums the potentials and dielectrics of surrounding grid points. However, the potential at each grid point influences the calculation of neighbouring grid points, so by repeating the calculation again with the updated charges, the value changes. Therefore this calculation is re-iterated until the values converge.

This finite-difference Poisson-Boltzmann method is used to calculate the electrostatic component of the solvation free energy by performing two sets of calculations using the same grids but with dielectrics set at 80 and then set at 1. These represent the electrostatic energy in solvent and in vacuum respectively. Subsequently, the difference between these values is indicative of the change in electrostatic energy in transferring from vacuum to solvent:

$$\Delta G_{pol}^{sol} = \frac{1}{2} \sum_{i} q_i \left(\phi_i^{80} - \phi_i^{1} \right) \tag{2.20}$$

where the electrostatic energy of a charge q_i in a potential ϕ_i equals $q_i \phi_i$. This summation is over all charges in the solute.

An alternative technique for calculating the polar contribution to the solvation free energy is the **generalised Born** method. This is an approximation to the PB method described above, and represents the solute molecule as a discrete set of overlapping spheres with a point charge in the middle. As with PB, the solute is imbedded in an

implicit polarisable dielectric continuum [39]. The ΔG^{sol}_{pol} is then calculated as:

$$\Delta G^{sol}_{pol} = \underbrace{\frac{1}{2}\sum_{i \neq j} \frac{q_i q_j}{r_{ij}}}_{vacuum} - \underbrace{\frac{1}{2}\left(1 - \frac{1}{\epsilon_w}\right)\sum_{i,j} \frac{q_i q_j}{f^{GB}\left(r_{ij}\right)}}_{solvent} \tag{2.21}$$

where ϵ_w is the dielectric constant of the solvent - set to 80; q_i and q_j are atomic charges; f^{GB}_{ij} is a smoothing function which depends on the inter-atomic distances and atomic radii. The first term is the energy in a vacuum, from which is subtracted the solvation energy term [49, 94].

If the GB method is implemented with the SASA method of calculating the non-polar solvation free energy, and the molecular mechanics method of calculating the internal energies of the components, the methodology is called MMGBSA. If, instead of the GB method, the PB method is implemented to calculate the electrostatic component of the solvation free energy, the methodology is called MMPBSA. Both were implemented in this thesis through the MMPBSA module of the AMBER9 software suite. Solvated ΔG_{bind} values were calculated across a trajectory for snapshots every 10ps apart. As the NAMD protocol (Section 2.1.2) ran for 1ns at a time, outputting 1000 configurations every 1ps over this time, 100 snapshots over the 1ns were analysed by the MMPBSA module. The 100 outputted ΔG_{bind} values were averaged to attain a mean value over the 1ns. The molecular mechanics energies ($\Delta G_{bind}(v)$) were calculated using the SANDER module of AMBER9 with no cut-off for the non-bonded interactions. The PBSA module was used to calculate both the PB and GB ΔG^{sol}_{pol} term with a grid-spacing of 0.5Å, the intramolecular dielectric set to 1, and the continuum dielectric set to 80. The potential was calculated after 1000 iterations. The MSMS module was used to calculate the ΔG^{sol}_{np} term, with a probe radius of 1.4Å used to determine the SASA, the surface-tension set to 0.00542kcal/molÅ^2 and the off-set set to 0.92kcal/mol [49, 94]. Furthermore, to decrease the computational requirements, only a single simulation was performed, rather than the three for each of receptor (which, in this thesis, was the protein HIV-1 protease), ligand and complex. The configurations for the separate protein and ligand were extracted from the complex simulations by deleting all unnecessary molecules. So to attain the ligand configuration, the waters and protein

were deleted. This methodology, while decreasing the required number of simulations by one-third, makes the assumption that there is no change in configuration upon complex formation by either the protein or the ligand. This is a fair assumption for this thesis because saquinavir is an inflexible molecule [124], and the protease's active site region is considered to be relatively static compared to the rest of the molecule [77].

2.3.3 Configurational Entropy

While the MMPBSA and MMGBSA methodologies include the change in configurational entropy of the solvent with the ΔG_{cav}^{sol} term of the non-polar contribution to the solvation free energy, they do not include the change in free energy due to the decrease in configurational entropy of the ligand and receptor upon complex formation. When a ligand binds to a receptor, both components undergo a decrease in configurational entropy as the non-bonded forces between the two restrict their degrees of freedom. This change in configurational entropy upon complex formation can be sub-divided into three components:

$$\Delta S_{conf} = \Delta S_{tra} + \Delta S_{rot} + \Delta S_{vib} \tag{2.22}$$

where ΔS_{tra} is the change in translational degrees of freedom; ΔS_{rot} is the change in rotational degrees of freedom; and ΔS_{vib} is the change in vibrational degrees of freedom. The free energies associated with each of these freedoms can be calculated according to the equations below [98]:

$$G^{tra} = \frac{3}{2}RT - RT\left[\frac{5}{2} + \frac{3}{2}ln\left(\frac{2\pi m k_B T}{h^2}\right) - ln(\rho)\right] \tag{2.23}$$

$$G^{rot} = \frac{3}{2}RT - RT\left[\frac{3}{2} + \frac{1}{2}ln\left(\pi I_A I_B I_C\right) + \frac{3}{2}ln\left(\frac{8\pi^2 k_B T}{h^2}\right) - ln(\sigma)\right] \tag{2.24}$$

$$\begin{aligned} G^{vib} &= \sum_{i=1}^{3N-6}\left[\frac{1}{2}h\nu_i + \frac{h\nu_i}{e^{h\nu_i/k_B T}}\right] \\ &- \sum_{i=1}^{3N-6}\left[\frac{h\nu_i}{e^{h\nu_i/k_B T}} - RTln\left(1 - e^{-h\nu_i/k_B T}\right)\right] \end{aligned} \tag{2.25}$$

These equations contain many constant terms that will not be described here. To calculate the change in free energy associated with the loss of rotational, translational and

vibrational entropy upon inhibitor binding, each of these three equations is applied to the three component molecules and then their difference calculated as shown in Equation 2.14. Of particular note is the v_i term in Equation 2.25, which is the vibrational frequency of the **normal mode**. The normal modes of a linear molecule such as a protein are the concerted set of harmonic motions that the covalently linked atoms perform. For example, the core backbone of a protein is the chain $[-C-C-N-]_n$. Though there are other atoms and side-chains that branch off this chain, these are ignored for this illustration. As the central carbon atom vibrates, it moves towards the nitrogen atom. As its proximity increases, repulsion effects cause the carbon atom to move away from the nitrogen atom, which in turn moves away. This carbon atom is now traveling in the opposite direction towards the other carbon atom and the same process happens. This results in all the atoms harmonically vibrating back and forth. The frequency of the system's harmonics is termed its normal mode. However, this calculation assumes the protein's oscillations occur within a single potential energy well, which is not a valid assumption when performing the analysis over a trajectory - as the trajectory proceeds, the protein's conformation changes which regularly results in movement across the configurational landscape to another potential energy well. This results in the protein's normal modes being different, which therefore alters the calculation of the vibrational degrees of freedom. For this reason, protein simulations show high vibrational entropy variability.

Calculation of the change in free energy of binding associated with the change in freedom was performed using the NMODE module of the AMBER9 software package. This calculates the vibrational, rotational and translational entropy as per Equations 2.23, 2.24 and 2.25, then sums these values to attain a G^{conf} for each of the ligand, protein and complex. The sum of the ligand and protein values is then subtracted from the complex to attain a $-T\Delta S^{conf}$ term which can be subtracted from the ΔG_{bind} term calculated through MM(PB/GB)SA to attain a more accurate change in free energy of binding that includes the reduced configurational entropy. An important note is that the NMODE program performs it own conjugate-gradient minimisation (see Section 2.1.2 for more details) of the inputted configuration to ensure that the structure, over which the normal mode analysis is performed, is at the bottom of the potential

energy well. For this thesis, the minimisation was performed until the gradient of the surrounding potential energy landscape was less than 10^{-4} kcal/molÅ. This means that the outputted normal mode value is not necessarily the same structure that the MMPBSA is performed on. Therefore caution must be taken when comparing the MMPBSA output against the NMODE output. This minimisation step also makes the NMODE calculation much more computationally-demanding than the MMPBSA calculation, and for this reason each 1ns trajectory was only entropically analysed every 50ps. The average of these 20 snapshots gave a mean $-T\Delta S$ value over the nanosecond, which was combined with the ΔG_{bind} value from MMPBSA to generate the final average free energy change upon complex formation over that nanosecond.

2.4 Supercomputing Resources

Running a molecular dynamics simulation on the 50,000 atom solvated HIV-protease system over nanoscale time-lengths, and the subsequent analyses performed, requires considerable computational resources. Running this system for 1 nanosecond using NAMD would take approximately 256 hours (10.67 days) if run on a single processor. Subsequent analyses would take approximately another 2 days. Under these circumstances, the turnaround of results for a 1 nanosecond simulation would take almost 2 weeks. This is unfeasible, especially if a comparison between two simulations was required. In this case, the equilibration protocol of 2 nanosecond followed in this thesis would take each simulation up to 3 nanoseconds, and would therefore take a total of over 2 months (64 days) if run on a single processor computer.

However, NAMD was designed to be highly parallelisable, and scale well if simultaneously run across multiple processors. High performance computing (HPC) resources, such as the Texas Advanced Compute Centre (TACC) in the USA, and the HECToR machine in Edinburgh, each contain thousands of networked processors that facilitate this parallelisation. By running the 2 aforementioned simulations across 64 processors (32 for each simulation), the time taken to run the simulation is reduced to 24 hours (1 day). Due to the requirements of processors communicating with each other, the increase in scalability plateaus at 32 processors for the 50,000 atom HIV-1 protease

system. The sheer number of total processors in current HPC resources means that not only can each simulation be parallelised, but multiple simulations can be run simultaneously. In the final chapter of this thesis, 3200 processors were utilised across multiple resources to simultaneously run 100 simulations of 12ns in length. While this is a lot of processors, it was only a fraction of the 82,120 processors that these resources collectively contain. Furthermore, the development of 'grid-networks', which are networks of connected HPC resources such as TACC and the National Grid Service (NGS) in the UK, means that processor availability is no longer a limitation to the turnaround of results. Rather, the scalability of the analysis software is more limiting. For example, the PCA and NMODE analyses performed by AMBER9 are serially-implemented, which means that they can only run on 1 processor at a time. Therefore, while multiple NMODE analyses can be run simultaneously, they will all take approximately 1.5 days to calculate an average configurational entropy from 20 snapshots. Even more limiting is the time taken to transfer data generated by the HPC resources to a backup or to a local storage for analysis. Using the NAMD configuration parameters described in this thesis, each simulated nanosecond generates 500Mb of data. As 'SCP' and 'SFTP' transfer data at approximately 200Kb/s, it takes around 45 minutes to transfer 1ns' worth of trajectory data. This was especially noticeable when transferring the 100-repetition 10ns ensemble data back to a local storage server from machines at TACC; while it only took approximately 4 days to generate all the data, it took over 2 weeks to transfer the 500Gb of data back.

Chapter 3

Development of a local relational database collating biochemical and structural data on HIV protease

3.1 Introduction

In order to investigate the use of molecular dynamics as a tool for calculating enzyme-inhibitor binding affinities, it was first necessary to acquire relevant experimentally-derived biochemical and biophysical data, and link where possible to genotypic and structural data to allow the selection of appropriate structures for molecular dynamics simulations. Atomic coordinates of proteins attained from X-ray crystallography and nuclear magnetic resonance methods are publicly archived in the online depository, the Protein Data Bank [87]. As of the 11th February 2009 this depository contained 55,795 structures, of which 174 were of HIV-1 and HIV-2 protease enzymes bound by substrate peptides or inhibitors.

One aim of the research was to compare the computationally-derived Gibb's free energies attained from the molecular dynamics simulations to the experimentally-derived energies seen *in vivo*. An online resource for enzyme-inhibitor kinetic data, such as

ΔG, ΔH and K_i values exists in The Binding Database, which contains data attained through Enzyme Inhibitor Assays and Isothermal Titration Calorimetry methods [8]. As with the Protein Data Bank, there are no restrictions on who may deposit data, nor are there restrictions on how the data is deposited. As a result there is no standardisation of the database; some entries contain logK values while others do not, some entries have uncertainties greater than 20% in their experimental values, some entries have IC_{50} values while others do not, and few assays are performed under standardised temperatures and pH values. A local database was designed and created that collated the information contained in the structural and biochemical databases such that database queries could be employed to extract relevant information. Sequences in the PDB are indexed according to an internal indexing system termed a 'PDB Identifier', rather than by the protein sequence, so cross-correlating the related biochemical data stored in the BindingDb to the relevant sequences in the PDB is indirect and requires comparison of the sequence genotypes. By organising the database in such a way that sequences are standardised into single-letter amino acid code and cross-mapped through an internal identifier, the relevant biochemical and biophysical data for a 'sequence of interest' with a starting structure can be extracted with a simple SQL query statement.

3.2 Populating the local relational database

HIV protease data was obtained primarily from the Protein Data Bank for atomic coordinates of protease from which to run MD simulations from, and the biochemical data from the Binding Database for protease-inhibitor kinetic data such as IC_{50}, K_i and ΔG. Also, experimental data published in journals provided additional kinetic data; however, as with the online depositories, the kinetic data is not presented in a standardised format. For example, enzyme-inhibitor kinetic data is presented as IC_{50} or IC_{95} values; as EC_{50} or EC_{95} values; as ΔG values; as K_i values; or in the context of mutant protease-inhibitor experiments, as fold changes with respect to their wild type sequence. These data are valuable and were converted into a comparative value to those calculated from the MD simulations. MMPBSA and normal mode analyses were employed to calculate the binding affinity between the protease and its inhibitor in an MD simulation; this outputs a change in free energy upon formation of the complex

($\Delta G_{complex}$). Therefore, the biochemical data were either converted into a ΔG value, or if presented as fold changes with respect to a specific sequence, the respective calculated ΔG values were converted into a comparable fold change. Equation 1.26 was followed to convert a K_i value into a ΔG value, and if the data were presented as $IC_{50/95}$ values, Equation 3.1 was employed to convert them to K_i values, which could then be converted into a ΔG value. However, this required knowledge of the substrate concentration and the Michaelis constant (K_M) of the reaction, which was not always available.

$$K_i = \frac{IC_{50}}{\left(1 + \frac{[S]}{K_M}\right)} \tag{3.1}$$

These conversions to ΔG values were performed manually on the data when necessary to compare to a computed value. The data entered into the local database was therefore exactly as deposited in the online depositories, and the calculations performed after extraction of the data.

Before the SQL database could be populated, the structure of the component tables needed to be determined to ensure that the maximum information could be extracted with the simplest query statement. Figure 3.1 shows the structure of the tables and their relationship to each other in the form of a UML diagram. A description of each of the tables in the database along with their relationship to each of the other tables is detailed below.

- **Sequence**. This table is the reference point for the whole database. Each table element is a unique sequence with an associated identifier which is used to reference all the relevant data for that sequence. In this way, the protease sequence of any biochemical or crystallographic data that is added to the database is checked against the sequences already present in the *Sequence* field, and if a match is found the data is added with the associated *Sequence_ID* identifier. If the protease sequence of the new data is unique, then the sequence is added as an element in this table and given a unique identifier equal to N + 1, where N is the previous number of elements in the table. The associated data is then inserted into the appropriate table with a *Sequence_ID* field whose value corresponds to that given to the protease sequence in this table.

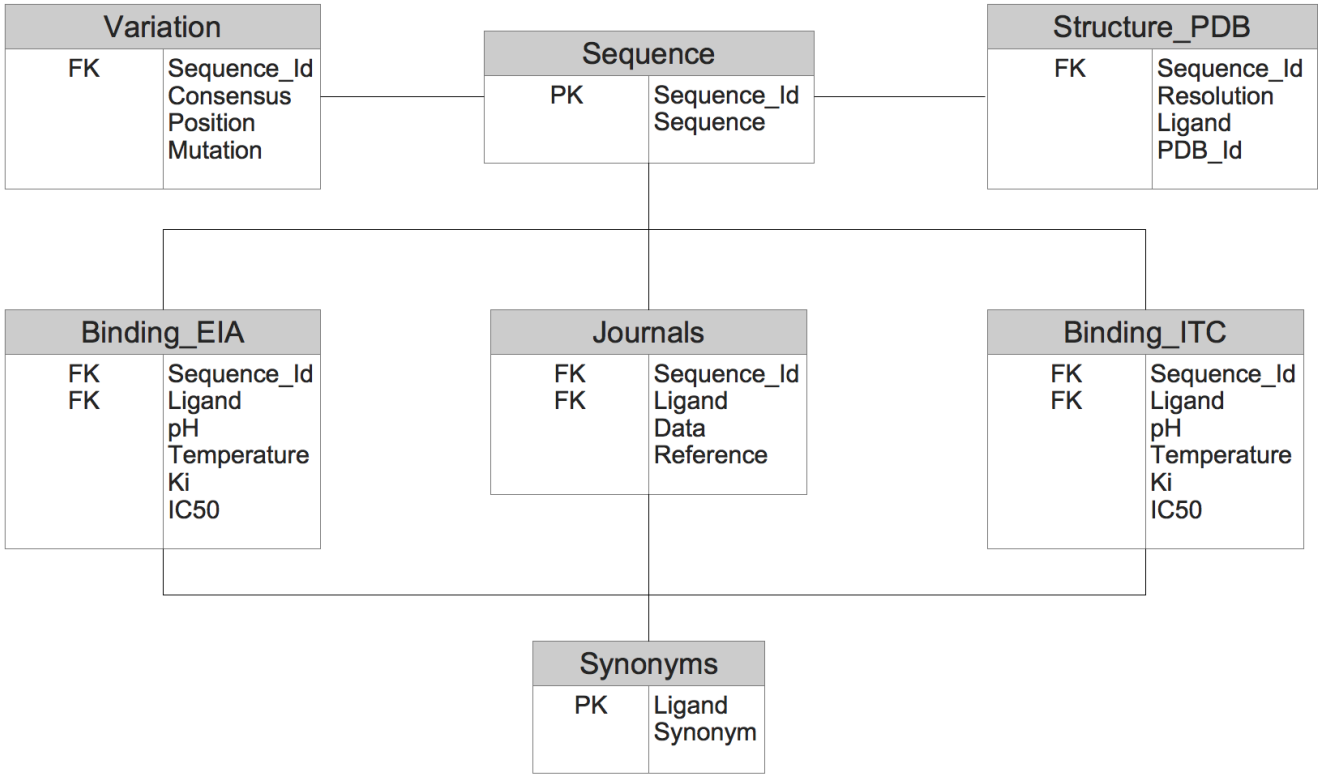


Figure 3.1: UML diagram representing the relationships between tables in the local database created to collate publicly-available biochemical and biophysical information on HIV-1 and HIV-2 protease complexed with inhibitors and peptide fragments.

- **Structure_PDB**. This table contains data extracted from the HIV-1 and HIV-2 protease PDB files. The fields in this table are *Sequence_ID*, *Resolution*, *Ligand*, and *PDB_Id*. The *Sequence_ID* field is directly associated with the field of the same name in the **Sequence** table, so that a simple SQL query could return the associated sequence with the PDB data. The *Resolution* field contains the resolution of the X-ray crystallographic diffraction data, indicating the quality of the atomic coordinates in the file. The *Ligand* field indicates the complexed molecule in the file, if present. It is important to note that *Sequence_ID* element values are not unique; each identifier does not appear just once in this table because proteases with the same sequence but different complexed molecules will each have an element in this table, and will have the same *Sequence_ID* value, but their other fields will differ. The *PDB_Id* field contains the identifier code used by the PDB depository to locate the file; each deposited file has a unique identifier that differentiates it from other files. This field is important because the atomic coordinates contained in the PDB files are not stored in the local database, and are manually extracted from the files in the PDB depository to attain the atomic coordinates. This is described in more detail (Section 3.3) along with example SQL queries.

- **Binding_EIA**. This table contains the biochemical information on HIV-1 protease attained through the Enzyme Inhibition Assay experimental method and archived in The Binding Database depository. The fields in this table are *Sequence_ID*, *Ligand*, *pH*, *Temperature*, K_i, and IC_{50}. As with the **Structure_PDB** table, the value of the *Sequence_ID* field corresponds to the unique *Sequence_ID* value of the associated protease sequence in the **Sequence** table. This allows for more complex SQL queries to be designed that can extract both crystallographic and biochemical data of a particular sequence from a single query, allowing for a broad overview of the available data much more easily than having to manually locate it within the various depositories. The *Ligand* field is the same as in the **Structure_PDB** table - it contains the name of the complexed molecule in the experiment. It is important to note that the name of the ligand entered into the database is not necessarily the same as in the original file. This

is because at the time of database generation there was no standardisation of nomenclature when depositing data into the online depository, so different authors gave different names to ligands depending on their preference; for example, for the HIV protease inhibitor saquinavir, some entries gave the drug name, others gave the chemical IUPAC name (2S)-N-[(2S,3R)-4-[(3S)-3-(tert-butylcarbamoyl)-3,4,4a,5,6,7,8,8a-octahydro-1H-isoquinolin-2-yl]-3-hydroxy-1-phenylbutan-2-yl]-2-(quinoline-2-carbonylamino)butanediamide, and others gave the drug's empirical formula $C_{38}H_{50}N_6O_5$. The Binding Database has since improved the layout of its entries, with the main ligand name given as the popular name, and with synonym fields for alternative names. However the **Synonyms** table was used in conjunction with specific Perl scripts, used to automate data entry, to standardise the ligand names in the *Ligand* fields of both the **Binding_ITC** and **Binding_EIA** tables when the database was created.

The *pH* and *Temperature* fields contained the pH and temperature, respectively, that the EIA experiments were performed under. These fields ensured that biochemical data compared between different protease sequences were performed under similar conditions. A change in either pH or temperature can affect an enzyme's biophysical properties, so it would not be possible to determine whether a difference in K_i was due to the difference in conditions or the difference in sequence, when comparing results from experiments performed under different conditions.

The K_i and IC_{50} fields contain the resultant data from the EIA experiments, but not every entry contains both K_i and IC_{50} values. As described earlier in this chapter, these values can be converted to ΔG and then directly compared to the ΔG values calculated from MD simulations. The result is the ability to create an SQL query that extracts both the PDB file identifiers and biochemical data associated with a particular protease sequence-inhibitor complex, so that the MD simulation can be run from the atomic coordinates in the PDB file and the results compared against the experimental data.

- **Binding_ITC**. This table's fields and purpose are the same as those of the **Binding_EIA** table. The only difference is that the populated data is generated

through the Isothermal Titration Calorimetry experimental method instead of the Enzyme Inhibition Assay method.

- **Variation**. This table contains each sequence's mutations from the experimental wild-type sequence Hxb2. The fields are *Sequence_ID*, *Consensus*, *Position* and *Mutation*. As with all the other tables, the *Sequence_ID* field provides a link to the full sequence it originates from. Each entry in this table represents a single mutation with respect to the experimental consensus sequence Hxb2. For each new sequence added to the **Sequence** table, each amino acid is compared to the Hxb2 sequence, and if a mismatch is found, an entry is added to the *Variation* table containing the Sequence_Id of the new sequence in the **Sequence** table along with the 1-letter code of the consensus amino acid, its position in the sequence, and the 1-letter code of the mutated residue. These are added to the *Consensus*, *Position* and *Mutation* fields respectively. Each difference to consensus sequence found in this way is added to this table, resulting in each Sequence_ID being represented in this table equal to the number of mutations in the sequence. For example, if a new PDB file is added to the database, the first step is to take the protease's sequence and compare it against all the other sequence entries in the **Sequence** table. If there is no match, then it is added to the table and the relevant data added to the **Structure_PDB** table with the corresponding Sequence_ID. The sequence is also compared to the Hxb2 consensus sequence residue-by-residue. If a mismatch occurs, then an entry is created in the **Variation** table; each subsequent mismatch generates an additional entry in this table. An example of the entries from a sequence with two mismatches to Hxb2 is shown in Table 3.1.

Table 3.1: Example output of **Variation** table for a sequence with 2 mutations from Hxb2

Sequence_ID[a]	**Consensus**	**Position**	**Mutation**
1	G	48	V
1	L	90	M

[a] Internal identifier. Number shown only for example purposes.

The purpose of this table is to be able to search through the collated data with respect to the number of mutations a sequence has from the experimental consensus sequence. For example, in Section 5.2, this was used to identify a series of thirteen protease-saquinavir complexes for simulation; each simulated complex contained one extra mutation compared to the previous one. So the first simulated-complex had a single mismatched residue to the Hxb2 sequence, the second simulated-complex had two mismatched residues, and the thirteenth complex had thirteen mismatched residues. By creating a table containing the mutations within each sequence located in the database, an SQL query can be created that outputs the PDB identifiers associated with sequences containing N number of mutations that also have experimentally-determined biochemical data in the database. This will be described in more detail in Section 3.3.

- **Synonyms**. As described above in the **Binding_EIA** table, the purpose of this table was to contain the synonyms for each of the complexed ligands in data attained from both PDB and The Binding Database depositories. The fields of this table are *Ligand* and *Synonym*. The *Ligand* field contains the 'common' name for the ligand, for example indivavir, saquinavir or ritonavir. The *Synonym* field contains alternative nomenclatures. Each entry in this table corresponds to a single synonym, so ligands have multiple entries determined by the number of different synonyms used. This table was used primarily through the Perl scripts written to automate population of the database.

- **Journals**. This table contains the experimentally-derived biochemical data published in journals rather than archived in online depositories. This table is less populated than the other tables due to the manual task of data acquisition and entry, compared to the automated process for the other tables. This table has the fields *Sequence_Id*, *Ligand*, *Data* and *Reference*. The *Sequence_Id* and *Ligand* fields functions the same as in the other biochemical-data tables, while the *Data* field contains the published data for that protease-ligand complex. Due to the various different ways in which authors publish their results (fold changes from Hxb2, fold changes from their designated wild-type, K_i values, $EC/IC_{50/95}$ values, or ΔG values) this field is not uniform; the populated data is the same as the

published data.

The collated data populating this database was substantial enough that manual entry was unfeasible; at the time of database creation there were approximately 230 HIV protease structures in the Protein Data Bank, approximately 100 HIV protease-ligand complexes analysed by Isothermal Titration Calorimetry in The Binding Database, approximately 2,500 complexes analysed by Enzyme Inhibition Assay in The Binding Database, and data for approximately 50 protease-inhibitor complexes in journals. Therefore Perl scripts were written to automate as much of the database population as possible, and to ease subsequent population when additional data is made available. A description of the Perl files is given below:

- **pdb_data_extractor.pl** was written to extract the important information from PDB files and write it to a text file in a standardised output, as shown in Figure 3.2. The reason for writing to a text file instead of immediately populating the database is so that the data can be manually checked for any irregular data; due to the non-standardised entry of data in PDB files, authors do not use the same field tags for information.

  Before the script could be run, a directory containing all the HIV protease PDB files was manually created by downloading each file from the online depository. Once done, the Perl script iterates over each file in the directory, and for each one it performs a regular expression pattern match to pull out the lines containing the PDB identifier, the ligand's name, the sequence and the resolution. The lines are split to extract the relevant information, which is then outputted to a text file and delimited by tabs. Before the ligand name is outputted to the text file, it is compared to the *Synonyms* field of the **Synonyms** table in the database with the following SQL query:

  SELECT Ligand FROM Synonyms WHERE Synonyms='$extracted_ligand_name'

  If a match is found, then the ligand name pulled from the database is entered into the text file instead of that in the PDB file. This is to ease analysis of the database when all data associated with a particular ligand is required; by creat-

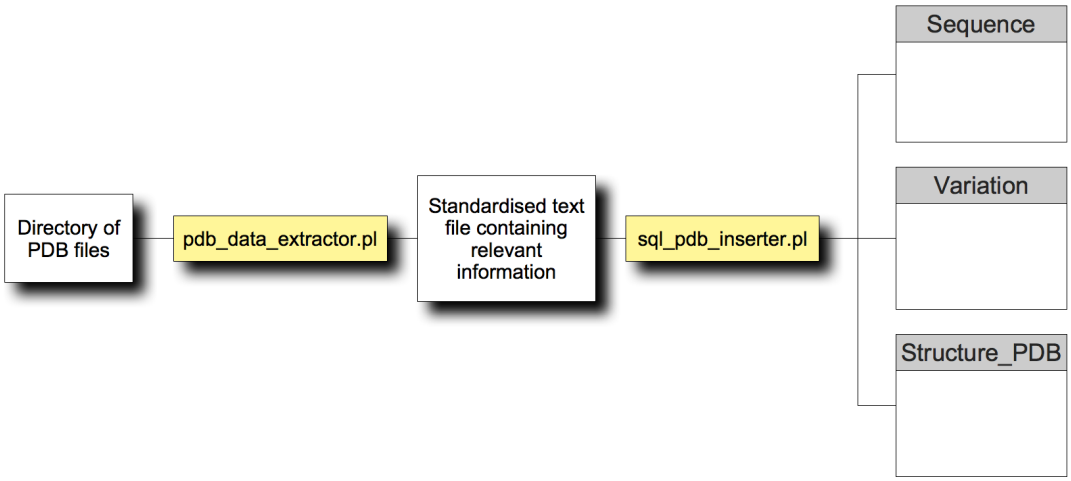


Figure 3.2: Flow diagram showing how the Perl scripts are involved in the population of the local database. A directory containing all the raw HIV protease PDB files is analysed by the pdb_data_extractor.pl Perl script, which outputs a text file containing tab-delimitated data to populate the local database. This text file is then analysed by a second Perl script, sql_pdb_inserter.pl, which converts the tab-delimitated data into the relevant SQL insert statements to populate the **Sequence**, **Variation** and **Structure_PDB** tables. A flow diagram showing the action of these scripts in more detail is shown in Figures 3.3 and 3.4.

ing an SQL SELECT query to standardise the population of the database, the SELECT queries required to extract the data from the database will be much simpler.

Furthermore, before the sequence is added to the text file, it is converted to a 1-letter code in the Perl script. This is achieved by having a hash variable whose keys are the 3-letter codes for each of the 20 amino acids, and whose values are the 1-letter codes, as shown below:

%amino_ acids = (
'Gly' => 'G'
'Tyr' => 'Y'
...)

Using this hash variable, each 3-letter amino acid can be converted to its 1-letter equivalent before being outputted to the text file. The actions of this script are

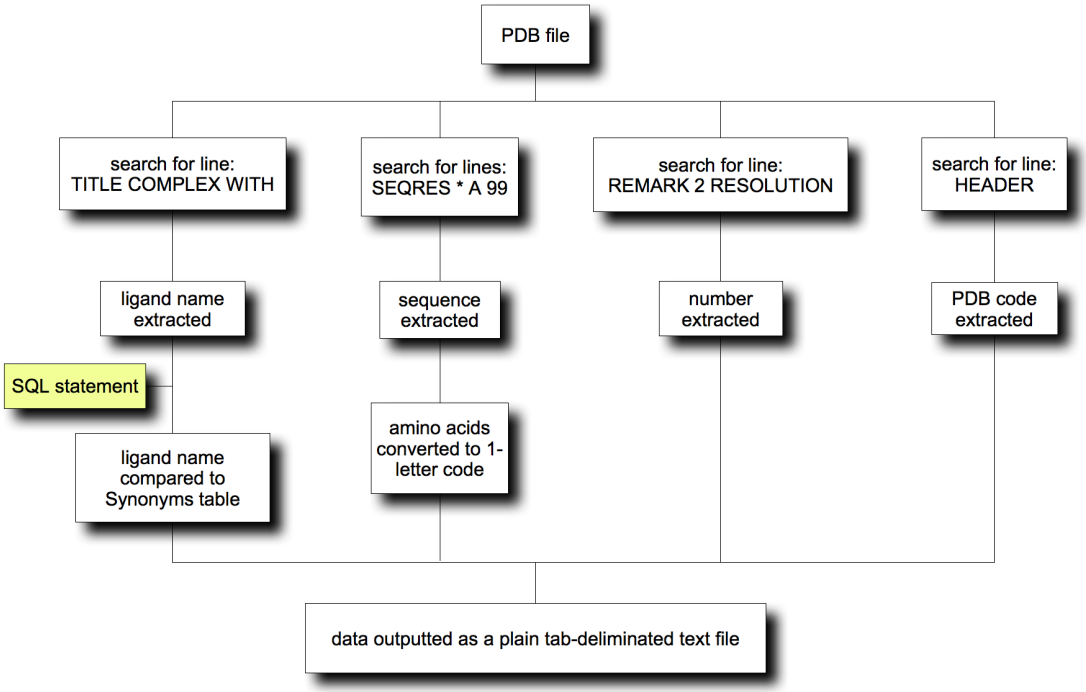


Figure 3.3: Flowchart showing the sequential actions of the Perl script *pdb_data_extractor*. This script requires as an input a directory containing raw PDB files downloaded from the online PDB depository. It outputs a text file containing the PDB identifier, the ligand, the sequence in 1-letter amino acid format, and the resolution of the crystallographic data. These data are tab-delimited, with each line comprising of data from a single PDB file. The context that this script operates in is shown in Figure 3.2.

shown diagrammatically as a flow chart in Figure 3.3.

- **sql_pdb_inserter.pl** is the second Perl script written to populate the database with data from the PDB files. It acts in concert with the *pdb_data_extractor.pl* Perl script, which is described above (Figure 3.2).

  The script takes as its input a text file containing tab-delimited data extracted from multiple PDB files by the *pdb_data_extractor.pl* script. For each line in the text file, this script splits the data and stores them into relevant variables so that it can generate SQL INSERT statements with the data to populate the database. This script also has a checking mechanism to ensure that data is not redundantly being entered into the database. Once the script has split up the data in the inputted text file, it generates the following SQL statement:

  *SELECT * FROM Structure_PDB WHERE PDB_Id='$pdb_id'*

  This checks to see if the PDB identifier is already in the database. If a match is found, then the data in that line is disregarded because the PDB identifier is unique and so the associated data must already be present in the database. If no match is found then the sequence is compared to the *Sequence* field of the **Sequence** table, using the following statement, to determine whether it is a novel sequence:

  SELECT Sequence_Id FROM Sequence WHERE Sequence='$sequence'

  If a match is revealed, then the *Sequence_Id* value is extracted and used with the remaining data to generate an SQL INSERT statement to enter the data into the **Structure_PDB** table of the database.

  INSERT INTO Structure_PDB (Sequence_Id, PDB_Id, Ligand, Resolution)
  VALUES ('$seq_id', '$pdb_id', '$ligand', '$resolution')

  If no *Sequence_Id* match is found then the sequence is not already present in the database. Therefore the sequence is added as a novel sequence with a unique

identifier, using a simple INSERT statement similar to those given above. In addition, the sequence's mutations with respect to the experimental consensus sequence (Hxb2) are determined. This is achieved by hard-coding the Hxb2 sequence into the Perl script and then dissociating the inputted sequence into individual amino acids. Each amino acid in turn is compared to the associated amino acid of the consensus sequence, and if they differ then the consensus amino acid, position in sequence, and mutated amino acid are added to the database with the following statement:

INSERT INTO Variation (Sequence_Id, Consensus, Position, Mutation)
VALUES ('$seq_id', '$consensus', '$position', '$mutation')

Once entered, the script moves onto the next amino acid and carries on checking. Any subsequent mutations are added to the database in the same way, such that N rows in the table are inserted corresponding to the N mutations to Hxb2; each row has the *Sequence_Id* attribute either extracted from the **Sequence** table or generated upon sequence insertion to the table to associate it to the particular sequence. The actions of this scripts are shown diagrammatically in a flowchart in Figure 3.4.

- **sql_itc_inserter.pl** was written to populate the database with the Isothermal Titration Calorimetry data from The Binding Database online depository. The curators of this database were kind enough to provide a raw dump of their data on HIV protease, both from EIA and ITC experimental methods. This acted as the starting input for both this Perl script, and the *sql_eia_inserter.pl* script described below. Figure 3.5 shows how these two scripts act together to populate the database with The Binding Database data, and which tables in the local database are populated.

  As can be seen in Figure 3.5, the *sql_itc_inserter.pl* Perl script is a semi-automative script that requires user-input to populate the database. The user is given a choice of the manner in which to populate the database; either automatically through a file, or through script-directed manual entry. The reason for this is because The Binding Database has no unique identifier that distinguishes experimental

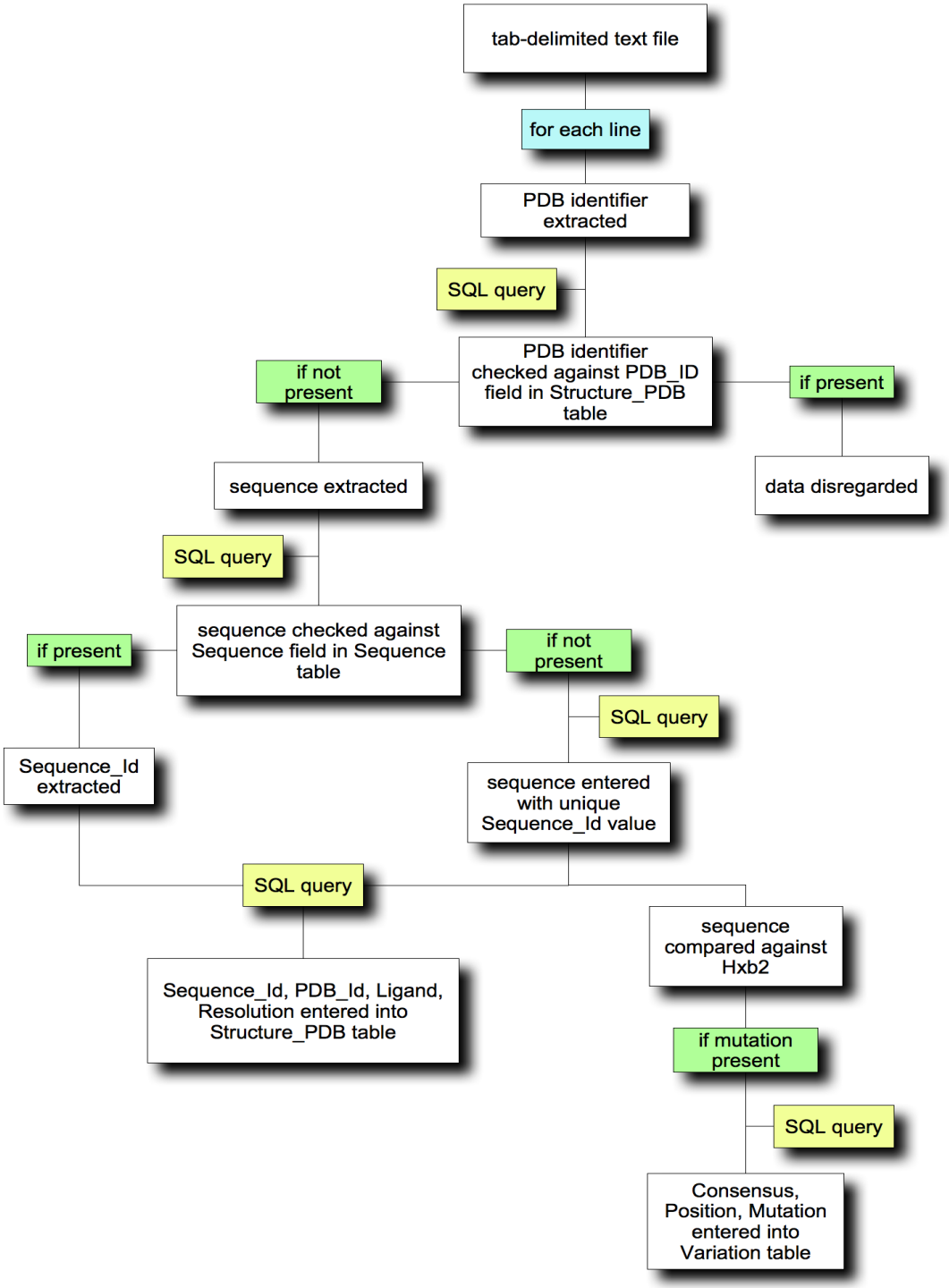


Figure 3.4: Flowchart showing sequential actions of the Perl script *sql_pdb_inserter*. The script takes as its input the tab-delimited text file outputted by *pdb_data_extractor* and checks whether the data is already present in the database. If not, it checks whether the sequence is novel, and either extracts the associated Sequence_Id value or generates one depending on the outcome. The remaining data is then inserted into the **Structure_PDB** table along with the Sequence_Id. A novel sequence is also compared against Hxb2 consensus sequence and the nature and position of each mutation inserted into the **Variation** table.

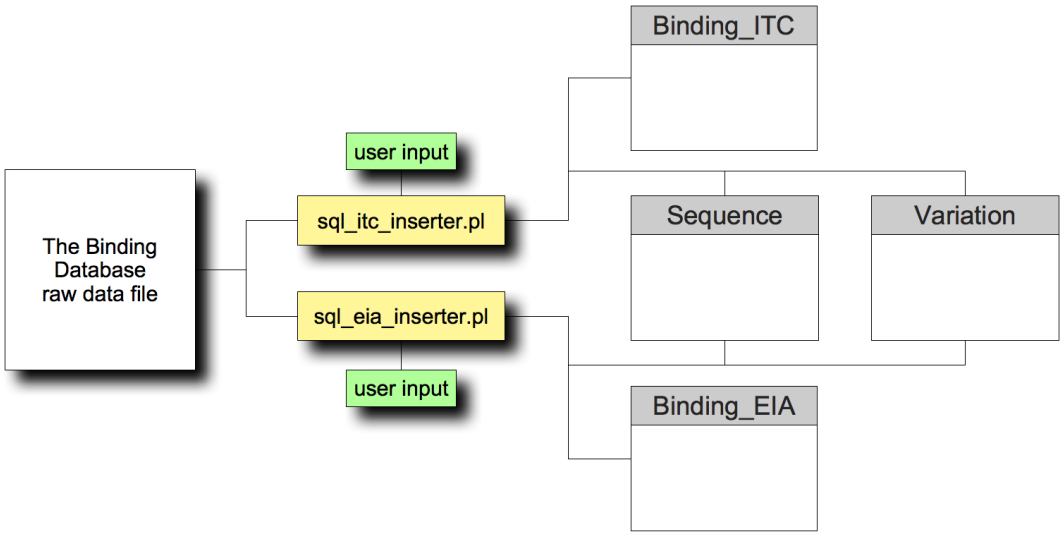


Figure 3.5: Flowchart showing how the data from The Binding Database populates the local database. The curators of TheBindingDb kindly gave us a file containing a dump of their HIV protease EIA and ITC data. The two Perl scripts *sql_itc_inserter* and *sql_eia_inserter* work in tandem to populate the **Binding_ITC** and **Binding_EIA** tables respectively. Any novel sequences are added to the **Sequence** table, and their mutations with respect to Hxb2 added to the **Variation** table. Individual entries can also be added to the necessary table through the script-directed manual-entry option of the associated script.

data from each other. Therefore it is not possible to determine whether two very similar entries are actually the same entry or two different entries that reinforce each other. As a result, the Perl script was written so that on initial population of the local database with the dump file, the user chooses the location of the file and the program automatically inserts the data into the database. Then on subsequent updates of the database, the user can choose to manually insert the data if there is only one or two entries to be added, or to automatically upload from a file if there are multiple entries to be added and they are sure that there are no duplicate entries. Figure 3.6 shows the actions of this script in more detail. The Binding Database data is dealt with by the script by the same method as *sql_pdb_inserter*: the sequence is extracted and compared against those in the **Sequence** table. If a match is found then the sequence's *Sequence_Id* is extracted from the database and combined with the ligand name, pH, temperature, K_i and IC_{50} data to update the **Binding_ITC** table as an SQL INSERT statement. If a match is not found for the sequence, then it is added to the **Sequence** table and its mutations with respect to Hxb2 added to the **Variation** table; the *Sequence_Id* given to the sequence is extracted and combined with the rest of the data as before.

When the Perl script is run, it first displays a simple command-line menu asking the user whether they want to enter data from a file or manually. If they choose to enter it manually then the program asks the user to enter the data with prompts. For example, it will ask for the ligand name and pauses until the user has entered it. Once the user enters it on the command line then it asks for the pH and pauses again, and so on. If nothing is entered, the script confirms that there is no data associated with that field and then proceeds without updating the associated variable. The SQL INSERT statement generated does not then contain the column or variable, and so the field in the table is left empty. The automatic method also performs these actions if there is a missing value in the file. This happens, for example, with K_i and IC_{50} values - usually either one or the other is given, leaving the other field empty.

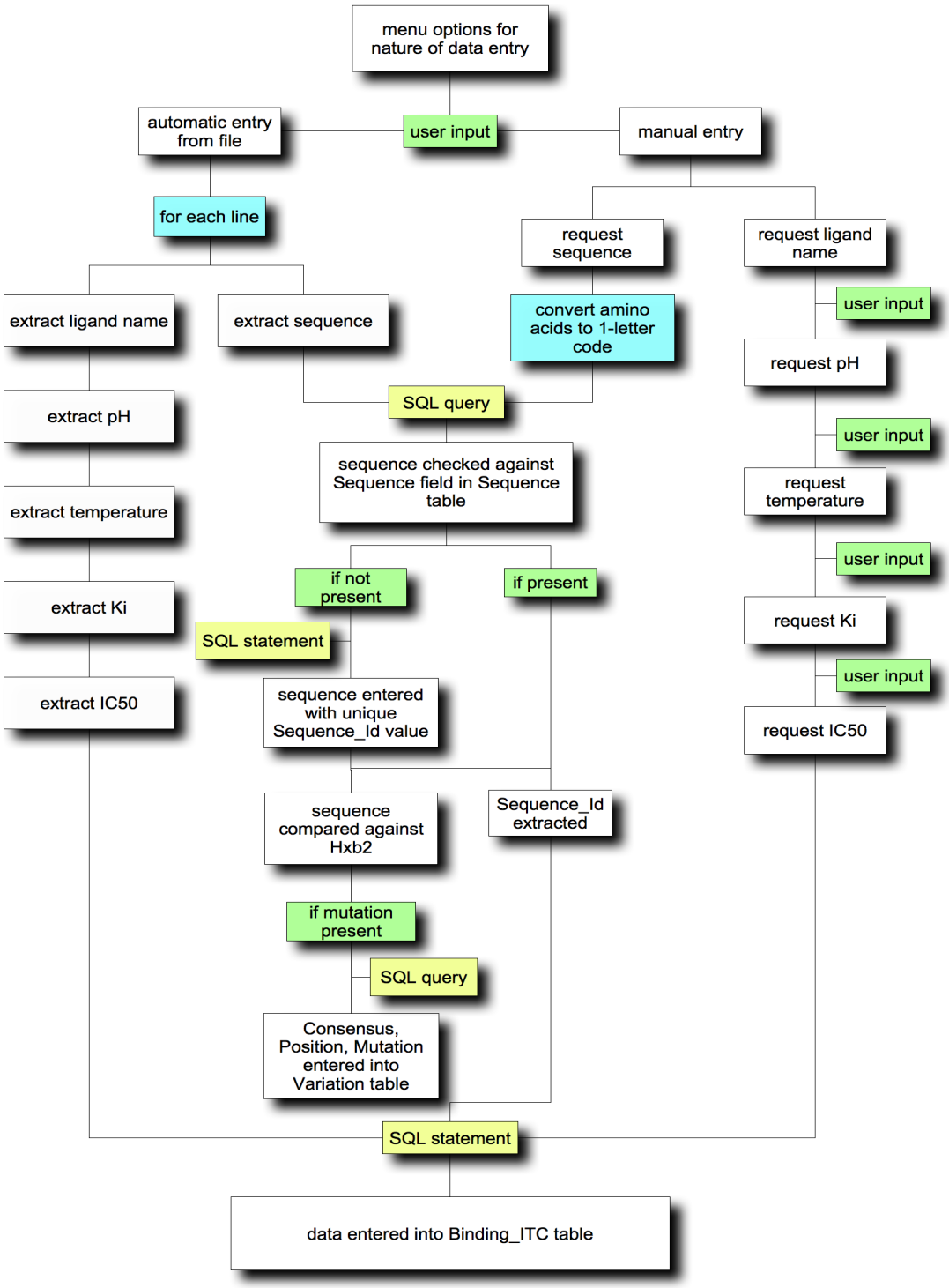


Figure 3.6: Flowchart showing the sequential actions of the Perl script *itc_sql_inserter*. The script takes as its input a file containing The Binding Database ITC information, which the script sorts and outputs to the associated field in the **Binding_ITC** table or, for the protease sequence, the **Sequence** and **Variation** tables. Novel sequences are treated in the same way as the *sql_pdb_inserter* Perl script. As the script has no mechanism by which to check if the inputted information is already present in the database, the script also allows for script-prompted manual entry of additional data.

- **sql_eia_inserter.pl** is almost exactly the same Perl script as *sql_itc_inserter*. The only differences are that it extracts the Enzyme Inhibition Assay data from The Binding Database's dump file rather than the ITC data. The data in the dump file was organised such that the first line just had 'EIA'. Following this was the data attained through EIA methods, which made up the majority of the file. Then there was a line with 'ITC' and following this was the data attained through ITC methods. Therefore the script only evaluated the lines between the lines 'EIA' and 'ITC', whilst the *sql_itc_inserter* script evaluated from 'ITC' to the end of the file. The second difference between these scripts, highlighted in Figure 3.5, is the tables in the database that the script outputted to. Instead of the populating the **Binding_ITC** table, the script generates SQL INSERT queries to populate the **Binding_EIA** table. The fields in these two tables are identical, so the script follows the same sequential actions as that shown in Figure 3.6, the only difference between these two scripts is that at the final box in the flowchart, the data is entered into the **Binding_EIA** table.

3.3 Extracting data from the local relational database

Having created a local depository containing both biochemical and crystallographic data on HIV protease, granting the ability to relate the biochemical data to the associated sequences of the crystallographic data, a series of Perl scripts were written to actualise this association. The reason for this is that although the biochemical and structural data are linked through the *Sequence_Id* field, SQL statements are still required to extract the related data. Therefore, Perl scripts were written to automate this extraction within user-defined parameters. For example, one of the Perl scripts will extract from the database, and display on the command-line, all of the associated biochemical data for the PDB identifier entered by the user. A detailed description of each of the Perl scripts is given below:

- **biochem_data_extractor.pl** was written to extract all of the biochemical data associated with the protease sequence of a PDB file. The reason for this is that molecular dynamics simulations use the atomic coordinates of a PDB file as the starting configuration, and so in order to compare the binding affinities calcu-

lated from the output of the simulation, this script will extract the corresponding experimental data that can be converted into an appropriate comparable value. When the script is run, it asks the user to input the PDB identifier they would like associative data for. The script then checks that the PDB identifier is present in the database, and if so it extracts the corresponding *Sequence_Id*. This is then used in the following SQL SELECT statement:

*SELECT * FROM Binding_EIA*
WHERE Binding_EIA.Sequence_Id = $seq_id
AND Structure_PDB.Sequence_Id = $seq_id
AND Binding_EIA.Ligand = Structure_PDB.Ligand

where *$seq_id* is the Perl variable containing the database's internal sequence identifier for the PDB file in question. This SQL statement extracts all the fields from the elements in the **Binding_EIA** table matching the *Sequence_Id* and containing the same inhibitor. This statement is then repeated for both the **Binding_ITC** and the **Journals** tables to extract all the biochemical data associated with the same sequence-ligand pairing. The script then prints the output of all three statements to the screen, or to a file if the user specifies a filename on the command-line.

The reason for restricting the extracted biochemical data to inhibitors matching that in the PDB file is because kinetic data from other inhibitors is not comparable to the kinetic data calculated from the protease-inhibitor complex in the MD simulation. The caveat to this is that the SQL statement formed by the script requires the ligand names to be identical in both tables; although the database-populating scripts have a measure of standardisation with respect to ligand name entry, some synonyms may still not be recognised, or may be misspelled. Therefore the user can choose to relieve this restriction, and so output all biochemical data associated with the *Sequence_Id*, through an option when running the script.

- **mutation_data_extractor.pl** was written to extract all the associated biochem-

ical data for a sequence N number of mutations away from a crystal structure. The reason for writing this Perl script was because although the PDB crystal structures provides a starting configuration for the MD simulations, there are too few distinct PDB sequences to cover the range of protease sequences observed clinically. Therefore it is necessary to computationally mutate a sequence away from the crystal structure prior to simulation and so this script allows the user to ascertain the validity of computationally-mutated MD simulations.

The script has two command-line arguments - the number of mutations from the baseline sequence, which is mandatory; and a PDB file identifier, which is optional. When the script is run, it first checks that the mandatory number of mutations has been given on the command-line. Once confirmed, it forms the following SQL SELECT statement:

*SELECT * FROM Variation*
GROUPBY Sequence_Id
HAVING COUNT(Sequence_Id) = $num

where *$num* is the Perl variable containing the number of mutations from the baseline sequence specified by the user. This statement extract all the elements from the **Variables** table that contain N number of rows with a specific *Sequence_Id* value. For example, in the database the sequence associated with the identifier *Sequence_Id=26* has two mutations with respect to Hxb2: I3V and N37S. Therefore the **Variables** table has two rows relating to this *Sequence_Id*, both of which would be extracted by the SQL statement if the user specifies two mutations when invoking the script. Any other sequences containing two rows in the **Variation** table would also be extracted, and outputted to either the screen or to a file if the user specifies so on the command-line. The nature of the mutations is also outputted along with the *Sequence_Id*, so an example output, if the user specified 4 mutations when invoking the script, would look like:

Sequence 12: I3V N37S G48V L90M
Sequence 47: I3V L10I N37S L90M

However, in addition to extracting the sequences with N numbers of mutations from the Hxb2 consensus sequence, it is also important to extract the sequences with N numbers of mutations from a crystal structure. For this reason the script takes a second optional command-line argument specifying the PDB identifier whose sequence is to act as the template sequence from which the number of mutations will be determined. If the second argument is given, the script forms the following SQL query to check that the PDB identifier is present in the database:

SELECT Sequence_Id FROM Structure_PDB WHERE PDB_Id = '$pdb_code'

where *$pdb_code* is the Perl variable containing the user-inputted PDB identifier. If the query returns no results then the script requests another identifier and repeats the query. If a result is returned, then the script takes the outputted *Sequence_Id* value and forms another SQL query to extract the sequence's mutations from the **Variation** table:

*SELECT * FROM Variation WHERE Sequence_Id = $seq_id*

where *$seq_id* is Perl variable containing the sequence identifier extracted from the previous SQL statement. The script stores each of the sequence's mutations from Hxb2 in an array. Then it forms the following SQL query to pull out all the *Sequence_Id* values present in the **Variation** table:

SELECT DISTINCT Sequence_Id FROM Variation

With the range of sequences stored in an array, the script iterates over each stored sequence in turn and extracts the mutations associated with it. These mutations are then compared against those in the PDB's sequence; if two opposite mutations are found (e.g. I3V in the PDB sequence, and V3I in sequence X) then these are cancelled out so that sequence X's number of mutations from the PDB sequence is one less than from Hxb2. If the PDB's mutations are not present within sequence X, then they are added such that sequence X's mutations from the PDB's sequence is N more than from Hxb2, where N is the number of mutations from Hxb2 in the PDB's sequence. The resultant mutations are then stored as

a value in a hash table as a concatenated string (e.g. I3V,N37S,L90M) with the *Sequence_Id* as its key. The number of mutations from the PDB's sequence is then entered into a second hash table, with the *Sequence_Id* as the key:

%mutations_pdb = (
1 => 4
2 => 15
...)

The script then compares the value in each of the elements in this second hash table against the mandatory user-determined number of mutations. If there is a match, then the script prints to the screen, or to a file of the user's specifications, the *Sequence_Id* associated with the match, along with the sequence's string of mutations extracted from the first hash table. Figure 3.7 shows the actions of this script as a flow diagram.

This collection of scripts, both for database population and data extraction, was developed over the course of the research in response to the needs of the current line of research.

3.4 Conclusions

There are presently no resources that collate together the structural data of proteins archived in depositories such as the Protein Data Bank, and experimental biochemical data on protein-ligand interactions archived in depositories such as the Binding Database. The local database described in this chapter was specifically designed to automate the retrieval of published structural and phenotypic data associated with a genotype-of-interest, with the aim of using the structural data to run molecular dynamics simulations and generate a biochemical value comparable to that in the database. This approach of associating phenotypic data, such as biochemical data, to high-throughput data, such as genotypic sequences, will become more necessary as advances in genome sequencing technology allow the characterisation of host and pathogen genetic variants at unprecedented levels [6, 119]. However, there is no standard measure for phenotypic assays; different assays give different measurements. For

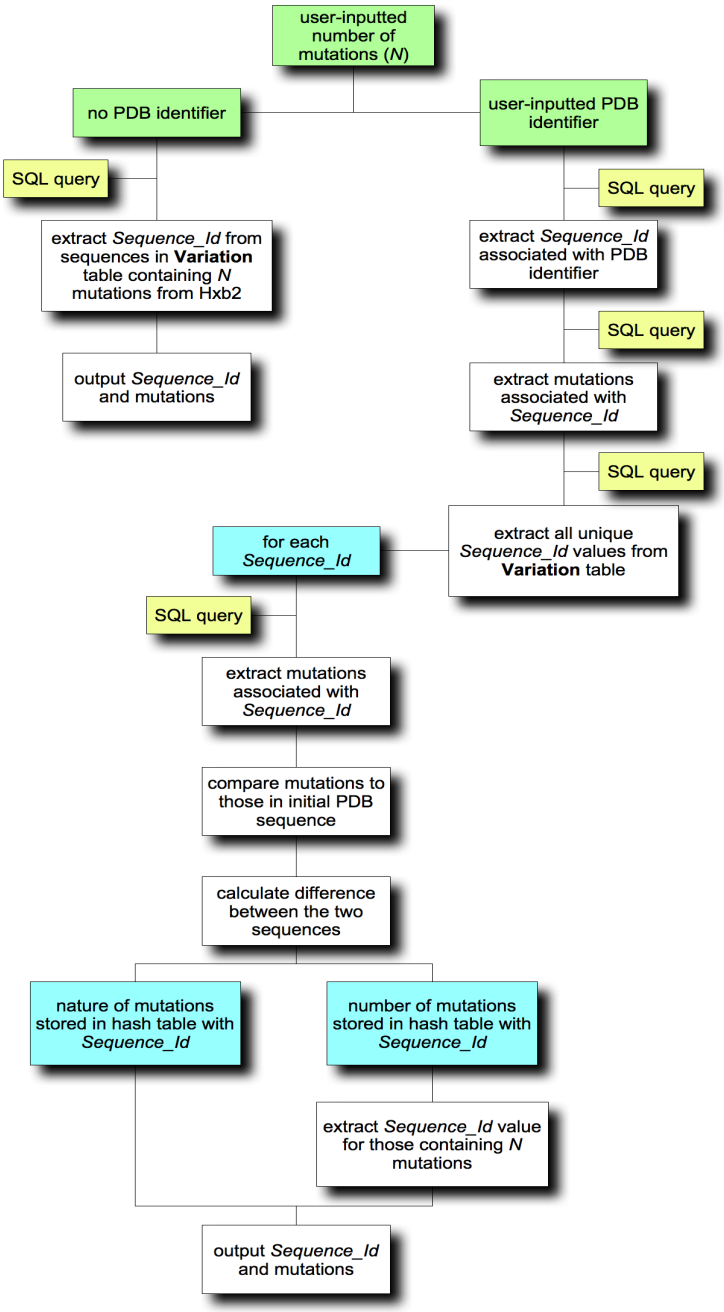


Figure 3.7: Flowchart showing the sequential actions of the Perl script *mutation_data_extractor*. The script takes as its input two arguments: a mandatory number of mutations, and an optional PDB identifier. If only the mandatory argument is given, the script extracts from the database all the *Sequence_Id* values of sequences with N mutations from the Hxb2 consensus sequence, and these sequences' associated mutations. If the optional PDB identifier is given, the script extracts all the *Sequence_Id* values of sequences with N mutations from the PDB's sequence, and these sequences' associated mutations.

example, drug resistance phenotype assays can result in measurements of IC_{50}, K_i or fold-change from a baseline sequence. While they all ultimately result in a measurement of the protein's sensitivity to a drug, these values cannot be directly compared against the ΔG value outputted by an MD simulation. The solution implemented in this research was to retain the measurements of the different biochemical assays in the local database, and convert them to ΔG values comparable to computational results when necessary.

The scripts written and implemented to ease population of the database and extraction of data result in a semi-automative method of maintaining and updating the local database. However, due to the non-standardisation of online depository data-formatting, it has not been possible to fully-automate this procedure. The manual proof-checking necessary for certain aspects of database population limits the scalability of the database. Expansion to include high-throughput data will require a standard nomenclature to be adopted and adhered to. Nevertheless, the database makes an important bridge between the computational assays which are playing more of a role in the post-genomic era, the experimental biochemical assays which are currently the gold-standard reference, and the structural data required to link the two.

Chapter 4

Structural comparison of proteases

4.1 Introduction

Currently, qualitative insights into proteins, such as HIV protease, are made through inferences from static structures attained from x-ray crystallography or NMR spectroscopy [109, 112]. For example, a study of 8 high-resolution crystal structures by Tie *et al.* (2005) proposed that two HIV-1 protease mutations V82A and I84V resulted in fewer *van der Waals* contacts and hydrogen bonds between the protease and clinical inhibitors, causing a reduced affinity between the two [112]. However, the stringent experimental conditions required to pack the HIV protease molecules into an ordered crystal complex may require the protein to adopt a unnatural configuration, which is unsuitable to make kinetic and dynamic structural inferences from. Mutations that confer reduced drug-sensitivity *in vivo* may not show altered phenotypes in a crystal structure because they either manifest in a protein's 'natural' configuration which is not conducive for crystallisation, or because they alter the *dynamics* between the enzyme and its inhibitor which will not be apparent in a static structure. For example, a molecular dynamics study carried out by Perryman *et al.* (2003) suggested a mechanism by which the V82F/I84V mutations contribute to drug resistance. Simulations over 22ns showed that the mutant's flaps were more flexible than the wild-type's, and spent a greater proportion of its time in a semi-open configuration. As a result, in-

hibitor binding would demand a greater enthalpic penalty in order to close the flaps. Therefore dynamic characterisation of proteins through computational methods such as molecular dynamics should usefully be applied as a novel analytical technique for the study of HIV drug resistance.

4.2 Comparison of static and dynamic structures

Research was initially concentrated on validating molecular dynamics as an analytical addition to x-ray crystallography and NMR spectroscopy by investigating whether two structurally-similar HIV-1 protease PDB structures would be shown to structurally-diverge into unique quaternary structures in a molecular dynamics simulation. The reasoning behind this is that the non-physiological conditions placed upon the protease enzyme during the crystallisation process combined with the restrictive number of crystallisable conformations able to be adopted by the protease would mean that two structurally-homologous crystallised HIV-1 protease enzymes would structurally-diverge through conformational relaxation in molecular dynamics.

Two HIV-1 protease structures (PDB identifiers 1HXB and 1BDQ) were extracted from the local database. These structures were chosen due to the large number of differences between their genotypes; 1HXB contained two mutations from the Hxb2 experimental consensus sequence, 1BDQ contained nine differences from Hxb2 consensus, and together 1HXB and 1BDQ were dissimilar at 10 amino acid positions (Table 4.1).

Table 4.1: Nature of 1HXB and 1BDQ's mutations with respect to Hxb2 consensus

PDB	**Mutations from Hxb2**[a]									
1HXB	I3V							N37S		
1BDQ		T31S	V32I	L33V	E34A	E35G	M36I	N37E	I47V	V82I

[a] Mutations presented as 'consensus'-residue-'mutation', *e.g.* I3V is isoleucine in Hxb2 at residue 3 mutated to a valine.

The reasoning behind selecting two structures with a large number of mutated residues between the two is that conclusions are made on static structures of multidrug-resistant

HIV protease enzymes as to how the mutations causes their effects on inhibitor-binding. Showing that two structures, with many mutations between the two, diverge into separate quaternary structures with little resemblance to the original crystal structure will highlight the potential for inappropriate conclusion of inferring the molecular basis of drug-resistance from a single crystal structure.

The two crystal structures were structurally-compared through profile RMSD (Section 2.2.1) of their C_α backbone atoms, providing an indication of the similarity of the proteins' quaternary structures (Figure 4.1). These RMSD graphs cover both monomers of the homodimer, with the first monomer covering residues 1 to 99 on the X-axis, and the second monomer covering residues 100 to 198. It can be seen that with the exception of residues 30-50, and the corresponding residues 130-150 of the second monomer, the RMSD between the two structures' backbone C_α atoms ranges between 0.1 and 0.7Å. These values are small, and show that the two structures have almost exactly the same quaternary structure. This was expected with both 1HXB and 1BDQ adopting the same hexagonal P 6_1 space group symmetry, and therefore a similar configuration during crystallisation.

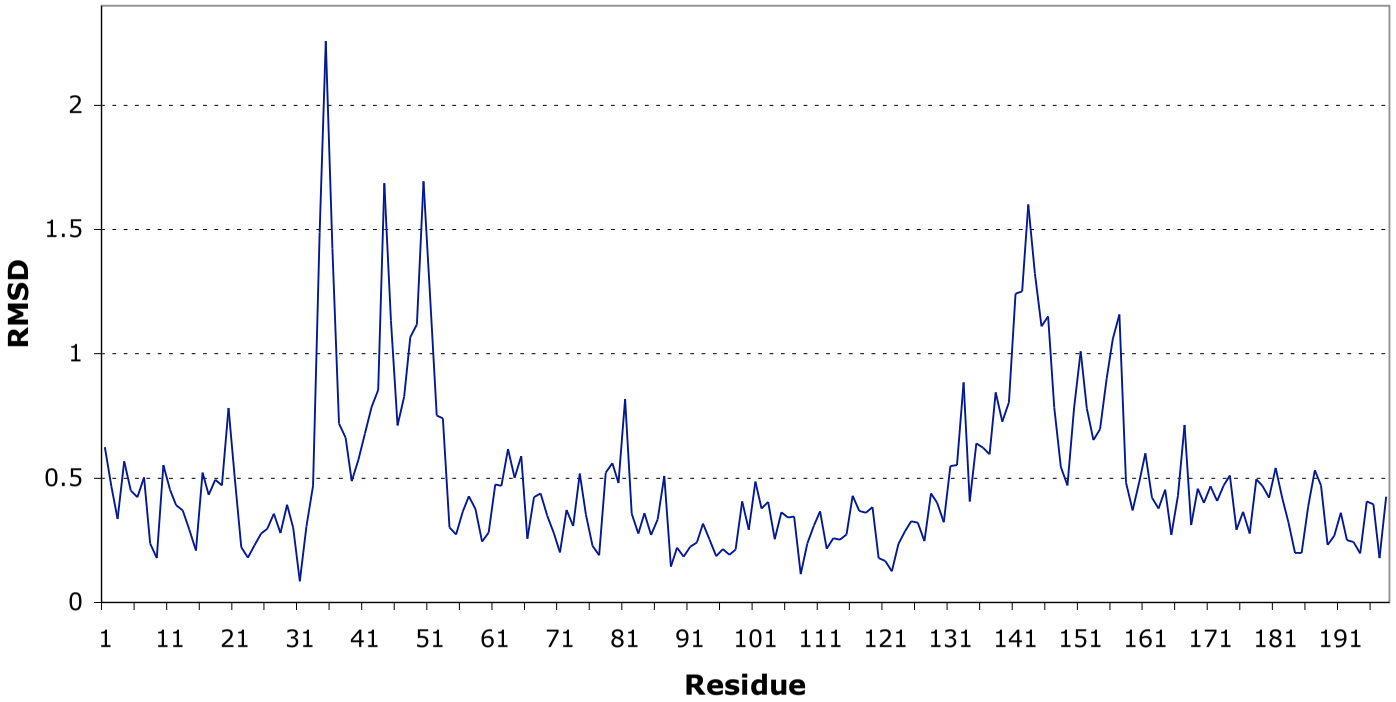


Figure 4.1: Profile RMSD comparison between the C_α backbone atoms in the crystal structures of 1HXB and 1BDQ. With the exception of residues 30-50 and 130-150, the RMSD between corresponding backbone atoms is $\sim$0.5Å which shows strong structural similarity between the two structures. The peaks map to the flexible cantilever and flaps regions of the protease, and contain 8 of the 10 different residues between the genotypes.

The RMSD peaks between residues 30-50 and 130-150 correspond to the flaps and flap elbow regions of the protease structure . These regions are known to be flexible [33, 34, 99], and in addition 8 of the 10 mutations between the two structures were located in this region, with resultant alterations in the structures (Figure 4.2). The backbone regions shown in red and orange represent residues 30 to 50 in each monomer for 1HXB and 1BDQ respectively. The small deviations in the two backbone structures are particularly apparent in comparison to the rest of the structure, where the two backbones superimpose almost perfectly. These two structures were then used as the

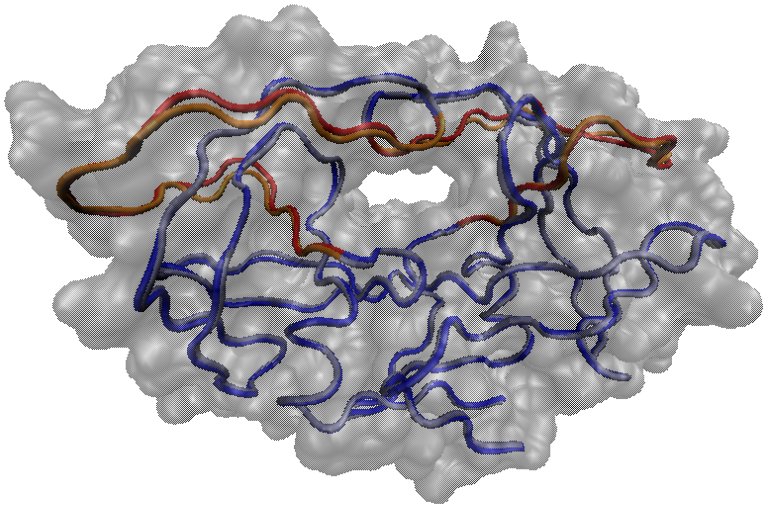


Figure 4.2: Superimposed crystal structures of 1HXB and 1BDQ. The backbone atoms and bonds are shown in dark-blue for 1HXB and light-blue for 1BDQ. Residues 30 to 50 in both monomers of each structure are shown in red for 1HXB and orange for 1BDQ, representing the regions of higher RMSD in Figure 4.1. The overall quaternary structure of 1HXB is shown in grey. Figure was created using VMD.

starting configurations for the molecular dynamics simulations. The protocol followed was similar to that described in Section 2.1.2: the protease molecule was placed in a neutralised 10Å water box and minimised through the conjugate gradient minimisation algorithm for 4 picoseconds. Following this, a force constraint of 25 kcal/mol/Å^2 was placed on all non-hydrogen protein atoms and the system simulated for 50 picoseconds while the temperature was increased from 50 Kelvin to 300 Kelvin. This allowed the

water in the surrounding water box to expand and fully solvate the protein. The temperature was increased slowly over this time to ensure that any kinetic energy added to the system was allowed to dissipate through the system. This ensured that the protein did not gain too much energy to adopt an unnatural configuration or unfold. Once solvated and heated, the constraints on the protein were lifted in 5 kcal/mol/Å^2 increments every 10 picoseconds for 50 picoseconds. Again, this was done to ensure that the sudden removal of constraint forces doesn't give the system enough freedom to adopt an unnatural configuration. The system was then simulated for a further 240 picoseconds for data collection.

The change in global RMSD with respect to the crystal structure were determined for 1BDQ and 1HXB (Figures 4.3 and 4.4 respectively). The global RMSD was calculated for each outputted configuration by superposition and comparison to the initial crystal structure's backbone. Instead of outputting a value for each residue's C_α atom, the sum across the whole backbone was calculated and then divided by the number of residues to generate an average RMSD across the whole protein. This acted as an indicator for the similarity of the two proteins as a whole. The results show that once the proteins were fully-relieved of the constraints placed upon them, they rapidly relaxed away from their constrained configuration to a configuration that they stably flexed around. For 1BDQ, relaxation took approximately 20 picoseconds, whence the gradient-plateau of global RMSD with respect to 1BDQ crystal structure showed that it finds a stable configuration, averaging 1Å away from the crystal structure, around which the natural thermal vibrations, side-chain rotations, and natural quaternary structure motions caused small ~ 0.1Å deviations, as indicated by the high-frequency oscillations. For 1HXB the relaxation time was approximately 40 picoseconds. However, although the gradient of change in RMSD became markedly less steep at this timepoint, the data showed that there was still a slight drift from ~ 0.8Å global RMSD at 140 picoseconds to ~ 1.0Å at 340 picoseconds. The simulation was therefore extended to 700 picoseconds to determine whether the drift in global RMSD was significant or whether it was a low-frequency natural movement that would indicate that the protease had relaxed into stable configuration. The results, shown in Figure 4.4, show that the system remains stable until approximately 450 picoseconds, when large-scale motions of the protease cause 0.4Å

deviations for 150 picoseconds, before settling again at ~ 1.1Å. Visual inspection of the protease through VMD showed that this sudden change in RMSD was caused by a small rotation in the backbone of the flaps region that resulted in subsequent flaps residues becoming misaligned which affected the average RMSD value. This suggested that the 1HXB system had relaxed by 140 picoseconds, so the outputted structures of 1HXB and 1BDQ at the 340 picoseconds time-point were compared through profile RMSD to determine whether the structures had diverged (Figure 4.5).

After 340 picoseconds, the proteases were much less ordered than their crystal structures, with different regions of the proteases showing different levels of similarity (Figure 4.6). It is also noticeable that the first monomer of each protease was significantly less similar than the second monomer, and the two monomers did not mirror each other. Therefore, although the monomers in the homodimeric protease were genotypically identical, random thermal fluctuations resulted in the two monomers fluctuating independently of each other. This is in contrast to the crystal structures which showed similar trends of similarity in each monomer; the peaks of deviation between the structures occured at residues 30-50 of both monomers, with surrounding residues being significantly similar, particularly towards the C-terminus (residues 80-99). The peaks of least similarity between 1HXB and 1BDQ at 340 picoseconds are between residues 32-50; 64; 92-104; and 164, with the rest of the dimer fluctuating between 1.0Å difference and 2.5Å. Figure 4.7 shows the locations of the residues with > 3.0Å RMSD, and, as can be seen, these regions are all superficial on the protease, where the residues have more freedom to move around. Comparing this figure to Figure 4.2, the reduction in superposition of the backbones is noticeable through the structures.

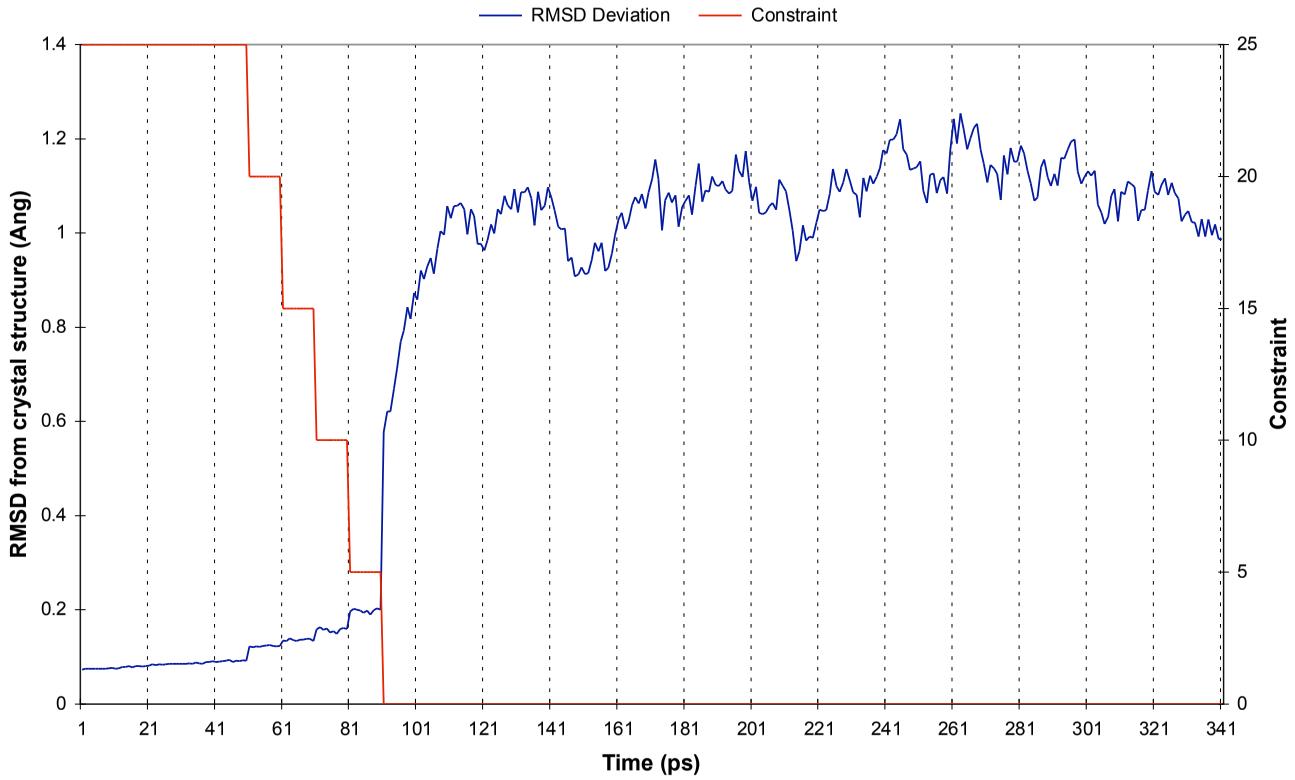


Figure 4.3: Change in global RMSD from 1BDQ crystal structure for each outputted configuration across the 700 picosecond MD simulation (the first 100ps of which involves the minimisation protocol). The results show a rapid change of configuration from initial structure to approximately 1.1Å average residue difference, where it remains for the course of the simulation (blue line). The red line depicts the strength of the constraint force placed upon the protein's non-hydrogen atoms, in kcal/mol/$\mathring{A}^2$.

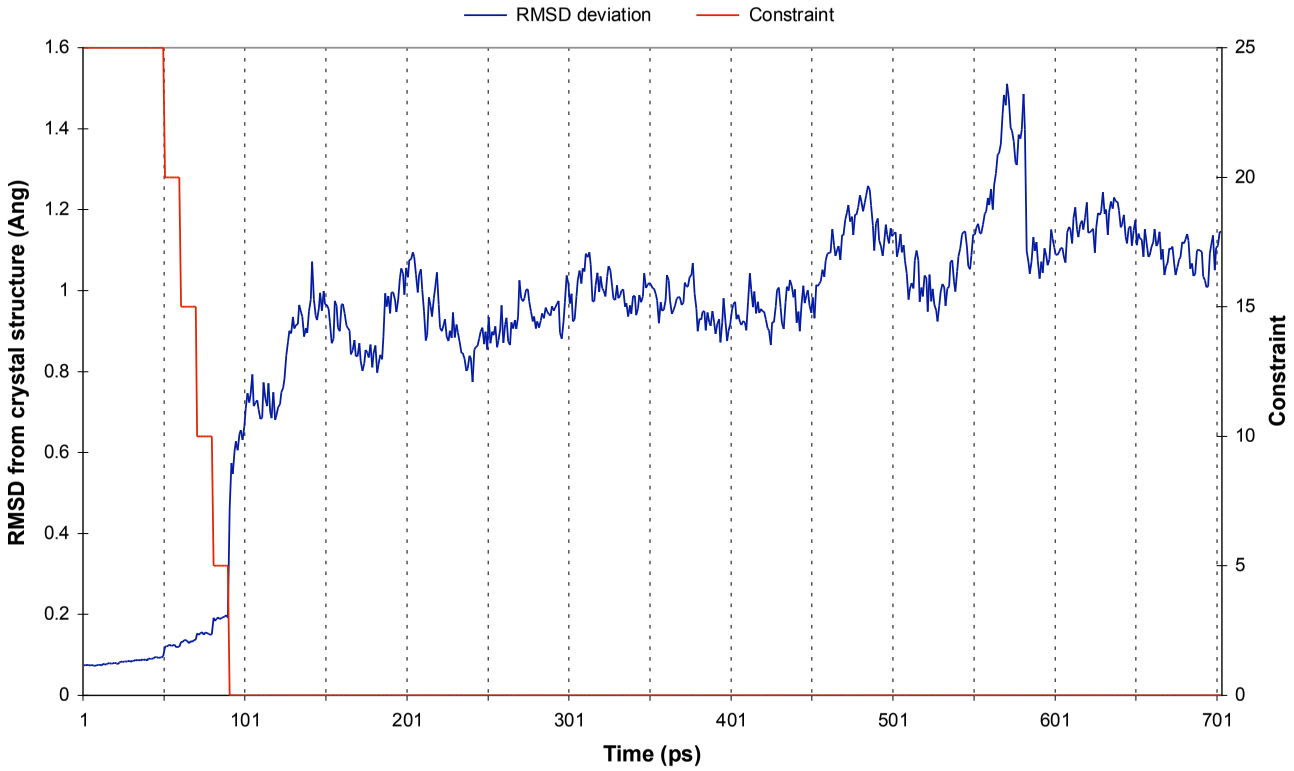


Figure 4.4: Change in global RMSD from 1HXB crystal structure for each outputted configuration across the 700 picosecond MD simulation (the first 100ps of which involves the minimisation protocol). The results show a rapid change of configuration from initial structure to approximately 1Å average residue difference, where it remains constant until approximately 500ps, whereupon a slight change in conformation to 1.1Å deviation (blue line). The red line depicts the strength of the constraint force placed upon the protein's non-hydrogen atoms, in kcal/mol/$\AA^2$.

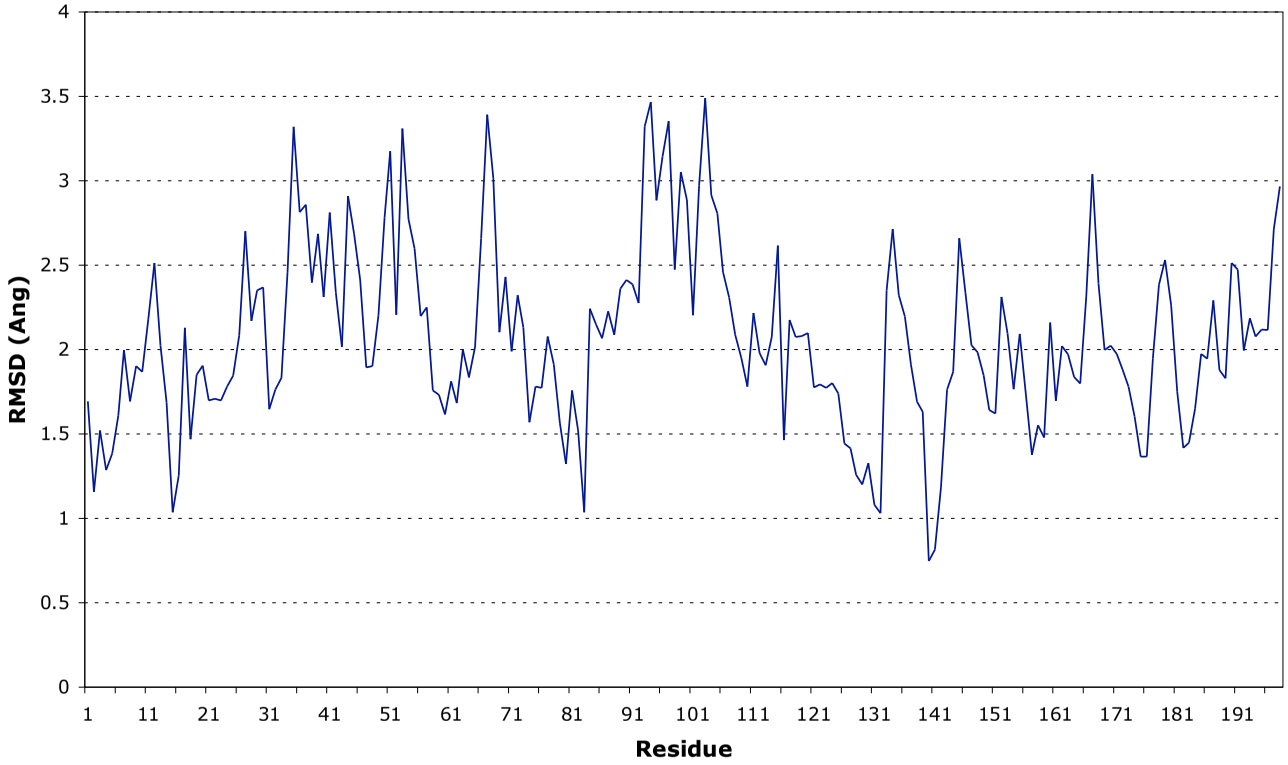


Figure 4.5: Profile RMSD comparison between the C_α backbone atoms of 1HXB and 1BDQ after 340 picoseconds of MD simulation. The results show a marked difference to the profile RMSD of the crystal structures shown in Figure 4.1; only residues 136-137 (residues 35-36 of the second monomer) have an RMSD of < 1.0Å, with many residues having > 3.0Å.

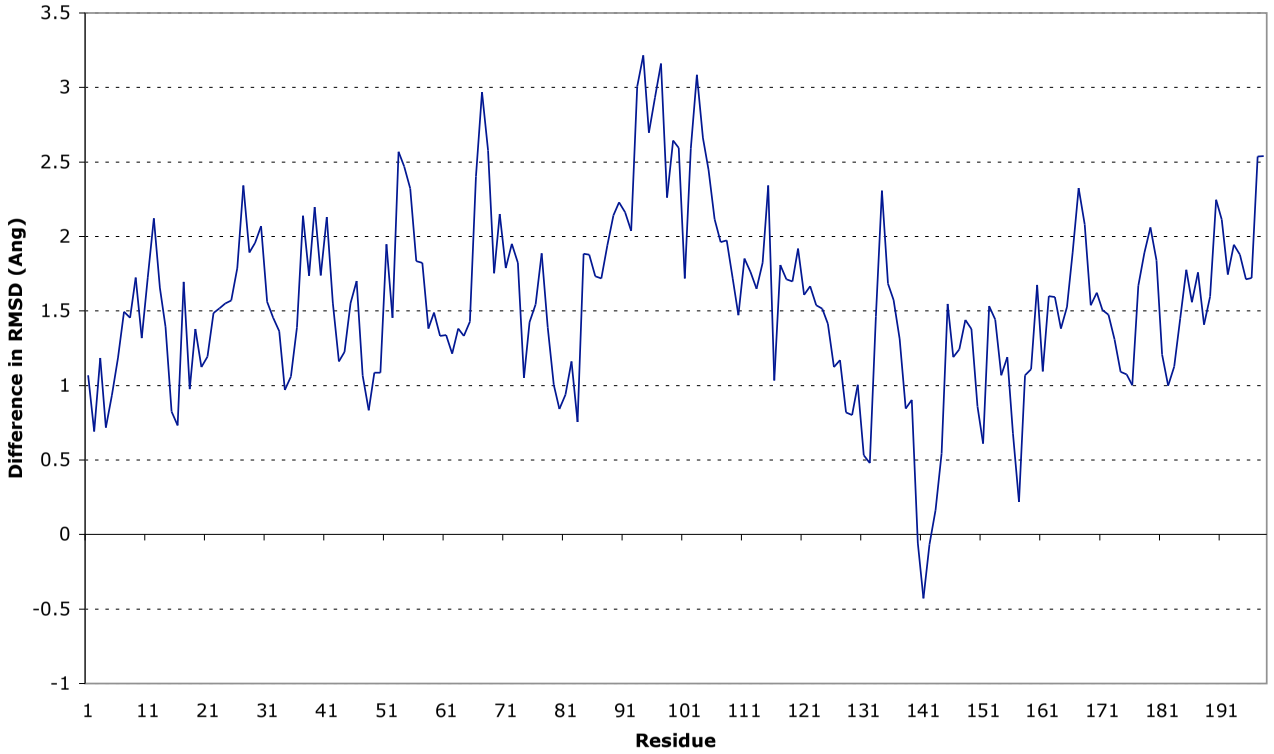


Figure 4.6: Difference between the profile RMSD plot between 1HXB and 1BDQ crystal structures and the profile RMSD plot between 1HXB and 1BDQ structures after 340ps simulation (Figure 4.5). The results show that except for residues 140, 141 and 142, the simulated proteases are more structurally-diverged than the crystal-structure proteases.

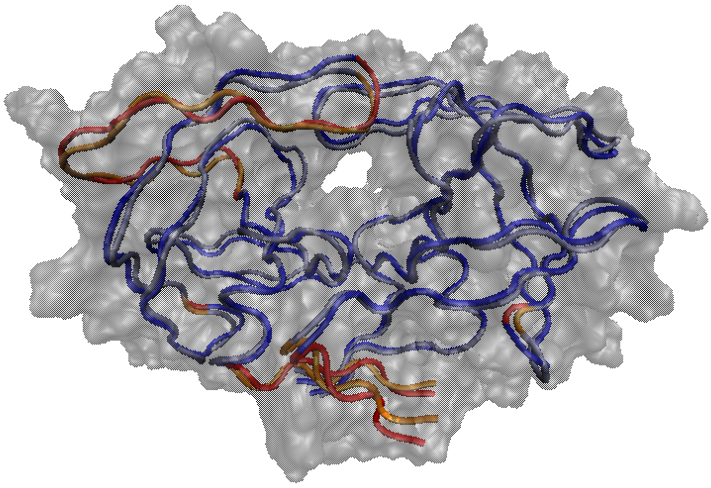


Figure 4.7: Superimposed structures of 1HXB and 1BDQ after 340ps of simulation. The backbone atoms and bonds are shown in dark-blue for 1HXB and light-blue for 1BDQ. Residues 32 to 50, 64, 92 to 104 and 164 are shown in red for 1HXB and orange for 1BDQ, representing the C_α residues that had an RMSD > 3.0Å. The overall quaternary structure of 1HXB is shown in grey. Figure was created using VMD.

With the exception of residues 140, 141 and 142, the simulated proteases have structurally-diverged away from their crystal structures, and away from each other (Figure 4.6). These three residues that appear more structurally-converged are located on the flap elbow of the second monomer, where they would be expected to have high flexibility due to their location on the extremity of the protease with their side-chains pointing out into the solvent. Combined with the fact that the immediately-surrounding residues show considerable divergence between proteases, these results suggest that the convergence of these individual residues is not an indication of similarity between the two when taken in the context of the tertiary structure in the cantilever region, but rather that this flexible region in the structures happened to overlap at this particular point in time when superimposed.

It can be seen from the fluctuations in Figures 4.4 and 4.3 that the protease structures are dynamic, with each fluctuating by up to 0.4Å around its average relaxed structure. It could therefore be argued that considering the similarity of two single structures at a single time-point in the simulation is not sufficient to conclude that the two structures have diverged; it could be argued that both 1HXB and 1BDQ diverged away from their

similar crystal structures to a common structure with a global RMSD of ~ 1.0Å away from starting structures, about which the two proteases fluctuate out-of-phase, giving the impression of different structures at a single point in time. Therefore, analysis of the simulations' structures was extended to include this dynamic nature; each snapshot of the 140 picoseconds between 220ps and 340ps of 1HXB's simulation was compared to each snapshot of the same timeframe in 1BDQ's simulation through global RMSD analysis. This covers the spread of configurations that each protease samples and will remove the effect of out-of-phase sampling where the two structures are sampling the same range of configurations at the same frequency but at different time-points. In addition, in order to get a more detailed indication of the configurational fluctuations of a protease, each snapshot between 200ps and 340ps was compared against all other snapshots in the same protease's simulation through global RMSD. This intra-protease simulation analysis acts as a 'background' fluctuation by which to compare the inter-protease fluctuations; if all three analyses (1HXB vs. 1HXB, 1BDQ vs. 1BDQ and 1HXB vs. 1BDQ) show similar ranges and means of RMSD values, then it can be concluded that they sample the same configurations. Conversely, if the inter-protease RMSD values are significantly higher than the intra-protease RMSD values, then it can be said that they have structurally-diverged. Three matrices were generated with notation as shown in Equation 4.1:

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \tag{4.1}$$

where m is the number of analysed snapshots in one simulation; n is the number of analysed snapshots in the second simulation; $a_{i,j}$ is the global RMSD value of the i^{th} snapshot of the first simulation superimposed to the j^{th} snapshot of the second simulation. For example, in the 1HXB vs. 1BDQ matrix, position $a_{120,110}$ is the global RMSD value between 1HXB's 120^{th} snapshot (at time-point 340ps) and 1BDQ's 110^{th} snapshot (at time-point 330ps). In the intra-protease matrices, the values down the main diagonal (entries $a_{i,j}$ where $i = j$) all equal 0 because the protease's snapshot is being compared against itself. These values were therefore not considered when eval-

uating the matrices because they would skew any statistical consideration of the set of values. For each matrix, the frequency distribution of RMSD values was generated with 0.05Å intervals (Figure 4.8, together with summary statistics (Table 4.2).

Table 4.2: Statistics of the inter- and intra-protease RMSD matrices

Statistics	**Inter- & intra-protease RMSD (Å)**		
	1HXB vs. 1BDQ	**1HXB vs. 1HXB**	**1BDQ vs. 1BDQ**
Range	0.90 - 1.53	0.43 - 1.22	0.45 - 1.17
Mean	1.21	0.83	0.81
Mode[a]	1.25	0.85	0.85
STDEV[b]	0.09	0.13	0.11

[a] Most populous interval used in the generation of the frequency distribution.
[b] is the standard deviation.

The results show that upon configurational relaxation, structures 1HXB and 1BDQ diverge into discrete configurations whose quaternary structures differ by an average of 1.21Å(Figure 4.8). Upon relaxation, each protease configurationally flexes around an average structure, as shown by the small variations in ΔRMSD in Figures 4.4 and 4.3 and also by the normal distribution of RMSD frequencies (Figure 4.8). These conformational fluctuations experienced by each protease are similar in magnitude, with a similar range of RMSD distribution (Figure 4.8 and Table 4.2). However, 1BDQ shows a smaller standard deviation around its mean, about which a higher proportion of the snapshots in its simulation adopt. 1HXB, meanwhile, has a similar mean RMSD frequency but has a more diffuse normal distribution about this mean, with a higher proportion of snapshots > 0.90Å apart. This indicates that 1BDQ spends more of its time around a particular configuration while 1HXB tends to sample other configurations more frequently.

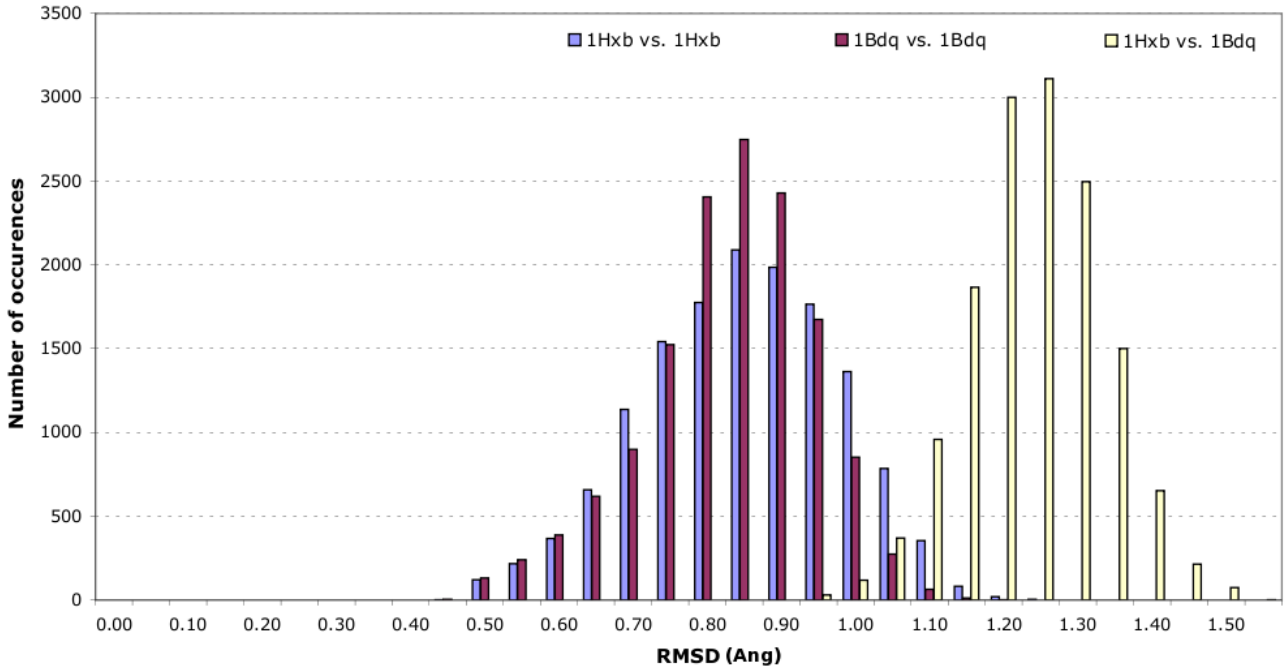


Figure 4.8: Frequency distributions of global RMSD values for intra- and inter-protease comparisons. The data used to generate the frequency distributions came from the RMSD values contained in the matrices described in Equation 4.1. The graph shows that the structural differences observed between 1HXB and 1BDQ are due to the proteins sampling different configurations, and that the range of conformational fluctuation is similar for each protein.

The results also show that the configurations adopted by each protease do not overlap, as the RMSD values between the two proteases range from 0.90Å to 1.53Å (Figure 4.8). This indicates that upon relaxation to the configuration about which they fluctuate, the proteases adopt a distinct range of configurations that never comes within 0.9Å of each other. This strongly suggests that the homologous structures adopted by two genotypically-variable proteases upon crystallisation relax into morphologically-distinct structures upon equilibration and subsequent simulation. This could have a significant impact on predicting how drug-resistance mutations cause their effect in genotypically-diverse, but crystallographically-similar, structures.

4.3 Further comparison of simulated structures

To ensure that the results observed between 1HXB and 1BDQ were not unique to the systems chosen, and that they are reproducible, the study was repeated by comparing 1HXB and 1A8G in the same manner. 1A8G was chosen because it is genotypically identical to the Hxb2 sequence, giving it two mutations from the 1HXB structure: V3I and S37N, putting it intermediate between 1HXB and 1BDQ.

As with the other two systems, the ligand was removed from 1A8G's structure so that the system was simulated as an 'apo-protein'. The same minimisation and equilibration protocol was followed as previously described for 1HXB and 1BDQ, and following this the system was simulated for 340ps for data collection. The final 120 picoseconds, between 220ps and 340ps was structurally analysed through global RMSD with respect to its crystal structure, which showed a rapid ΔRMSD away from the crystal structure upon constraint-removal, which reached a plateau at 1.0Å and fluctuated around this structure. Following this, the matrix shown in Equation 4.1 was generated between the same 120 snapshots of 1HXB, and the 120 snapshots between 220ps and 340ps for 1HXB, giving a frequency distribution graph (Figure 4.9).

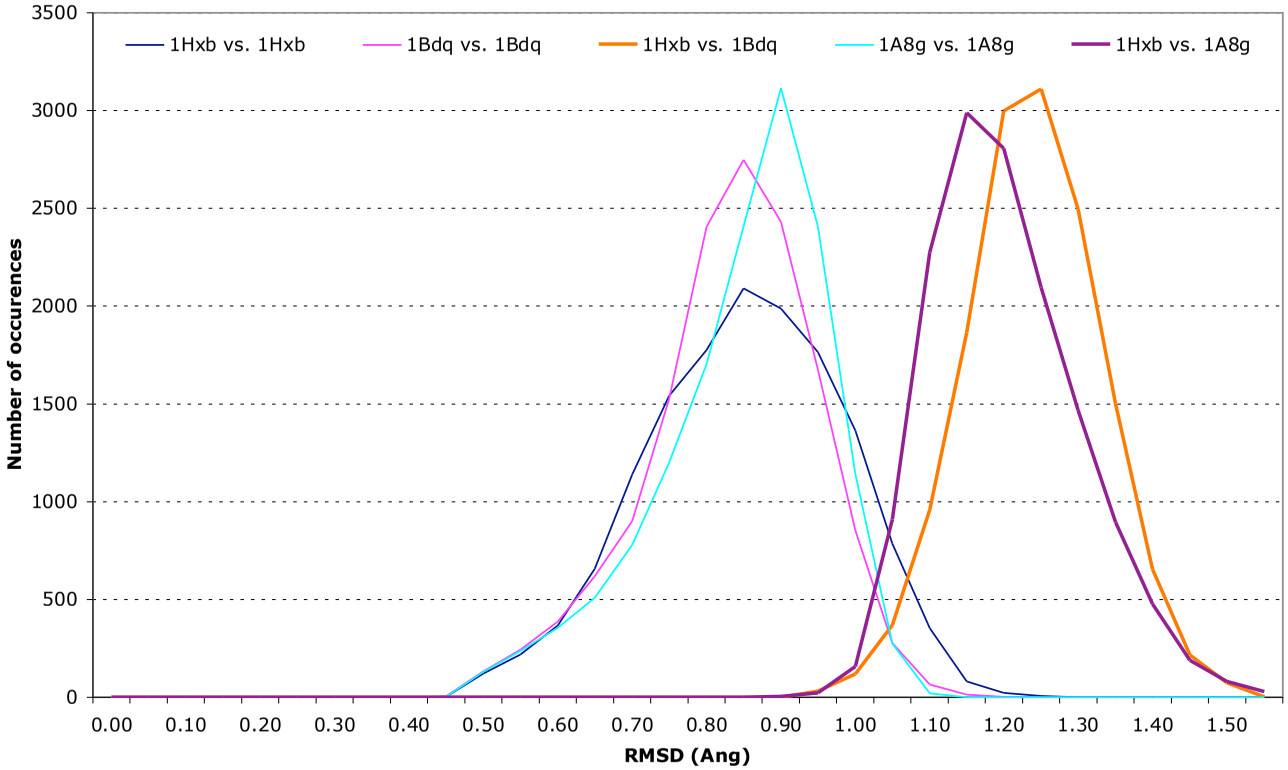


Figure 4.9: Frequency distributions of intra- and inter-protease global RMSD values for 1HXB, 1BDQ and 1A8G. The data used to generate the frequency distributions came from the RMSD values contained in the matrices described in Equation 4.1. The graph shows that the 1HXB-1A8G inter-protease RMSD values concur with the 1HXB-1BDQ inter-protease results, and the 1A8G intra-protease frequency distribution concurs with the other intra-protease distributions.

These data show a continuation of the pattern seen between 1HXB and 1BDQ (Figure 4.9), with the intra-protease RMSD values forming normal distributions with similar means but varying standard deviations, and the inter-protease RMSD values forming positively-translated taller, tighter normal distributions (Table 4.3).

Table 4.3: Summary statistics of the frequency distributions of 1HXB-1A8G inter- and 1A8G-1A8G intra-protease matrices

Statistics	**Inter- & intra-protease RMSD (Å)**	
	1HXB vs. 1A8G	**1A8G vs. 1A8G**
Range	0.85 - 1.55	0.44 - 1.10
Mean	1.18	0.82
Mode[a]	1.15	0.90
STDEV[b]	0.10	0.11

[a] Most populous interval in the generation of the frequency distribution.
[b] is the standard deviation.

Comparing the summary statistics in Table 4.3 to Table 4.2, and the frequency distributions in Figure 4.9 to Figure 4.8 it can be seen that the 1A8G results tie closely to the 1HXB-1BDQ results, and therefore support the conclusions drawn from the 1HXB-1BDQ analysis. Interestingly, the 1HXB-1A8G inter-protease distribution is slightly translated towards the intra-protease distributions, with a mean value of 1.18Å compared to 1.21Å for the 1HXB-1BDQ distribution. This could be because 1A8G has less mutations than 1BDQ, which would suggest that the more mutations a protease has from 1HXB, the further from 0Å the mean RMSD value between the two becomes. It is important to note that the range of RMSD values between the 1HXB and 1A8G are the same as for 1HXB and 1BDQ, but in the 1A8G comparison the distribution is more negatively skewed.

4.4 Conclusions

The results in this chapter can be best described through the proteases' potential energy surfaces, also called a hypersurface or a 'configurational landscape'. The potential energy of a protein is a function of its configuration, such that as atomic movements

cause bonds to rotate and stretch, the potential energy of the protein changes. By reducing the dimensionality of the protein's configuration down to two functions, these can be plotted on perpendicular axes. The potential energy of the configuration is then plotted on the third perpendicular axis, creating a configurational landscape (Figure 4.10). This configurational landscape has peaks and troughs, where peaks indicate higher-energy configurations, and troughs indicate lower-energy configurations. The underlying thermodynamics dictate that proteins will adopt configurations that minimise its potential energy.

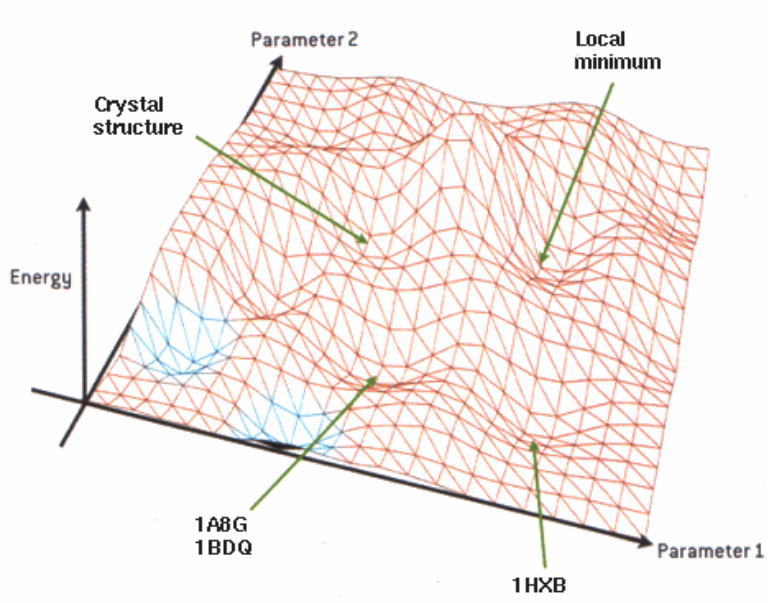


Figure 4.10: Hypothetical configurational energy-landscape. Parameters 1 and 2 on the X- and Z-axes are a function of the protein's configuration such that the position on the X-Z plane is a function of the protein's configuration. The configurational plane is stretched in the Y-axis to represent the potential energy of the protein's configuration. Troughs in the landscape indicate local configurations with minimum energy, the lowest of which is termed the global minimum. The arrows give an example of hypothetical sampling locations of the simulated structures. Figure adapted from a Scientific American image titled "The Landscape" and downloaded from <*http://singularidad.wordpress.com/category/humanidades/filosofia/*>.

If a fourth axis of *time* is added, then as time proceeds, the position of the protein will tend towards local energy minima, about which they configurationally sample. In order to go against the gradient and head towards the 'peaks', energy needs to be provided to adopt the necessary configurations. This energy is provided in the form of thermal

energy; the higher the thermal energy of the protein, the greater the probability of adopting high energy configurations. This is important because it allows the protein to reach other local minima in its configurational landscape. Results attained from *in vitro* protein experiments such as ITC are averaged from approximately 10^{16} molecules (for a 1.6ml protein solution at 30μM concentration [37]), therefore the entire range of configurations available to the protease will be thoroughly represented. To accurately replicate experimental data, the same range of conformations needs to be sampled computationally.

However, 1HXB, 1A8G and 1BDQ do not share a common configurational landscape as they do not share the exact same structure; each system contains atoms and bonds not possessed by the other two systems, so therefore adopt configurations unattainable by the other two. However, if the X-Z configurational plane is restricted to a function of just the backbone atoms, then the results can be described across a shared configurational landscape. While this loses configurational information of the side-chains, the overall global quaternary structure is contained in the backbone, which is sufficient to configurationally distinguish the proteases [92]. The crystal structures of the proteases can be postulated to be located at or near the zenith of an energy peak (Figure 4.10). The process of minimisation and equilibration allows the protease to relax into a configuration with less potential energy, which would be reflected by a movement down the peak. Due to the slight variations in the crystal structures combined with variation introduced through the random assignment of thermal energy at the first timestep of the molecular dynamics simulation, each protease is unlikely to take the same path down the peak. This can lead to the proteases residing in different local energy minima. In these simulations it is suggested that the 1HXB structure resided in a different energy minima than that reached by 1BDQ and 1A8G. As a result, the mean RMSD values between 1BDQ/1A8G and 1HXB were approximately equal at 1.2Å. The differences in range, standard deviation, and mean RMSD values of the 1BDQ and 1A8G intra-protease results (Figure 4.9) may have been a result of differences in the distribution of thermal energy in the systems leading to different conformational sampling of the minimum. These results suggest that the systems only sample a single energy minimum over 340 picoseconds, indicating that they need to be simulated for longer.

Chapter 5

Validation of molecular dynamics simulation protocol

5.1 Introduction

It was argued in Chapter 4 that MD simulations can complement static structural analysis of proteins, and subsequently shown that MD simulations can reveal novel structures unavailable through crystallography. However, MD simulations are still reliant on these static structures as initial configurations from which to start, and this severely limits the range of protein genotypes that can be simulated. In the case of HIV-1 protease, only 150 unique genotypes were present in the local database described in Chapter 3, which is significantly less than the potential, and observed, number of clinically-observed genotypes. This is due in part to the difficulty of crystallising proteins, and also the complete inability of some proteins to crystallise due to their flexibility, shape, or polarity. As a result, in order to be able to simulate every possible HIV-1 protease genotype, the crystal structures need to be computationally-mutated into the sequence of interest. This, however, introduces side-chain atomic coordinates that are not verified experimentally, with the potential to cause abnormal movements or configurations that would not be observed *in vivo*. For example, computationally mutating a small residue, such as glycine (-H) into a larger residue, such as lysine ($[-CH_2]_4 - NH_3{}^+$) in the interior of a globular protein without changing the surrounding tertiary structure to compensate could result in atoms coming in too close contact

and as a result the energies of the protein being distorted. The same principles apply when changing residue polarity upon mutation, and these must all be taken into consideration with the computational mutation-protocol. Therefore, it was investigated whether the mutational protocol (Section 2.1.1) was robust. The limit of the number of residues that can be computationally-mutated while still retaining structural integrity was also determined. Further research was then undertaken to computationally-mutate a crystal structure into a second structure with a different genotype. By simulating the crystal structure against the homo-genotypic mutated structure and comparing structural and energetic results, an indication of the quality of the mutational protocol could be determined.

5.2 Limits of computational residue-mutation protocol

Initial investigations on the limits of the computational mutation protocol concentrated on how many residues could be mutated in the dimer before affecting the protease's structural integrity. In comparison to the simulations performed in Chapter 4, these simulations were run with an inhibitor complexed so that the binding affinity between the two could be calculated from the simulation. By comparing this computed binding affinity to experimentally-determined binding affinities, the quality of the mutated-protease's simulation could be ascertained. Therefore, the chosen genotypes to which the crystal structures will be mutated must have associated experimentally-determined biochemical data. In order to achieve this, the local database described in Chapter 3 was explored with the Perl script *mutation_data_extractor.pl* (Section 3.3). This script was repeatedly invoked with its two command-line arguments: potential PDB starting structures, and an integer N where $1 \leq N \leq 13$. For each invocation, the script outputted a file containing the sequences with N mutations from the specified crystal structure and associated experimentally-derived data for that genotype.

Using this Perl script, a series of thirteen protease genotypes were identified in the database, each with a different number of mutations from a crystal structure complexed with the protease inhibitor saquinavir, and each with an associated experimentally-derived binding affinity value for that genotype complexed with saquinavir. The output

was a list of PDB structures from which to start the simulations, and for each system a list of the required mutations to convert it to a genotype with associated binding affinity data (Table 5.1).

Table 5.1: Initial PDB and nature of mutated residues for each of the 13 mutational-chain systems

Mutations	**Nature of mutations from initial PDB**	**Initial PDB**
0	-	1HXB
1	V48G	1FB7
2	Q7K, I84V	1FB7
3	Q7K, V48G, M90L	1FB7
4	V3I, V48G, V82A, M90L	1FB7
5	Q7K, L33I, V48G, L63I, M90L	1FB7
6	V3I, Q7K, L10I, L33I, V48G, L63I	1FB7
7	Q7K, L33I, V48G, L63I, V82F, I84V, M90L	1FB7
8	V3I, Q7K, L33I, V48G, L63I, V82A, I84V, M90L	1FB7
9	Q7K, I13V, E35D, M36I, S37N, R41K, R57K, H69K, L89M	1HXB
10	V3I, Q7K, L33I, M46I, V48G, I54V, L63I, V82A, I84V, M90L	1FB7
11	Q7K, I13V, E35D, M36I, S37N, R41K, V48G, R57K, H69K, L89M, M90L	1FB7
12	Q7K, L10I, M36I, S37D, M46I, V48G, R57K, L63P, A71V, G73S, I84V, I93L	1FB7
13	Q7K, I13V, E35D, M36I, S37N, R41K V48G, R57K, H69K, V82F, I84V, L89M, M90L	1FB7

Having determined the nature of each computationally-mutated system, each initial PDB structure was mutated accordingly by following the pre-simulation protocol described in Section 2.1.1. This protocol differed from the equilibration protocol described in Chapter 4 due to the increased caution that needed to be followed to ensure the mutated residues reorientate into a more natural, relaxed configuration. To achieve this, before the global constraints were gradually lifted from the protein complex, the constraints were lifted for a 5Å sphere surrounding each mutated residue in sequential order for 50 picoseconds. The reason for lifting the constraints on the immediate surrounding region was because the mutated residue may need to displace surrounding atoms in order to 'un-trap' itself and reach a comfortable orientation. This is most necessary for

any mutations internalised within the globular protease, where the close-packed tertiary structure means that a larger side-chain of a mutated residue could get trapped in a high-energy configuration. Once each mutated residue had been given time to reorientate itself, it is re-constrained while the next mutated residue is released. Once the final mutation has been re-constrained, the constraints on the whole protein are gradually released and the protein given time to equilibrate. The whole process of minimisation and equilibration was increased from the 220 picoseconds protocol in Chapter 4 to a 2 nanosecond protocol to allow sufficient minimisation. Following on from the equilibration protocol, data was collected for a further 2 nanoseconds to give the systems time to sample more of their configurational landscape. Simulations were therefore run for a total of 4 nanoseconds, which was significantly more computationally-demanding than previous simulations. Therefore these simulations were parallelised across 32 processors on the Leeds site of the National Grid Service (NGS). Four simulations were run concurrently across 32 processors at a rate of 8 hours/ns. Therefore each simulation took 32 hours (1.3 days), requiring 1024 computer-hours to achieve this. The entire mutational-chain was completed in 128 hours (5.3 days), requiring 14,336 computer-hours. Trajectory data was transferred back to a local server for analysis via 'GSISCP'.

Before the data could be analysed, it needed to first be ensured that any variations between systems were not a result of differences between the two different starting structures; 1HXB and 1FB7. Therefore three extra 2 nanosecond simulations were therefore run on the Leeds site of the NGS: 1FB7 crystal structure, 1HXB mutated into 1FB7, and 1FB7 mutated into 1HXB. By cross-comparing the results of these simulations with the un-mutated system in the mutational chain (1HXB crystal structure), any adverse effects on the simulations due to the choice of starting crystal structure could be ascertained. Profile-RMSF comparison of the four simulations showed significant overlap, suggesting that there is no significant difference in the dynamics of the 2 crystal structures, nor that the mutation protocol affects these dynamics. Furthermore, calculation of the ΔG_{bind} for each system resulted in all systems falling within error of each other (Table 5.2). These results suggest that the choice of starting structure does not affect either the structural dynamics or the energetics of the protease-saquinavir complexes. Therefore, any deviations within the mutational chain will not be attributed to the

differences in starting structure or the mutational protocol employed.

Table 5.2: Binding free energy values for the control-simulations

System	**ΔG_{bind} (kcal/mol)**
1HXB crystal	-17.53 ± 1.05
1FB7 crystal	-16.55 ± 2.12
1HXB -> 1FB7	-16.38 ± 1.98
1FB7 -> 1HXB	-16.72 ± 1.59

Once all the simulations had been completed, their structural dynamics were compared through profile RMSF of the backbone C_α atoms (Figure 5.1). For each system, the time-averaged position of each C_α atom was determined, and the RMSF of this atom calculated across the simulation. This generated a value for each of the system's 198 C_α atoms that represented the average movement of that atom around the average structure across the simulation. The results show considerable overlap between the systems' profile RMSF values, with all systems fluctuating to approximately the same degree across their simulations, and the regions of relative flexibility and rigidity concurring between systems. This suggests that the mutation protocol is robust enough to mutate and configurationally relax up to 26 residues in the protease dimer whilst retaining the structural dynamics. Furthermore, the residues mutated in this study were spread throughout the protease's quaternary structure (Figure 5.2(d)), indicating that the location of the mutated residue is not a concern in this system.

The profile RMSF of the 1HXB un-mutated system (Figure 5.1) gives an indication of the structural and dynamic constraints of the protease's quaternary structure. The residues with RMSF values below 0.5Å can be considered as 'rigid regions' which show little fluctuation over the simulation; those with fluctuations between 0.5Å and 1.0Å are less restrained in their movements; and those with fluctuations greater than 1.0Å are flexible regions that show considerable movement over the simulation. Figures 5.2(a) and 5.2(c) display these RMSF gradings superimposed onto the protease's backbone structure for the un-mutated system and the system with 13 mutations respectively. This shows that the regions of low fluctuation were restricted to the core of the glob-

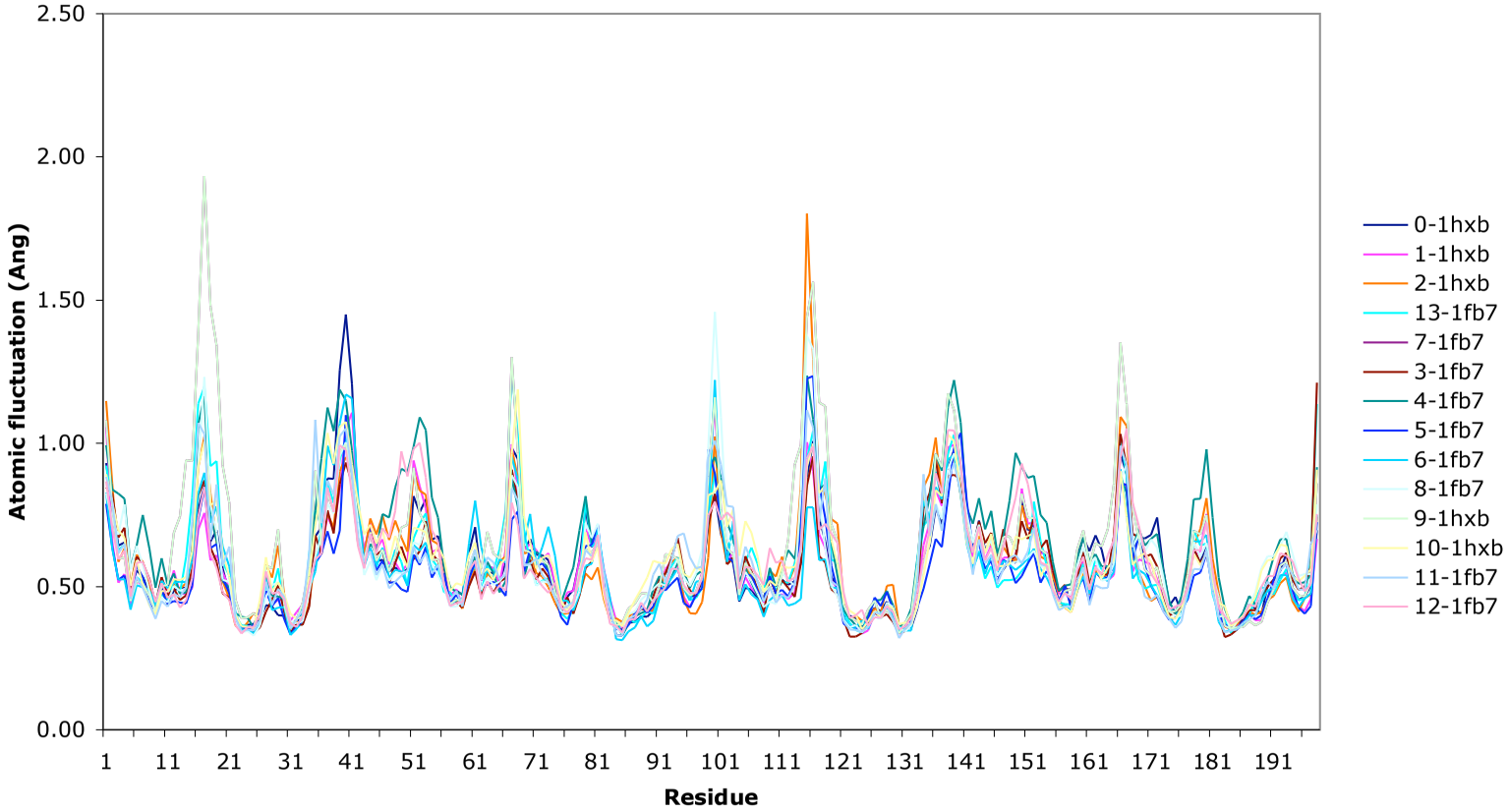
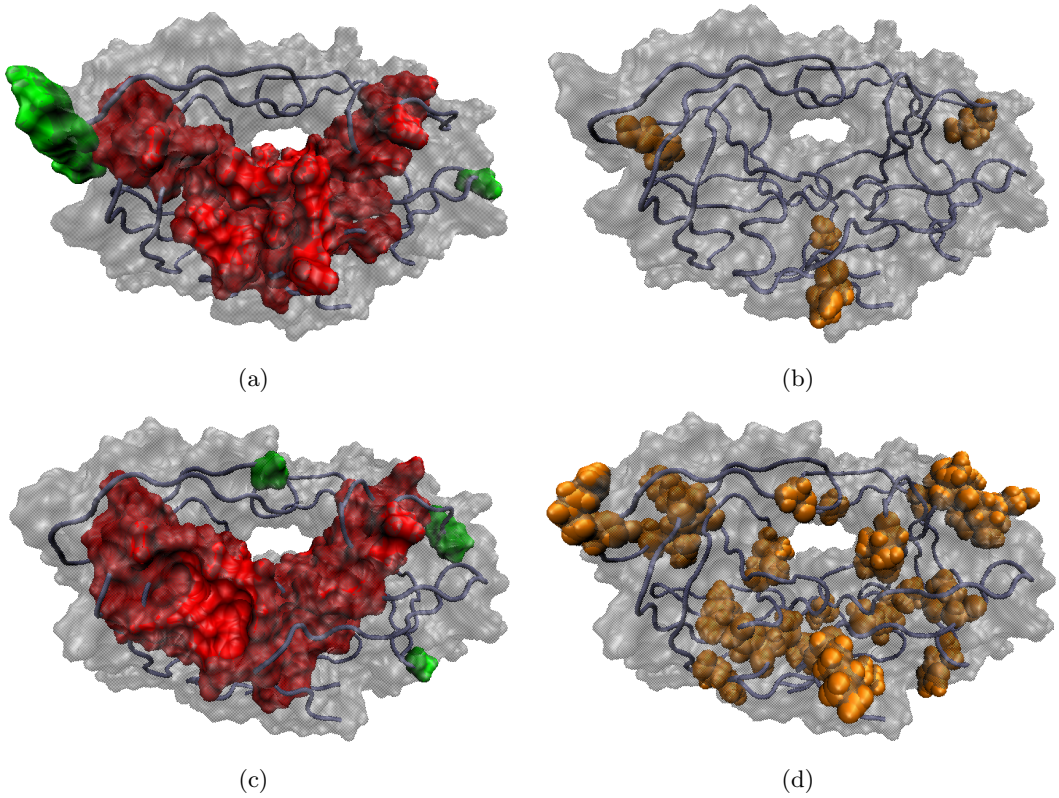


Figure 5.1: Comparison of profile RMSF values for the 14 mutational-chain systems. For each system, the time-averaged position of each backbone $C_\alpha$was determined, and the RMSF across the simulation calculated for each. The results show significant similarities in the profile RMSF of each system, indicating that the mutational protocol does not adversely affect the structural dynamics across a simulation.

ular protease for both systems. This is to be expected because these residues are tightly-packed, with many surrounding atoms, restricting their torsional, rotational and translational freedom of movement. The regions of higher RMSF are distal to these inflexible core residues, and generally are solvent-accessible, while the residues with the highest fluctuations are located at the external β-turn structures around the flaps, flap-elbows, cantilever and fulcrum. These regions are known to be highly mobile in molecular dynamics simulations [77, 33]. Therefore these results show that the structural dynamics across the simulations were not impeded through the mutational protocol and that the motions observed were in concurrence with theoretical dynamics. Minor differences observed between RMSF profiles of each system and the un-mutated system (Figure 5.1) were compared against the location of the mutated residues to determine whether they were correlated. The results showed that the regions of RMSF variation did not significantly coincide with the location of the mutations. For example, the RMSF differences observed between the un-mutated system and the 13-mutations system (Figure 5.2) occur around the extremities of the protease structure, but these do not coincide with the location of the distal mutations. This suggests that the observed RMSF variations were not an artefact of the mutation but rather due to natural configurational fluctuations of the protease.

An important note about the results of this mutational chain is the mutational protocol was not performed on just a single PDB starting structure; the chain contained systems with mutations in both 1HXB and 1FB7 crystal structures. Therefore, as all the systems showed overlapping profile RMSF values, this suggests that mutational protocol is transferrable between crystal structures. However, this does not indicate that all protease crystal structures can withstand 13 mutations in each monomer, nor that other crystal structures can be confidently mutated at all. Nevertheless, 1FB7 and 1HXB currently represent the only two crystal structures complexed with saquinavir, so these results suggest that the mutation protocol is robust enough for the protease systems complexed with saquinavir. The genetic diversity between the HIV-1 protease amino acid sequences of different subtypes is approximately 5%-6% of the total sequence, which means that HIV-1 subtypes differ from each other by approximately 6 residues in their protease protein sequence [45]. The mean number of saquinavir-associated mu-



(a) (b)

(c) (d)

Figure 5.2: Cartoon representation of un-mutated HIV-1 protease ((a) and (b)), and 1FB7 with 13 mutations in each monomer ((c) and (d)). (a) and (c) show the regions of average RMSF > 1.0Å (in green) and < 0.5Å (in red) across the proteases' 2ns simulation. Intervening regions (0.5Å $\leq$ RMSF ≤ 1.0Å) are shown in grey. (b) and (d) show the locations of the proteases' mutations with respect to Hxb2 consensus sequence as orange *van der Waals* spheres. This is to give an indication of the relationship between the ΔRMSF and the location of the mutations in the quaternary structure. Images were created using VMD from the final outputted configuration of each simulation.

tations in protease during anti-retroviral therapy is 3.8 [57], of which only one or two key mutations are necessary for resistance to saquinavir [32]. Therefore the ability to mutate up to 13 mutations in a crystal structure means that a large proportion of the important mutations across multiple subtypes can be represented computationally.

Having determined the dynamical stability of the mutated systems, the binding affinities were calculated over the 2ns production-phase through MMPBSA and MMGBSA methodologies combined with normal mode analysis (Section 2.3.2). These were then compared against the equivalent experimentally-derived K_i values, which were extracted from the local database and converted into a comparable ΔG value (Equation 1.26). The MMPB(GB)SA ΔG values for each system were calculated as the average of 200 equally-spaced snapshots over the 2ns simulation, and the normal mode analysis calculated as the average of 10 equally-spaced snapshots. The error bars of the resultant ΔG value were calculated as:

$$\sigma_T = \sqrt{\sigma_H^2 + \sigma_S^2} \tag{5.1}$$

where σ_T is the standard deviation of the ΔG value; σ_H is the MM(PB/GB)SA standard deviation; and σ_S is the normal mode standard deviation. Experimentally-derived data have no error bars because there were no standard deviations associated with the ITC or EIA data in The Binding Database (Figure 5.3). Therefore there were no experimental errors by which to compare the computational results.

The results showed that the MMPBSA methodology consistently returned average binding affinity values closer to experimental values than MMGBSA. In every system the MMGBSA results were $\sim$ 11kcal/mol more negative than the corresponding MMPBSA results, and shared the same error bar size. This is to be expected because the Generalised-Born equation implemented in MMGBSA for calculating the accessible surface area of the protein to the solvent is an approximative model to the exact, but more computationally-demanding, Poisson-Boltzmann equation implemented in MMPBSA [73]. The error bars were similar in magnitude because the majority of the standard deviation was due to the configurational entropy calculated by normal

mode analysis, which was common to both methods. As both methods themselves are heuristic, and therefore trade absolute accuracy of other methods, such as Thermodynamic Integration (TI), for an increase in turnaround of results, both continuum solvation methods were implemented in this study to determine whether the more relatively-computationally demanding MMPBSA method would yield more accurate results than MMGBSA. The results showed that in all circumstances the MMPBSA method performed considerably better - in three cases attaining values whose corresponding experimental values fell within its error bars. In no circumstances did this happen with the MMGBSA results, with the closest value 7.0kcal/mol more negative than experimental. Therefore, due to the Poisson-Boltzmann equation outperforming the Generalised-Born methodology in all circumstances, combined with the increase in computational resource available due to greater access to Grid computers and supercomputers described in Section 2.4, the MMPBSA methodology was utilised for all subsequent energy calculations, and the MMGBSA methodology was not pursued any further.

In comparing the computed MMPBSA results to the experimentally-determined values (Figure 5.3), two main trends can be observed: for the first 7 systems (un-mutated to 6 mutations) the computational results fluctuate around the experimental results, ranging from 5kcal/mol more negative than experimental at system 0, to 5kcal/mol less negative than experimental at system 3. For systems 2 and 4, the experimental values lie within the computational error bars, and for systems 5 and 6 they fall just outside of computational error. However, the second 7 systems show a markedly different trend, with all MMPBSA values consistently 7kcal/mol more negative than experimental. The correlation coefficient between MMPBSA results and experimental data across the entire chain is $r = -0.24$, but when considering the first and second 7 systems independently, the correlation coefficients become $r = -0.14$ and $r = 0.75$ respectively (Figure 5.4). This shows that the second 7 systems' ΔG values were strongly correlated with the experimental data, whilst the first 7 were not. In order to investigate the cause of these results, the MMPBSA and configurational entropy components of the computational ΔG were examined (Table 5.3).

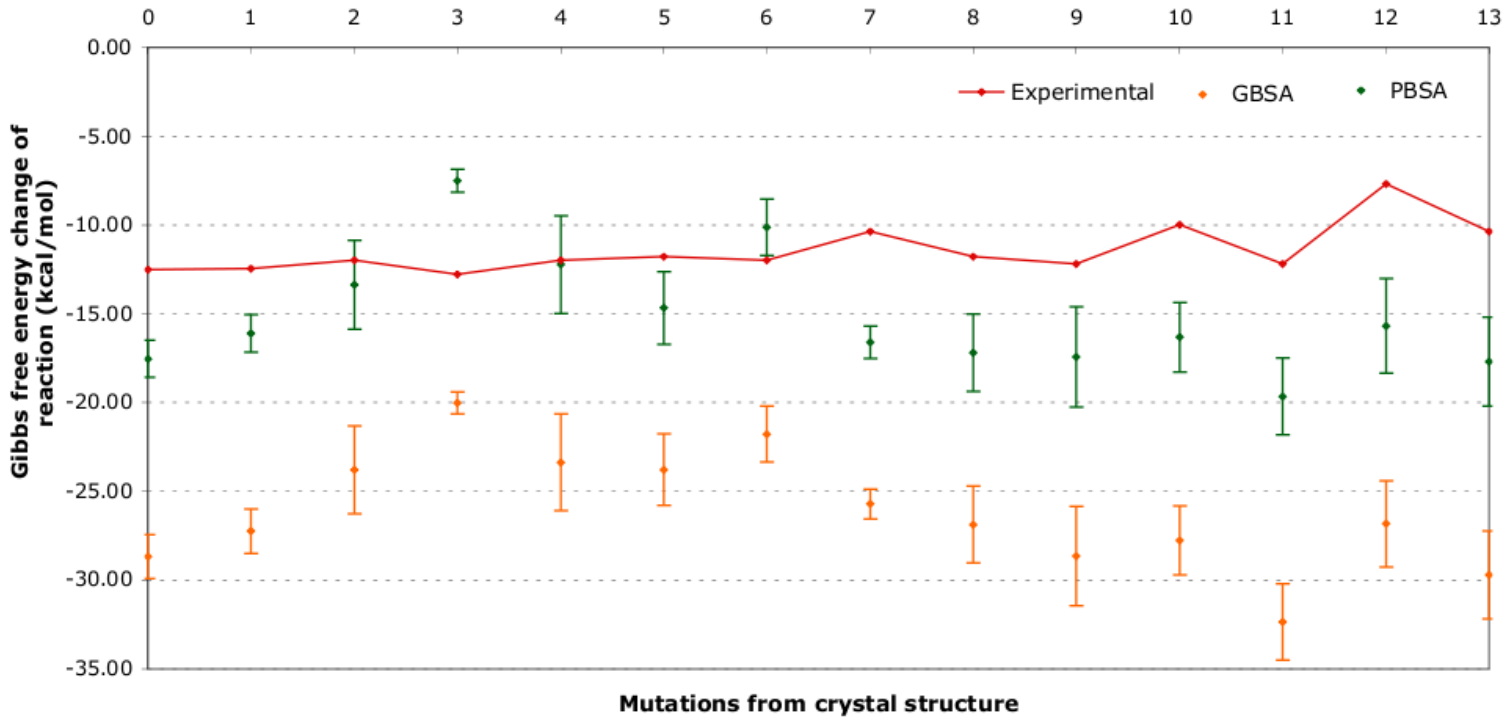
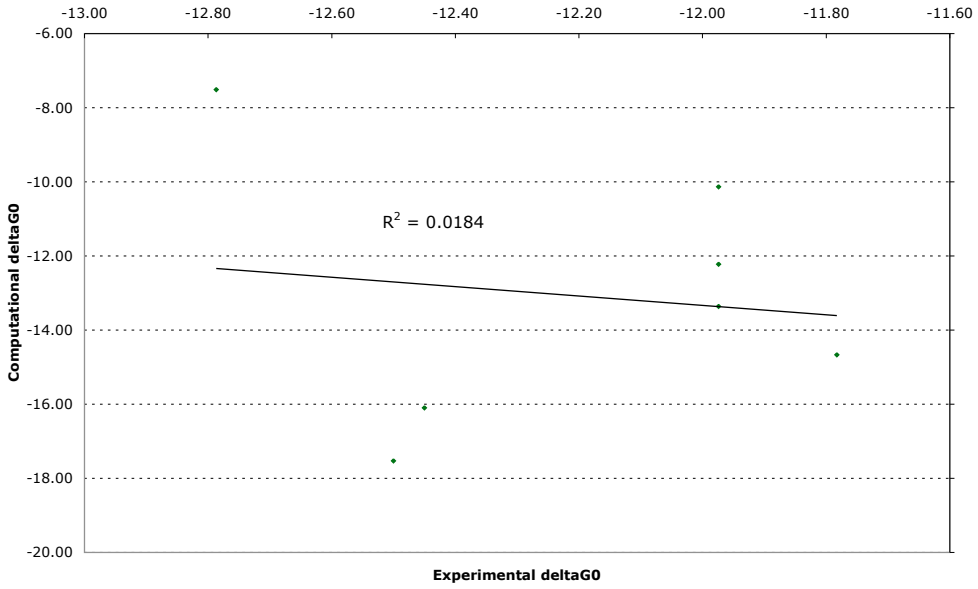
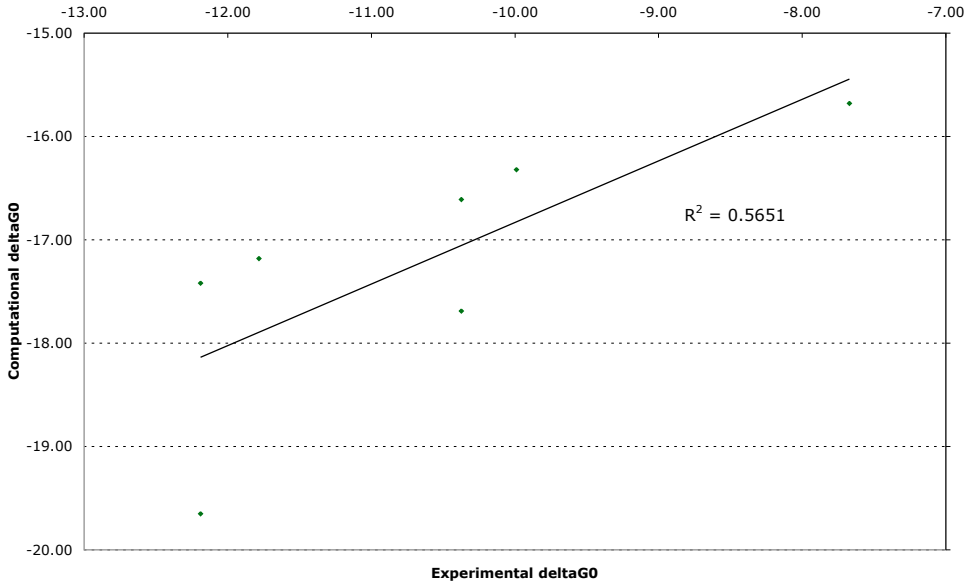


Figure 5.3: Comparison between computationally-determined MMPBSA (green) and MMGBSA (yellow), and experimentally-derived ΔG values (red) for each of the 14 mutational-chain systems. The results show consistent outperformance of the MMPBSA methodology over MMGBSA, with values $\sim$ 11kcal/mol closer to experimental. Two trends can be distinguished in comparison of the MMPBSA to experimental data; between systems 0 and 6 the computational results fluctuate around the experimental values, but are unable to follow the changes in ΔG between systems (Figure 5.4(a)). From system 7 to 13, the trend changes such that the MMPBSA values are consistently $\sim$ 7kcal/mol more negative than experimental, but consistently reproduce the change in ΔG between systems (Figure 5.4(b)).



(a)



(b)

Figure 5.4: Correlations between computational free energy changes, calculated through MM/PBSA, and the corresponding experimental free energy changes for systems 0 to 6 (a) and 7 to 12 (b) of the mutational-chain. The linear regression lines show the lines of best fit between the correlative values. Each correlation's coefficient of determination is also given; this is the square of the correlation coefficient, which is -0.14 for (a), and 0.75 for (b). This indicates there is no correlation between computational and experimental results for the first 7 systems, and a strong positive correlation for the second 7 systems.

Table 5.3: Comparison of the thermodynamic components of computed ΔG to the experimental ΔG

System	Thermodynamics (kcal/mol)			
	$\Delta H_C{}^a$	$T\Delta S_C{}^b$	$\Delta G_C{}^c$	$\Delta G_E{}^d$
0	-42.22	-24.69	-17.53	-12.50
1	-45.30	-29.20	-16.10	-12.45
2	-39.32	-24.06	-13.36	-11.97
3	-39.64	-32.13	-7.51	-12.79
4	-38.81	-26.59	-12.22	-11.97
5	-45.30	-30.63	-14.67	-11.78
6	-45.38	-35.25	-10.13	-11.97
7	-42.26	-25.65	-16.61	-10.37
8	-49.51	-32.33	-17.18	-11.78
9	-43.22	-25.80	-17.42	-12.19
10	-46.74	-30.42	-16.32	-9.99
11	-46.96	-27.31	-19.65	-12.19
12	-45.37	-29.69	-15.68	-7.67
13	-48.40	-30.71	-17.69	-10.37

[a] Computationally-determined using MMPBSA methodology.
[b] Computationally-determined using normal mode analysis.
[c] Computationally-determined as $\Delta H_C - T\Delta S_C$.
[d] Experimentally-derived through EIA or ITC.

The decomposition of the ΔG reveals that the ΔH and ΔS components are equally variable across the mutational chain, with ranges of 10.70kcal/mol and 11.19kcal/mol respectively. However, the strong correlation previously observed between the second 7 systems' computational ΔG and experimental ΔG is lost upon decomposition. The correlation between the second 7 systems' ΔH and experimental data was $r = 0.12$, and for the systems' TΔS the correlation was $r = -0.25$. Therefore the strong correlation observed for the ΔG was due to the combination of the two components rather than a particular contribution.

The cause of the change in trends between the first and second 7 systems could not be explained through the stabilising effect of a mutation common to the second 7 systems. Visual inspection of the systems through VMD also did not reveal any significant structural differences between the systems, such as as more open configuration of the flaps

Table 5.4: Decomposition of the enthalpic energy values for the mutational-chain systems

System	Enthalpic energy (kcal/mol)				
	VdW^a	Ele^b	$PB_{sur}{}^c$	$PB_{cal}{}^d$	$PB_{tot}{}^e$
0	-68.50	-32.90	-8.35	67.53	-42.22
1	-71.93	-44.52	-8.43	79.57	-45.30
2	-69.15	-25.88	-8.31	64.02	-39.32
3	-73.03	-42.40	-8.19	83.99	-39.64
4	-66.46	-35.39	-8.43	71.47	-38.81
5	-75.75	-32.89	-8.30	71.64	-45.30
6	-72.13	-42.80	-8.42	77.96	-45.38
7	-73.88	-37.02	-8.28	76.92	-42.26
8	-73.21	-42.86	-8.53	75.09	-49.51
9	-76.39	-32.72	-8.49	74.37	-43.22
10	-73.27	-43.89	-8.45	78.87	-46.74
11	-73.91	-47.59	-8.49	83.02	-46.96
12	-71.87	-40.64	-8.58	75.72	-45.37
13	-73.18	-51.84	-8.68	85.29	-48.40

[a] Non-bonded *van der Waals* energy.
[b] Non-bonded electrostatic energy.
[c] Non-polar contribution to solvation free energy.
[d] Polar solvation contribution to solvation free energy.
[e] Change in enthalpy upon inhibitor binding.

[77], or the lateral movement of saquinavir out of the active site [96], and the profile RMSF analysis showed that the dynamics of all 14 systems were similar. Therefore, the underlying cause of the strong correlation to experimental data in systems 7 to 13 remains unclear.

5.3 Conclusions

Simulations of particular HIV protease genotypes are either run from crystal structures of the genotype of interest, or, less commonly, an alchemical mutation of 1 or 2 residues is performed on the nearest crystal structure [77, 100, 96]. Both methods are strongly reliant on the experimental generation of crystal structures. However, there are currently only approximately 230 HIV protease structures in the Protein Data Bank, which is not enough to cover the range of possible genotypes *in vivo*. Therefore, to extend the range of simulations beyond the key primary mutations that are currently simulated to

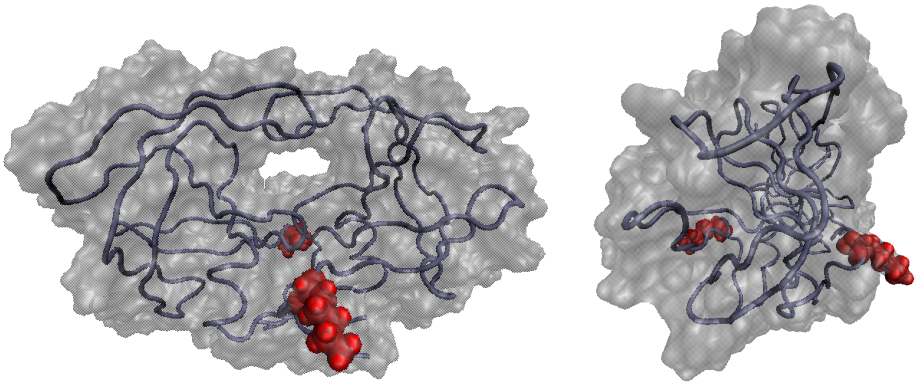
Table 5.5: Decomposition of the entropic energy values for the mutational-chain systems

System	Configurational entropy (kcal/mol)			
	$\mathbf{TS_{tra}}$	$\mathbf{TS_{rot}}$	$\mathbf{TS_{vib}}$	$\mathbf{TS_{tot}}$
0	-13.58	-11.81	0.70	-24.69
1	-13.58	-11.82	-3.79	-29.20
2	-13.58	-11.79	1.31	-24.06
3	-13.58	-11.81	-6.74	-32.13
4	-13.58	-11.85	-1.16	-26.59
5	-13.58	-11.81	-5.24	-30.63
6	-13.58	-11.83	-9.84	-35.25
7	-13.58	-11.82	-0.25	-25.65
8	-13.58	-11.85	-6.90	-32.33
9	-13.58	-11.81	-0.40	-25.80
10	-13.58	-11.84	-5.00	-30.42
11	-13.58	-11.82	-1.91	-27.31
12	-13.58	-11.83	-4.28	-29.69
13	-13.58	-11.88	-5.25	-30.71

[a] Change in translational entropy upon inhibitor binding.
[b] Change in rotational entropy upon inhibitor binding.
[c] Change in vibrational entropy upon inhibitor binding.
[d] Change in the total entropic contribution to free energy upon inhibitor binding, calculated as $TS_{tra} + TS_{rot} + TS_{vib}$.

genotypes containing multiple accessory mutations, a greater number of residues will need to be alchemically-mutated. However, there has not yet been a study examining the effect of mutating multiple residues in HIV protease.

The results in this chapter showed that the mutation protocol followed was able to mutate up to 26 residues in the homodimer without adversely affecting the proteases' quaternary structures. Furthermore, the dynamics of alchemically-mutated proteases were almost identical to those of un-mutated proteases over 2ns. Two different starting crystal structures were mutated within the mutational-chain with structural and dynamic results comparable to un-mutated. Therefore the mutation protocol was shown to be robust enough to be transferrable between HIV protease systems. However, these results cannot be interpreted for alternative proteins, such as HIV reverse transcriptase, where residues may have less freedom of movement, or adverse interactions between the mutated side-chain and surrounding residues results in a non-physiological structure.



(a) (b)

Figure 5.5: Anterior (a) and lateral (b) cartoon views of HIV protease, showing the backbone structure in ice-blue and the surface outline in grey. The side-chain of lysine at residue 7 in each monomer is shown as red *van der Waals* spheres. The location of this residue is at a β-turn near the dimer interface, distant to the active site, and superficial to the protease. The side-chain points out into the solvent. Figure was created from the final outputted configuration of system 12 (See Table 5.1) at 2ns production-phase using VMD.

The MMPBSA binding affinity results revealed 2 different relationships between the mutational-chain and experimental data; the first 7 systems gave results that fluctuated around their corresponding experimental values, but showed no correlation to the $\Delta\Delta$G between consecutive experimental datum. The second 7 systems, conversely, showed a strong correlation to the experimental $\Delta\Delta$G, but was consistently 7kcal/mol more negative than the corresponding experimental value. A possible cause of change in relationship was shown to be the *van der Waals* interactions between the protease and ligand, however, this could not be localised to a common mutation between the second 7 systems. Although the underlying cause of the differing relationships to the experimental data was not ascertained, these results were still encouraging as 8 of the 13 consecutive computational $\Delta\Delta$G values gave the same trend as the equivalent experimental $\Delta\Delta$G. Furthermore, only systems 11 and 13 were incorrectly calculated as having a more negative ΔG value than the 1HXB un-mutated system, which contains 2 differences to wild-type: I3V and N37S, both of which are natural polymorphisms that confer no drug resistance on their own [105, 77]. These results suggests that the methodology is able to recognise the alchemical mutations' effect on the strength of

interaction between the protease and saquinavir. Therefore, even though the methodology is unable to quantify the absolute binding affinity value correctly, if it is able to correctly identify the directional effect of mutations on the binding affinity, then it has clinical applications. As a result, it was decided to apply the simulation protocols and MMPBSA/normal mode analyses from this chapter to a smaller chain of protease genotypes containing an increasing number of drug-resistant mutations observed *in vivo* to determine whether it could correctly order their binding affinities to saquinavir.

Chapter 6

Computational reproduction of a series of increasing drug-resistant proteases

6.1 Introduction

In Chapter 5.2 it was shown that up to thirteen residues could be computationally mutated in each of protease's monomers whilst still retaining backbone fluctuations that correlate with an un-mutated simulation. This has practical implications because it validates the ability to mutate the constrained set of starting 3-dimensional structures into potentially any HIV protease genotype observable *in vivo*. However the binding affinity values calculated between HIV protease and saquinavir were not able to consistently replicate *in vitro* values calculated through biochemical techniques such as enzyme inhibition assay (EIA) or isothermal titration calorimetry (ITC). The results showed that on mutating up to six residues in each monomer, the binding affinity values calculated through MMPBSA and normal mode analysis produced values within the bounds of experimental error, but inconsistently. Beyond six mutations, however, there appeared to be a shift such that the computational results followed the trend of changes in binding affinity, but the absolute values were consistently around 7kcal/mol more negative than experimental. Though the underlying cause for these observations was not determined, each half of the mutational chain contained qualities that were

promising for the continuation of the methodology to determine whether MMPBSA and normal mode analysis can replicate the decrease in drug sensitivity of a series of protease systems containing an ordered increase in protease-inhibitor resistance mutations.

The motivation for this study came from a paper by Ohtaka *et al.* (2003) who performed high-precision isothermal titration calorimetry to titrate one of six different protease inhibitors (indinavir, nelfinavir, saquinavir, ritonavir, amprenavir and lopinavir) into a solution of one of six HIV protease genotypes. From the outputted data they calculated the change in free energy, enthalpy, and entropy of inhibitor binding for each of the 36 combinations of inhibitors and proteases [69]. From this data-set the 6 saquinavir systems were most pertinent to this study, as previous binding affinities had been calculated with respect to saquinavir. Table 6.1 shows the the subset of the data used as reference values for the computational calculations.

Table 6.1: Mutations and associated thermodynamic values for the 6 HIV-1 protease genotypes complexed with saquinavir published in Ohtaka *et al.* (2003).

Protease genotype	**AKA**	**Thermodynamics (kcal/mol)**		
		ΔG	ΔH	$-T\Delta S$
wild-type	WT	-13.0	1.2	-14.2
L10I, L90M	DM	-12.0	3.6	-15.6
M46I, I54V	FL	-11.9	5.1	-17.0
V82A, I84V	AS	-11.8	3.7	-15.5
M46I, I54V, V82A, I84V	QM	-10.0	8.5	-18.5
L10I, M46I, I54V, V82A, I84V, L90M	HM	-8.5	10.4	-18.9

This table shows not only the change in binding affinity for each genotype, but also the thermodynamic mechanism by which the mutation causes a reduction in drug affinity. For example, the [L10I, L90M] genotype has a more positive change in free energy upon formation of a complex with saquinavir than the wild-type, indicating that this binding is less likely to happen spontaneously. Examining the enthalpic and entropic contributions shows that the cause of this less negative ΔG is a more positive change in enthalpy. The entropic contribution is actually more negative, but this is not enough to offset the positive increase in enthalpy, so overall the free energy change is less negative.

For clarity, each protease genotype is given a two-letter short-hand code (see Table 6.1): the wild-type genotype is termed WT, the [L10I, :L90M] residues are located in the dimerisation (DM) region of the protease, the [M46I, I54V] residues are located on the flaps (FL), the [V82A, I84V] residues are located in the active site (AS), the [M46I, I54V, V82A, I84V] genotype is termed the quadri-mutant (QM), and the final genotype is termed the hexa-mutant (HM). The location of these mutations on the quaternary structure is shown in Figure 6.1.

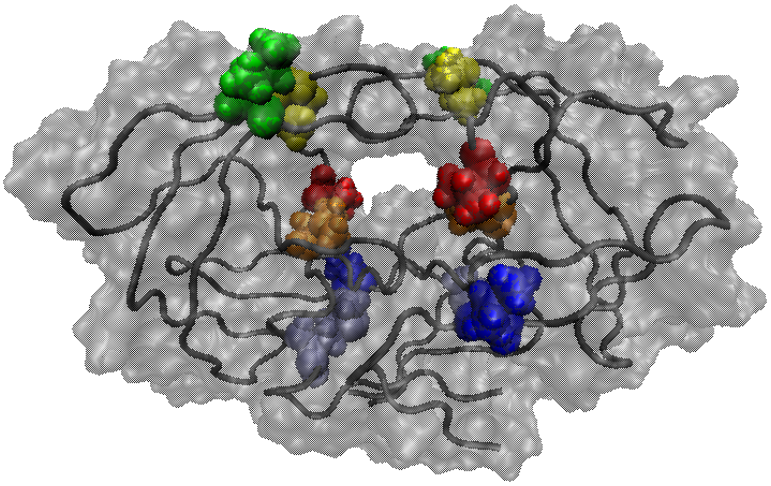


Figure 6.1: Location of the 6 mutations associated with Table 6.1 on HIV-1 protease's quaternary structure. The backbone structure and surface-area outline are shown in grey; for each monomer: residue 10 is shown in dark-blue; residue 46 is shown in green; residue 54 is shown in yellow; residue 82 is shown in red; residue 84 is shown in orange; and residue 90 is shown in ice-blue. Image was created in VMD from structure 1FB7.

Combinations of these mutations occur frequently in patients undergoing HAART therapy, with the HM genotype resulting in multi-drug resistance to the 6 inhibitors described in the paper by Ohtaka *et al.* [69]. This gives these mutations clinical relevance and therefore a good template by which to determine whether MMPBSA and normal mode analysis of MD simulations is able to replicate trends in drug-resistance caused by specific mutations.

6.2 Reproduction of trend of drug-resistance

Each of the 6 genotypes was created by applying the mutation protocol (Section 2.1.1) to the α-chain mono-protonated 1FB7 crystal structure. Thermodynamic analysis of the four different possible protonation states of the two catalytic aspartic acids showed that the most thermodynamically-favourable protonation state was with the α-chain protonated and the β-chain un-protonated (data not shown), consistent with the literature [80, 125]. Therefore this protonation state was followed for generation of the proteases' initial configurations. Following the 2 nanosecond equilibration protocol, simulations were run for a further 10 nanoseconds for data analysis. The reason for the increase in the length of the data analysis beyond the 2 nanoseconds previously run was because data from the mutational chain study (Section 5.2) showed that extending the simulations to 2 nanoseconds was insufficient for consistently calculating a binding affinity value comparable to *in vitro* experimental data. This may have been due to insufficient configurational sampling over 2ns. Data published in the literature also suggested that low-frequency motions that involve important configurations only occur across longer timescales [33]. This coincided with access to a greater number of Grid computers, including TeraGrid resources, that made running simulations for this length of time feasible. Each of the 6 protease genotypes was therefore simulated in complex with saquinavir for 12 nanoseconds (2 nanoseconds equilibration protocol followed by 10 nanoseconds production-phase) and then the average binding affinity value across a simulation's production-phase calculated using MMPBSA to compute the enthalpic component of the free energy, and normal mode analysis to compute the configurational entropic component. The MMPBSA calculation was averaged from 1000 snapshots over the 10 nanosecond simulation, with 100 equally-spaced snapshots analysed every nanosecond. The normal mode calculation was performed on 50 snapshots, with only 5 equally-spaced snapshots per nanosecond due to the increased computational expense of calculation.

Prior to comparing the computational binding affinities to the experimental ones, determination of equilibration was achieved through examination of the stability of the ΔH and TΔS components of the binding free energy across each simulations, and also

the stability of the energy terms that comprise the ΔH such as the electrostatic and *van der Waals* forces between the protease and saquinavir. If the values have not stabilised then the protease has not reached equilibrium and so any thermodynamic value may be erroneous. Figure 6.2 shows the evolution of the component terms that comprise the enthalpic value outputted by MMPBSA. It can be seen that in all systems the component energies are stable across the entire 10 nanosecond simulation. This shows that the mutation and equilibration protocol are sufficient for each system to reach a stable value by the start of the data collection. Also of note are the absolute values for each system; each Y-axis has the same range, and it can be seen that while the non-polar solvation free energies, *van der Waals* energies, and ΔH values are equivalent between systems, the electrostatic energies (red) and polar contributions to solvation free energies (yellow) are the most variable between system. This supports the results seen in Section 5.2 where the major differences observed between systems arose from these same two enthalpic components. However, unlike in the mutational-chain study, none of the systems contain mutations to a lysine residue. In fact, none of the systems undergo changes in side-chain polarity: in the DM system the two neutral non-polar leucine residues are mutated into neutral, non-polar isoleucine and methionine residues; in the FL system the neutral, non-polar methionine and isoleucine residues are mutated into neutral, non-polar isoleucine and valine residues respectively; in the AS system the neutral, non-polar valine and isoleucine residues are mutated into neutral, non-polar alanine and valine residues respectively; and the QM and HM systems are comprised of variants of each of the double-mutants. Therefore the simulated systems' differences in electrostatics with respect to the wild-type system are not a direct indication of the change in non-bonded electrostatic force between the drug-resistance mutation and the inhibitor. However the change in electrostatic force between the protease as a whole and the inhibitor causes the major source of difference in ΔG between simulated protease genotypes.

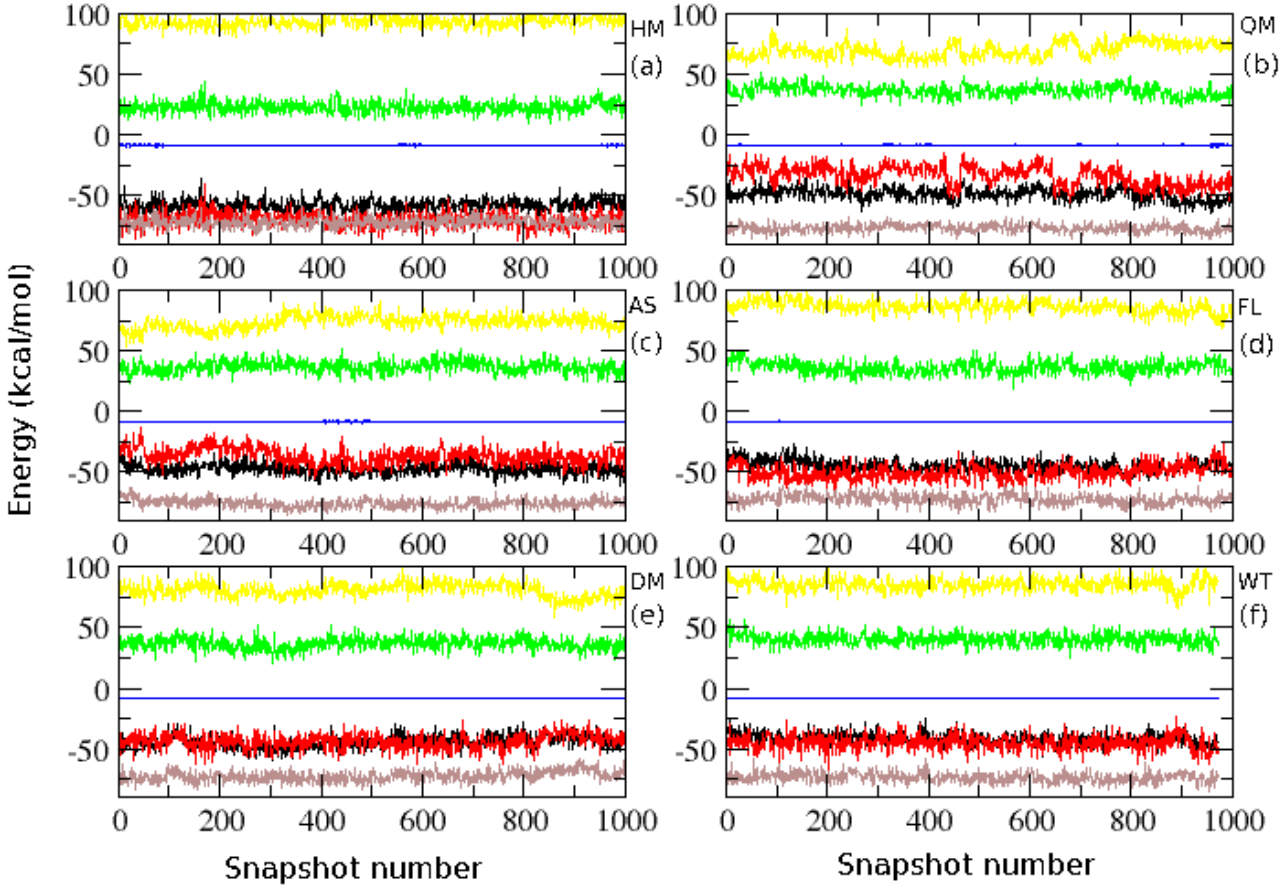


Figure 6.2: Decomposition of the enthalpic contribution to the binding affinity for HM (a), QM (b), AS (c), FL (d), DM (e) and WT (f) across the 10 nanosecond simulations. Decomposed energies are: eletrostatic (red), EleTOT (green), non-polar solvation energy (blue), polar solvation energy (yellow), *van der Waals* (brown), and ΔH (black).

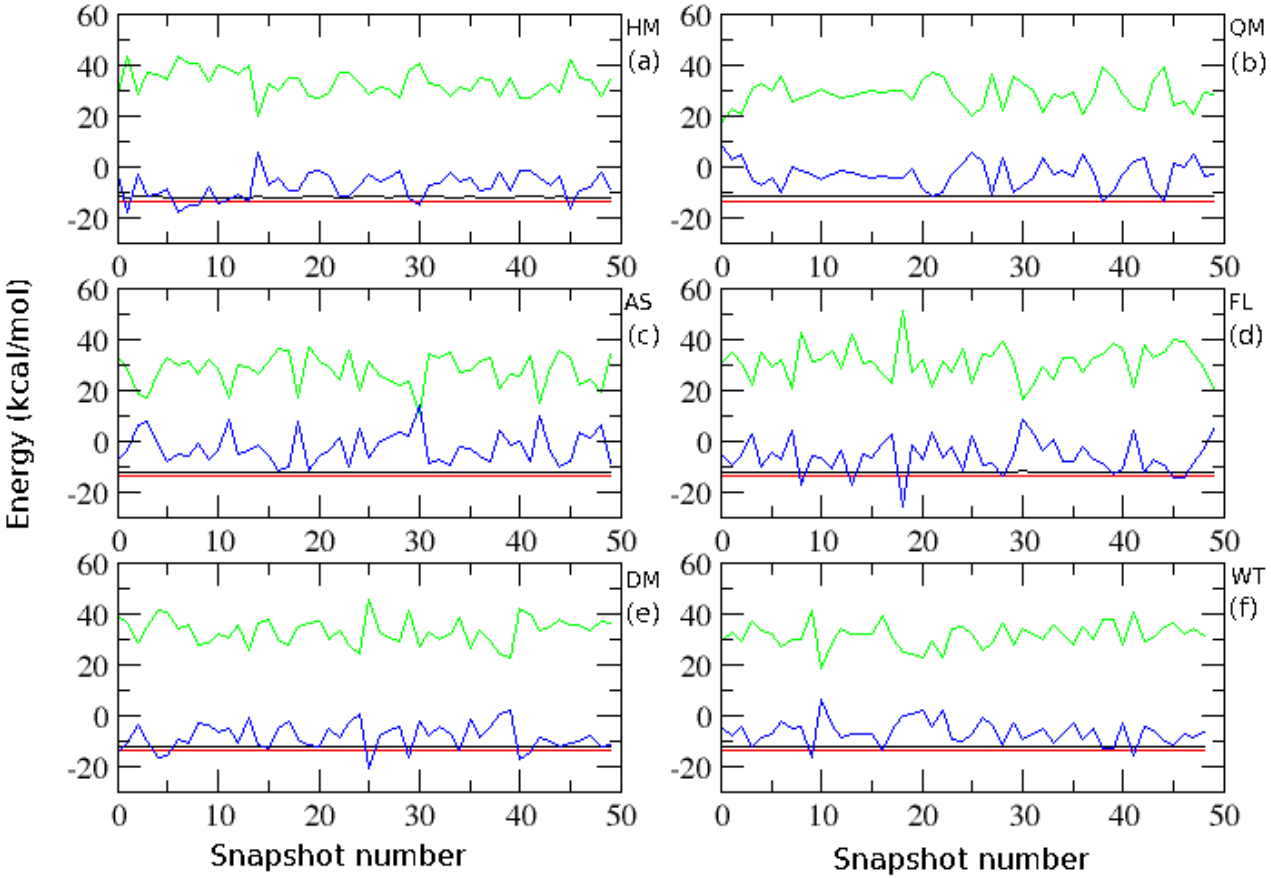


Figure 6.3: Decomposition of the entropic contribution to the binding affinity for HM (a), QM (b), AS (c), FL (d), DM (e) and WT (f) across the 10 nanosecond simulations. The component entropies are: $T\Delta S_{tra}$ (red), $T\Delta S_{rot}$ (black), $T\Delta S_{vib}$ (blue), and $-T\Delta S_{total}$ (green). The $-T\Delta S_{total}$ is the sum of the other component terms.

The component entropic terms ($T\Delta S_{vib}$, $T\Delta S_{tra}$, $T\Delta S_{rot}$, and $-T\Delta S_{total}$ which is the configurational entropy), as with the entropic component terms in the mutational-chain study (Section 5.2) and the rotational and translational components of the configurational entropy, are almost constant across the 10 nanosecond simulation, and show very little variation between systems (Figure 6.3). The vibrational component however is more variable between systems, and unlike the rotational and translational components, is not constant across a simulation. The variability between the systems is expected because their differing genotypes results in an altered numbers of atoms in the protein, and so subsequently a differing number of possible configurations that the protease can adopt. The variation in the vibrational entropy within a simulation is not unique to HIV-1 protease, and does not indicate that the entropies have not converged. The $T\Delta S_{vib}$ variation across a simulation is attributed to the methodology of the normal mode analysis, which is implemented to calculate it. This does not calculate the vibrational entropy of the protein itself, but rather of the most accessible local minimum. This therefore assumes that the protein resides in a single local-minima across the whole simulation, which causes variation to occur when the protein configurationally flexes, resulting in a different local minima to become most accessible to it. Calculation of the rotational and translational entropies uses a different methodology that does not rely on the energy landscape, and so does not vary in the same way [106, 74, 26]. Even though the vibrational entropy shows variation between snapshots within a simulation, the mean value around which they vary does not drift through the course of the simulation.

In addition to determining the convergence of the energy terms as a measure of simulation reliability, the stability of each system's quaternary structure across their respective simulation was determined through global RMSD and global RMSF calculations. RMSD was performed with respect to each system's post-mutation, pre-minimisation structure, and results showed that each system rapidly adopted a structure ~ 1.0Å away from this initial structure, about which they stably fluctuated around. RMSF was performed with respect to each system's average structure, which was generated from the time-averaged position of the C_α atoms across the 10ns data-collection phase using VMD. The results showed that all systems fluctuated stably around their average

structure, with an average peak-to-peak amplitude of $\sim 0.5\mathring{A}$. Having ascertained that the data-collection phase of the simulations showed structural and energetic stability, the computed binding affinity values can be confidently compared against the experimental values.

The ΔG results from the simulations are shown in Table 6.2 compared to the corresponding experimental results [69]. The computed ΔH_C and $T\Delta S_C$ are not compared to experimental values because the MMPBSA methodology calculates the solvation electrostatic term as a *free energy* which is not decomposed into the enthalpic and entropic contributions, meanwhile the TΔS value is only a calculation of the change in configurational entropy upon complex formation [49].

Table 6.2: Calculated ΔG_C and its constituent ΔH_C and $T\Delta S_C$ in comparison to the experimental ΔG_E published by Ohtaka *et al.* (2003).

System	**Thermodynamics (kcal/mol)**				
	ΔH_C[a]	$T\Delta S_C$[a]	ΔG_C[a]	ΔG_E[b]	$\Delta\Delta G$[c]
WT	-41.90	-31.74	-10.16	-13.00	+2.84
DM	-44.39	-33.59	-10.80	-12.00	+1.20
FL	-44.50	-31.31	-13.19	-11.90	-1.29
AS	-47.25	-27.68	-20.62	-11.80	-8.82
QM	-49.35	-28.43	-20.92	-10.00	-10.92
HM	-57.99	-33.23	-24.76	-8.50	-16.26

[a] computationally-determined. [b] experimentally-determined.
[c] ΔG_C - ΔG_E.

Comparison of the computed ΔG to the experimentally-derived ΔG shows an anti-correlation, with the computational results calculating the ΔG of the HM system to be 14.60kcal/mol more negative than the WT system. Therefore the computational results are presenting the multi-drug resistant HM system to be *more* strongly attracted to the inhibitor drug than the wild-type system. The computational results also show the three double-mutant genotypes to be more attracted to saquinavir than the wild-type, but less than the MDR genotype. The correlation coefficient between the computational and experimental results is $r = -0.86$ (Figure 6.4). That such a strong anti-correlation is formed over 6 different systems suggests that the methodology is able to recognise

an effect caused by the mutations, but is unable to quantify it properly. These results are particularly interesting in the context of concurrent simulations run and analysed by collaborators using the same protease genotypes and pre-simulation protocols, but complexed with the alternative inhibitors published in Ohtaka *et al.* (2003): ritonavir, lopinavir, indiavir, nelfinavir, and amprenavir. MMPBSA and normal mode analysis averaged over 10ns resulted in correlation coefficients of $r = 0.93$, $r = 0.81$, $r = 0.67$, $r = 0.44$, and $r = -0.79$ respectively for the above inhibitors complexed with the 6 protease genotypes. These represent a spread of results ranging from strongly positively correlated to strongly negatively-correlated, but none show a significant lack of correlation implying that the ability of the computational methodology to quantify the mutations' effects are inhibitor-specific.

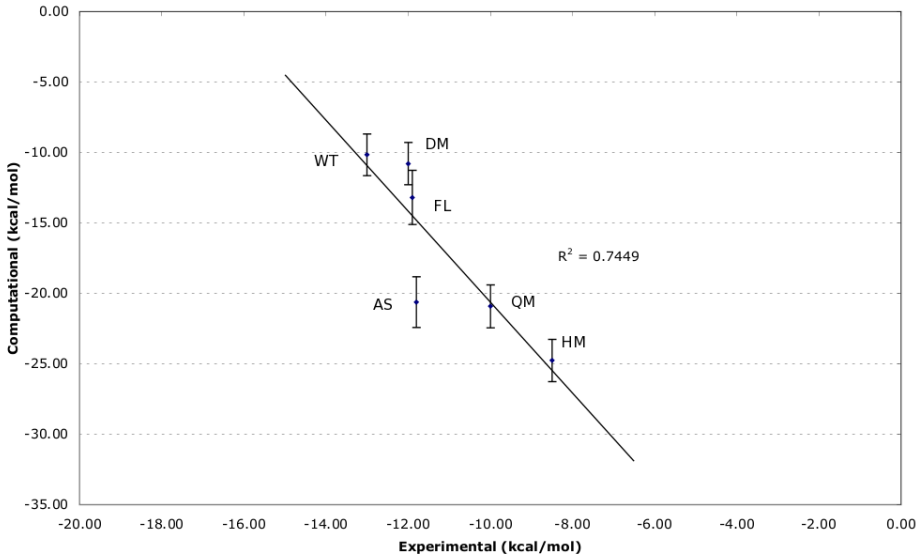


Figure 6.4: Correlation between computational and experimental ΔG values. The graph shows strong anti-correlation between computational and experimental results, with a correlation coefficient $\sqrt{R^2} = r = -0.86$. The experimental results do not contain error bars as none were published [69].

Closer examination of the decomposed binding affinity values (Table 6.2) show that the anti-correlation observed between ΔG_C and ΔG_E is due to the enthalpic component of the calculated ΔG_C, with a correlation coefficient of $r = -0.97$ to the experimental data, while the configurational entropy component has a correlation coefficient of

$r = -0.07$ representing no correlation to the experimental data. The enthalpic component of the binding energy was therefore decomposed further to ascertain whether the correlation could be attributed to any specific protein-inhibitor interaction (Table 6.3).

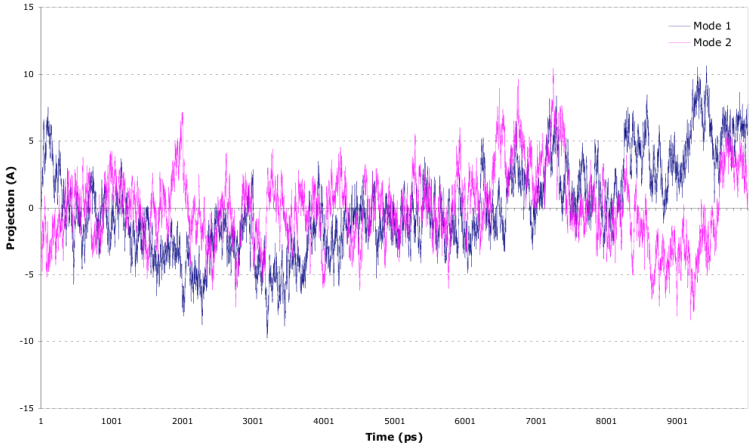
Table 6.3: Correlation between computed enthalpic energy terms and experimental binding affinities

System	**Thermodynamics (kcal/mol)**					
	Ele[a]	PB_{ele}[b]	PB_{sol}[c]	PB_{cal}[d]	VdW[e]	ΔG_E
WT	-44.93	40.24	76.77	85.17	-73.74	-13.00
DM	-44.06	36.51	72.12	80.58	-72.45	-12.00
FL	-50.05	36.15	77.97	86.20	-72.41	-11.90
AS	-36.07	36.88	64.51	72.95	-75.69	-11.80
QM	-33.60	36.08	61.26	69.67	-77.01	-10.00
HM	-70.27	22.70	84.39	92.97	-72.12	-8.50
CC[f]	-0.53	**-0.89**	0.17	0.18	-0.02	N/A

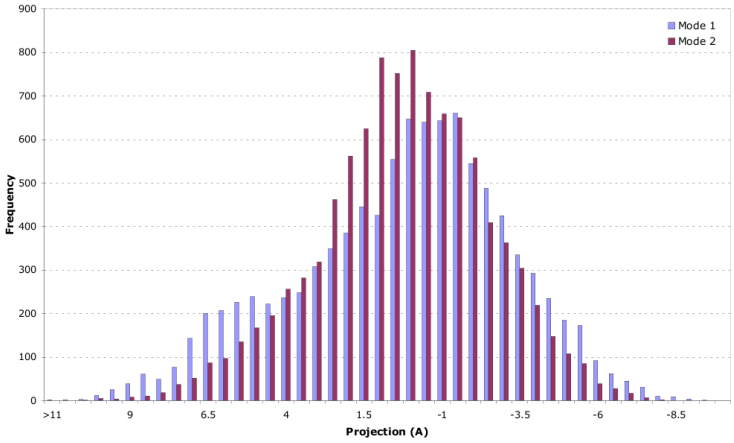
[a] is the non-bonded electrostatic energy between the protease and inhibitor.
[b] is the sum of PB_{cal} and *Ele*.
[c] is the solvation free energy calculated by the Poisson-Boltzmann method.
[d] is the hydrophobic component of PB_{sol}.
[e] is the non-bonded *van der Waals* energy between the protease and inhibitor.
[f] is the correlation coefficient (r).

The results show that differences in the PB_{ele} component of the ΔH_C value between systems is the cause of the strong anti-correlation to ΔG_E, with a correlation coefficient of $r = -0.89$. This PB_{ele} component is the sum of the electrostatic component of the solvation free energy and the non-bonded electrostatic energies between the ligand and protease. It was previously observed that the enthalpic differences between the systems were due to the non-bonded electrostatic energies (Figure 6.2), but this data shows that these differences are actually anti-correlated to the experimental binding affinities. It must be noted, however, that while this anti-correlation is strong, the error bars of the WT, DM and FL systems overlap and so the anti-correlation is manifested mainly in the AS, QM and HM results, which are significantly more attracted to saquinavir than WT (Figure 6.4). This may be due to a greater configurational landscape that needs to be sampled as the number of mutations increases in order to accurately calculate the binding affinity value. In order to determine whether the latter proteases insufficiently sampled their configurational landscape, principle component analysis (PCA)

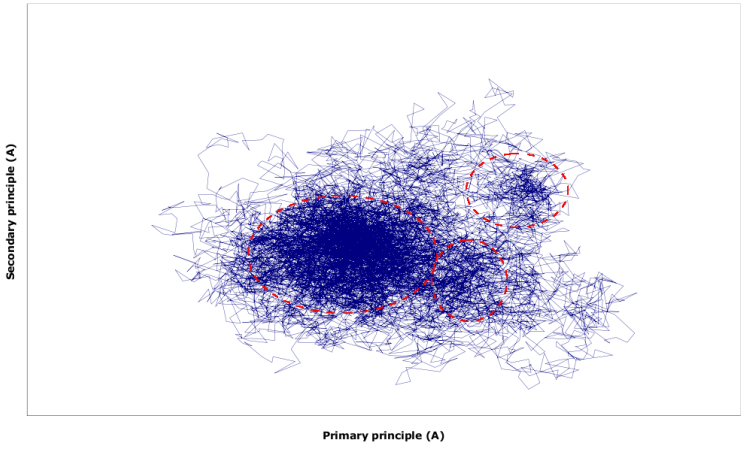
was performed on the WT and HM systems. This method is able to separate out the random high-frequency thermal fluctuations from the important lower-frequency correlated structural motions which allow analysis of sampled configurations over their trajectories. The projection of WT and HM's two most significant modes were plotted against time (Figures 6.5(a) and 6.6(a)), revealing that neither system were sufficiently sampling their configurations: the WT system stably sampled configurations for approximately 6ns, whereupon the secondary mode begin to show drift towards previously unsampled conformation. Meanwhile, the HM system's primary mode took approximately 3ns to begin stably sampling configurations, whilst its secondary mode begins drifting by 8ns. These results are best described by the frequency at which the projections are sampled over the simulation. The frequency distribution of the primary and secondary PCA modes for the WT system reveal a normal distribution for the secondary mode's projections, but with the primary mode's projections showing a slight bimodal distribution (Figure 6.5(b)). The frequency distribution for the HM system, however, shows a much less ordered sampling, with the primary mode's projections showing positive skew and the secondary mode showing negative skew (Figure 6.6(b)). This suggests that the 10ns simulation is insufficient for the HM system to effectively sample its conformational landscape. Although Figures 6.5(c) and 6.6(c) are not directly comparable to one another, the increased number of distinct conformations in the HM system's landscape over the WT system is apparent. The PCA results indicated that while the WT system showed a near-normal distribution of conformational sampling across its 10ns simulation, the HM system was much more disordered and required longer simulation time in order for it to effectively sample its much larger conformational landscape. Therefore the research needed to be directed towards improving this conformational sampling, particularly for the AS, QM and HM systems, in order to hopefully attain a more accurate binding affinity value. Two methods were concurrently pursued to improve the sampling: extension of the simulations to 50ns, and repetition of the simulations, such that they are expanded into an ensemble.



(a)

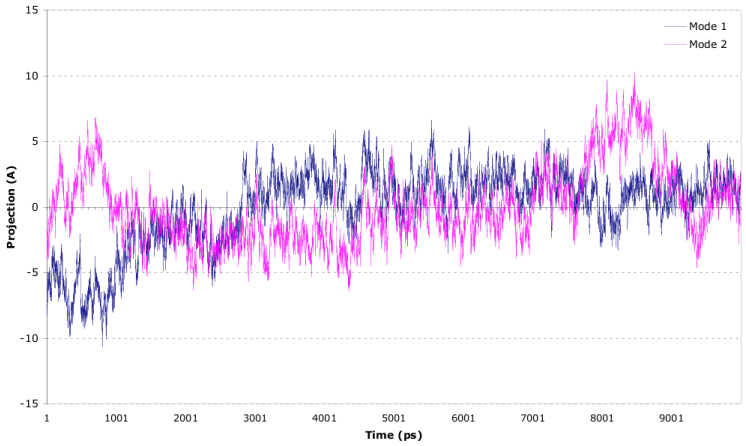


(b)

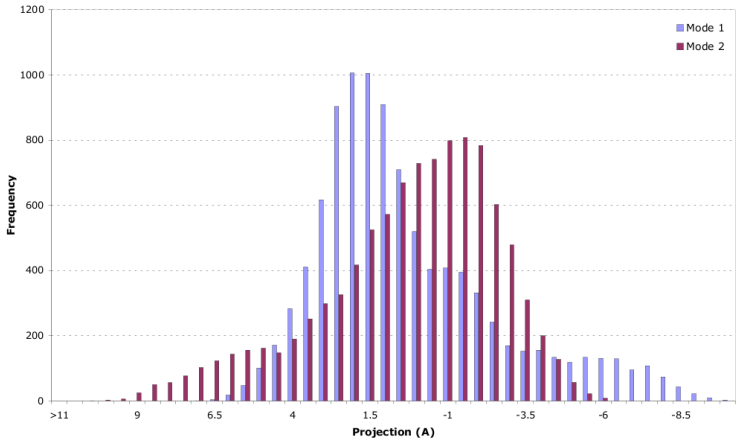


(c)

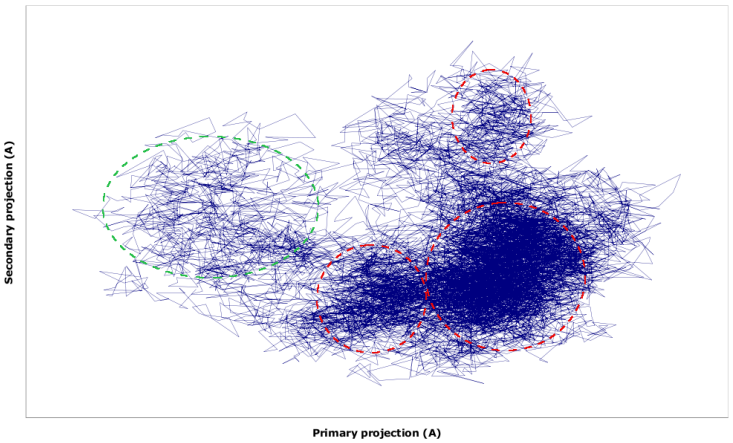
Figure 6.5: PCA on 10ns WT simulation. (a) Evolution of first 2 modes' projections across time. (b) Frequency distribution of the first 2 modes' projections shown in (a). (c) The first 2 modes' projections plotted against each other, with the blue line directing its configurational sampling through the simulation. Red dashed lines indicate sampled conformations.



(a)



(b)



(c)

Figure 6.6: PCA on 10ns HM simulation. (a) Evolution of first 2 modes' projections across time. (b) Frequency distribution of the first 2 modes' projections shown in (a). (c) The first 2 modes' projections plotted against each other, with the blue line directing its configurational sampling through the simulation. Red dashed lines indicate sampled conformations, and the green dashed line a poorly sampled conformation.

6.3 Extended single-trajectory simulations

The results in Section 6.2 indicated that the 10 nanosecond simulations were still not sufficiently sampling configurations, which may have resulted in the anti-correlation to experimental results published by Ohtaka *et al.*. Therefore the WT and HM systems were extended to 50ns and subsequently analysed to determine whether the sampling and calculated binding affinity to saquinavir improve. The reason for choosing these two systems was because they represent the two extreme ends of the drug-resistance trend, and therefore they have the greatest difference in binding affinity. If the relationship between these two systems cannot be reversed upon simulation extension then there is no need to extend the other simulations because the overall trend still has not been altered.

In order to accomplish 100 nanoseconds of MD simulations, the Lonestar machine at the Texas Advanced Compute Center (TACC) was utilised. This machine is able to perform the required NAMD force calculations at a rate of approximately 7 hours/simulated-ns for a 50,000 atom system when parallelised across 32 of its processors. Therefore, as both simulations were able to be run concurrently, the extension of 80 nanoseconds took 11.67 days and required 17,920 computer-hours. In comparison, the two 22 nanosecond simulations performed by Perryman *et al.* (2004) took 314.29 days to run on two computers, each running on a single 2-GHz Xeon processor [77]. This clearly shows the importance of Grid-computing at the 'peta-scale' range in order to attain a turnaround of results to allow such studies to become tractable [95].

Structural analysis of the simulations through RMSD show slight yet distinct dynamic differences between WT and HM (Figure 6.7). While both proteases rapidly adopt a conformation ~ 1.0Å away from the initial structure, the long-timescale dynamics show motions hitherto unseen in the 10 nanosecond simulations. The WT protease displays a very gradual drift (0.006Å/ns) away from this adopted configuration to another ~ 1.3Å away from the initial structure, with little oscillation except for the high-frequency thermal motions of the protease. Conversely, the HM protease displays a low-frequency oscillation with a peak-to-peak amplitude of 0.2Å, peaking at ~ 1.0Å and ~ 1.2Å, and

a period of approximately 5 nanoseconds. In order to determine whether HM was oscillating between two major configurations across this simulation, PCA was performed on the simulation's trajectory, and the first two principle components plotted against each other (see Figure 6.8(a)). The results show that for the trajectory between 22ns and 36ns, where two full oscillations of the global RMSD occur, the HM system samples between two ends of a constrained region of the configurational landscape, and that its sampling of this region follows a normal distribution (6.8(b)). Therefore the structures of both WT and HM can be considered structurally equilibrated across the 50ns.

Having ascertained that both WT and HM structures were stable across their respective simulations, the binding affinity values were calculated. Before comparing the computed values to the experimental data, the evolution of the component terms that comprise the binding affinity values across the simulations needed to be determined to ensure the systems are energetically stable. The results revealed that for both WT and HM there was little difference in both their enthalpies and entropies compared to their respective 10ns simulations. The components of the enthalpic contribution to WT's ΔG value remain fluctuating around the same values (Figure 6.9(a)), and the vibrational entropy retains its variability around the same value through to 50ns (Figure 6.9(b)) whilst the rotational and translational entropies remain constant. The same can be seen in the enthalpic (Figure 6.10(a)) and entropic (Figure 6.10(b)) components of the HM system. While this shows energetic stability across the whole 50ns, it also suggests that the computed ΔG values are not going to change for either system, and therefore will not reverse the anti-correlation.

The MMPBSA and normal mode analysis of the extended systems calculate the ΔG values to be -8.20kcal/mol and -26.17kcal/mol for the WT and HM systems respectively (Table 6.4). Therefore the results of the 50ns simulations show no improvement upon the results across 10ns; in fact, when considered in the context of the results of the 10ns simulations for AS, DM, FL and QM, the correlation actually becomes more strongly anti-correlated, with a correlation coefficient of $r = -0.90$. However, although the HM ΔG becomes more negative, and the WT ΔG becomes less negative, the HM system fell within the error bounds of the original 10ns results, and the WT system

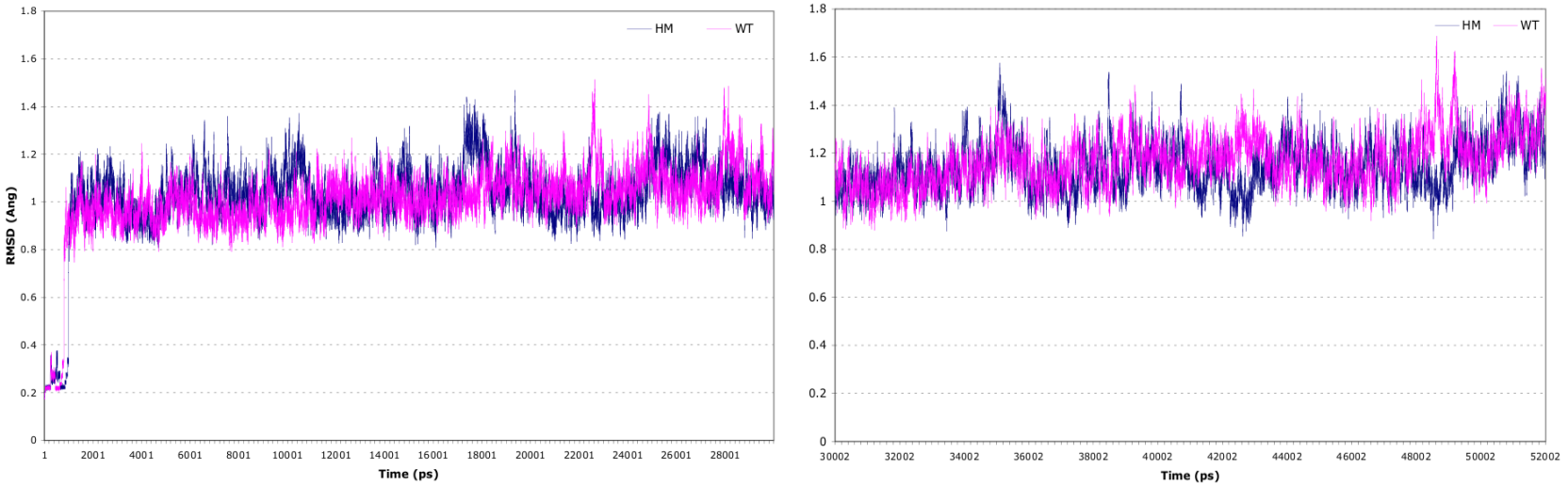
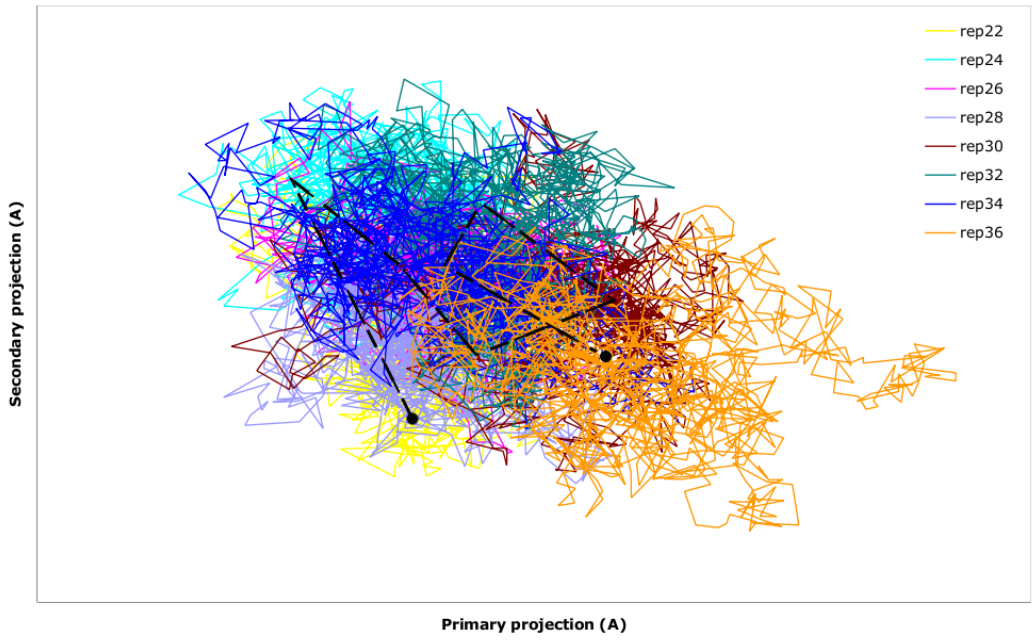
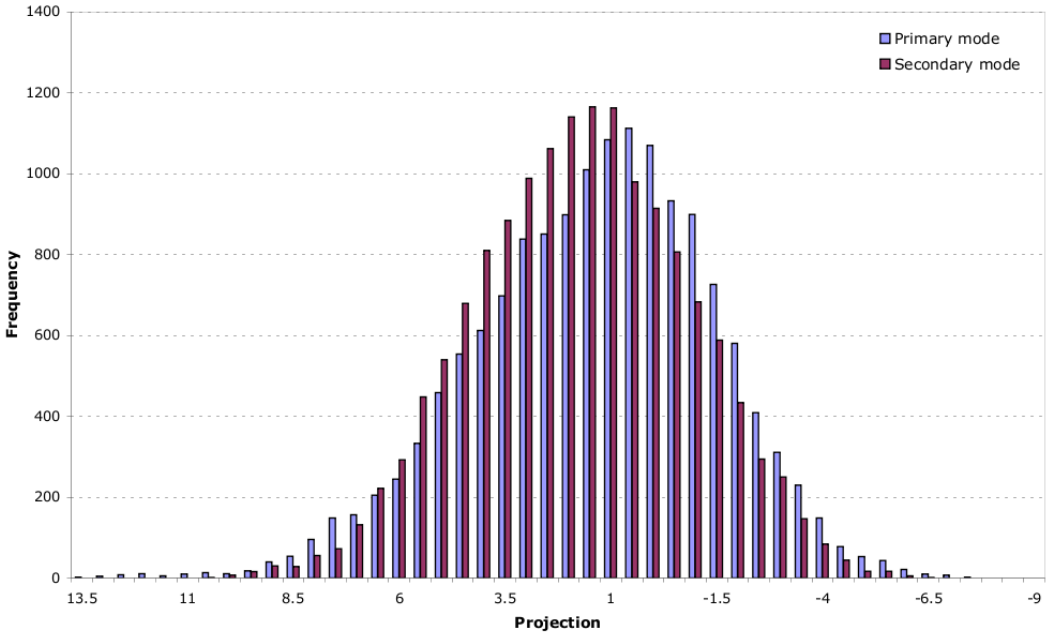


Figure 6.7: Evolution of the global RMSD with respect to the initial starting structure for the HM and WT protease systems complexed with saquinavir. WT is shown in pink, and HM is shown in blue. Each line shows the progression of the protease system across the 52 nanosecond simulation (2ns equilibration followed by 50ns production-phase). The RMSD for HM is shown in dark-blue, while that of WT is shown in pink. The RMSD is split across two graphs due to the maximum number of data points Microsoft Excel will plot on a single graph. The results show that the WT system quickly evolves away from the starting configuration to another 1.0Å away. Over the next 50ns it slowly drift to a configuration $\sim$ 1.3Å away. The HM system shows low-frequency oscillation between configurations 1.0Å away and 1.2Å away.

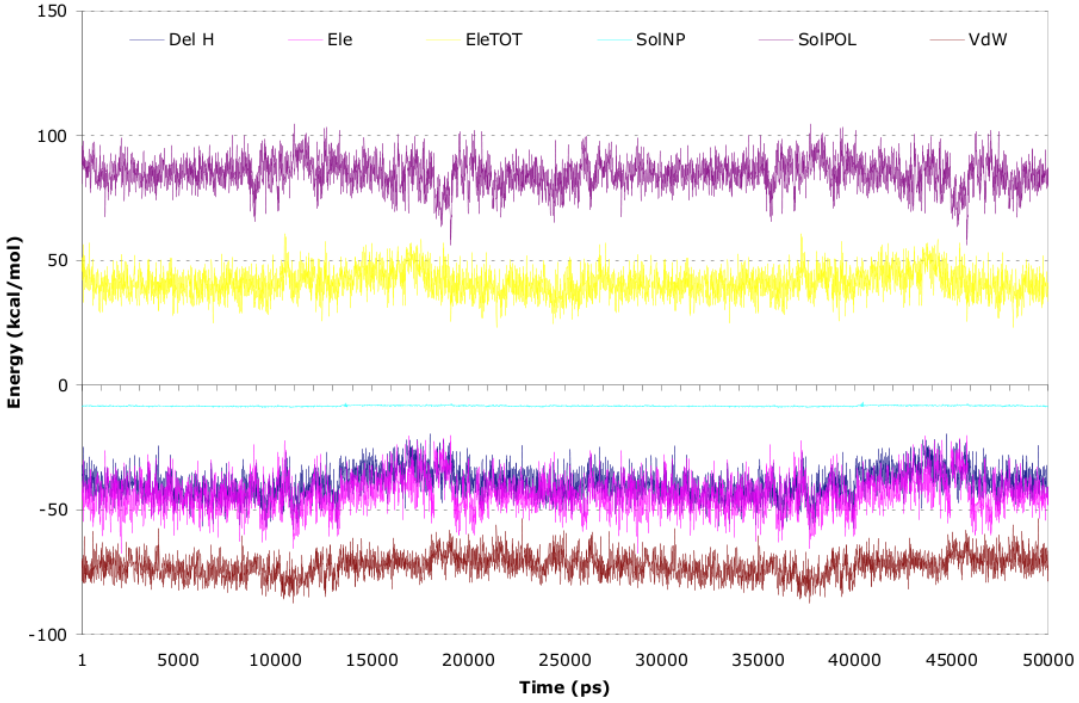


(a)

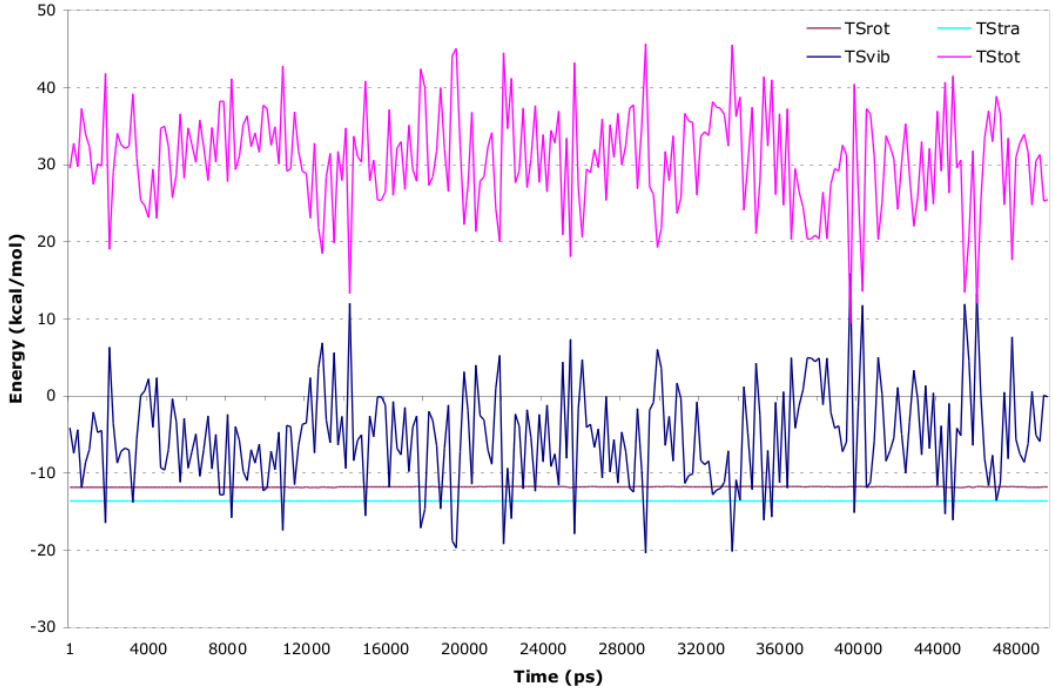


(b)

Figure 6.8: PCA of HM's trajectory between 22ns and 36ns, corresponding to 2 RMSD oscillations in Figure 6.7. (a) The primary and secondary modes' projections plotted against each other. Configurations for each nanosecond are shown in different colours, and a black dotted line depicts the overall direction of configurational sampling. (b) Distribution of frequencies of the primary and secondary modes' projections. The normal distribution indicates stable sampling of configurations.

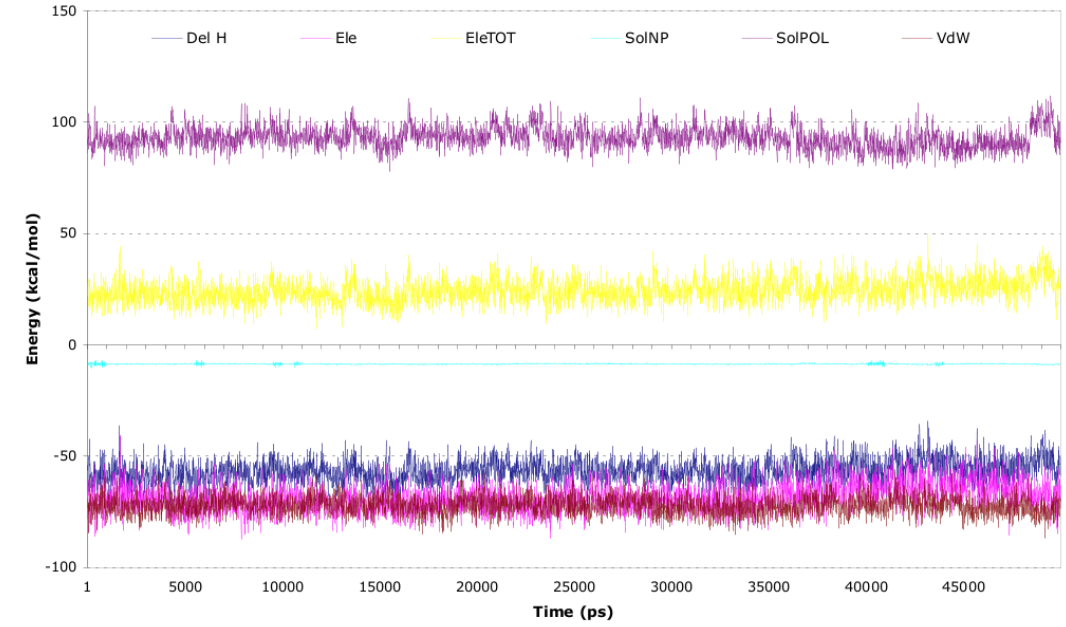


(a)

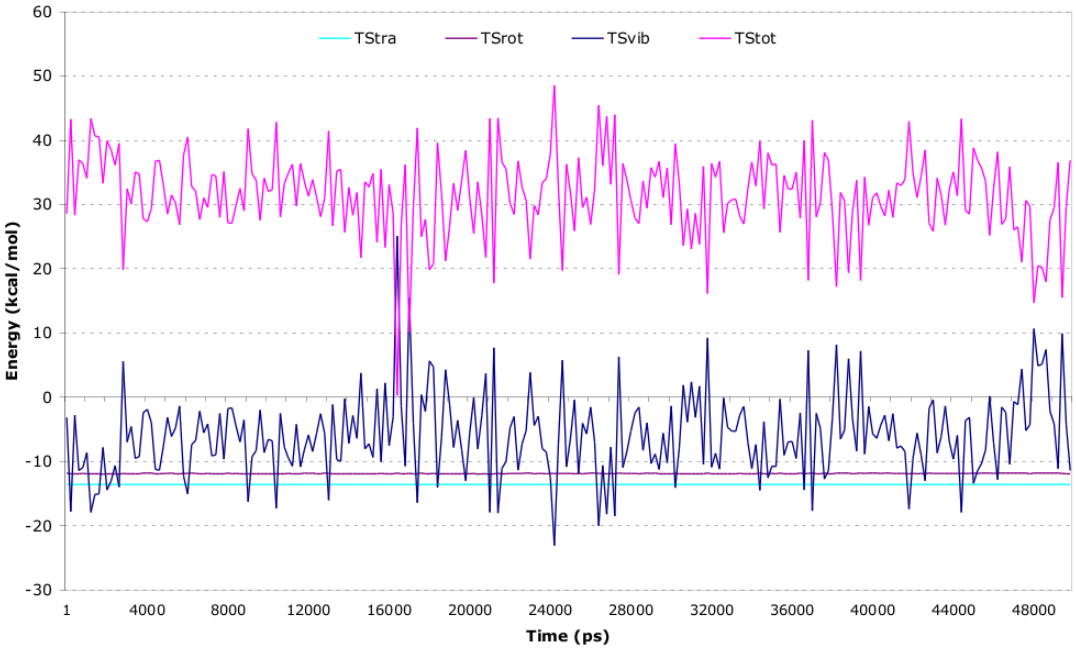


(b)

Figure 6.9: Evolution of the energy components that make up the enthalpic (a) and entropic (b) contributions to WT system's free energy change upon saquinavir binding across the 50ns simulation. All components are consistent across the simulation, showing the system was energetically at equilibrium.



(a)



(b)

Figure 6.10: Evolution of the energy components that make up the enthalpic (a) and entropic (b) contributions to HM system's free energy change upon saquinavir binding across the 50ns simulation. All components are consistent across the simulation, showing the system was energetically at equilibrium.

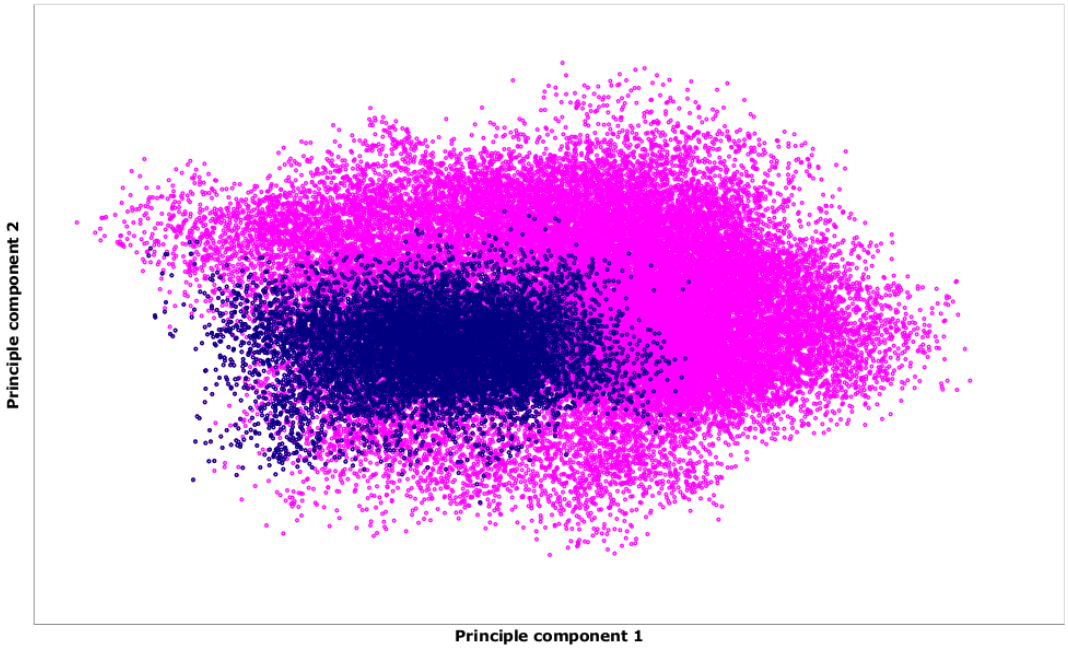
was 0.2kcal/mol outside of the original 10ns results. Therefore it could be argued that the extension to 50ns had no effect on improving the correlation to experimental values for the HM system, and that it worsened the WT correlation. This suggests that the proteases did not undergo the significant improvement on configurational sampling that was hypothesised to improve the accuracy of the computed binding affinity value. In order to determine the extent of the configurational sampling undertaken through the 50ns compared to the 10ns simulation, PCA was performed across the 50ns trajectory, and subsequently the data for the first 10ns compared against the dataset as a whole.

Table 6.4: ΔG values for the 50ns WT and HM simulations

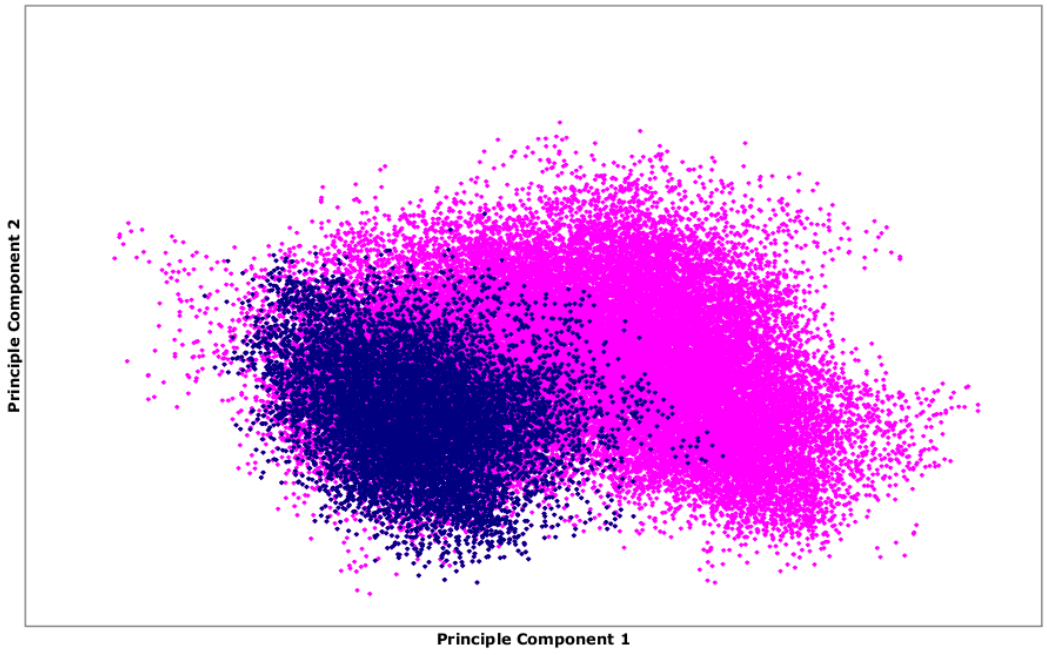
Pr	**Binding affinities (kcal/mol)**						
	Exp	**1-10ns**	**STDEV**[a]	**11-50ns**	**STDEV**[a]	**1-50ns**	**STDEV**[a]
WT	-13.00	-10.16	1.94	-6.24	4.99	-8.20	4.63
HM	-8.50	-24.76	3.12	-27.58	3.90	-26.17	3.73

[a] Standard deviation. Calculated between the ΔG values outputted each nanosecond from the MMPBSA and normal mode analyses.

The comparison between the configurational sampling of the original 10ns simulation and the extended 50ns simulation is shown for the WT system (Figure 6.11(a)) and the HM system (Figure 6.11(b)). The results for the WT system show that between 10 and 50ns, which correspond to the pink regions not covered by the blue regions in Figure 6.11(a), the protease samples an extended range of the motions sampled within 10ns. This is best described by visualisation of these projections on the protease structure. Figure 6.12 shows the superposition of two protease structures that represent the extremes of the correlated motions comprising the WT system's primary eigenvector. So for example, the fulcrum and cantilever regions of the protease show correlated dynamics such that when the fulcrum moves from the pink structure to the grey structure, there is a corresponding movement in the cantilever from the pink structure towards the grey structure. Upon extension of the simulation from 10ns to 50ns, Figure 6.11(a) shows that the maximum extent of correlated motions increases. So the additional structures sampled were those where pink and grey structures deviated by an even greater amount at the ringed regions.



(a)



(b)

Figure 6.11: The projections of the first 2 principle components plotted against each other for the WT system (a) and the HM system (b). In pink are shown the adopted configurations for the 50ns simulation, and in dark-blue the configurations adopted in the first 10ns.

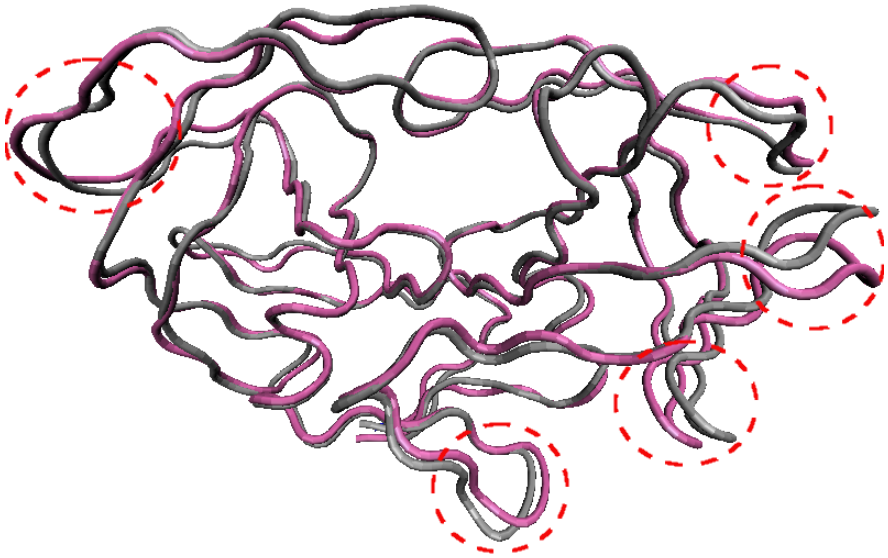


Figure 6.12: Visualisation of the primary eigenvector on the WT protease's quaternary structure. The two superimposed structures represent the minimum and maximum ranges of the WT system's correlated motions over the 50ns simulation that comprise the primary eigenvector. The red dashes highlight the major regions of correlated motions.

As can be seen, the correlated motions corresponding to the primary eigenvector all occur at the extremities of the structure. The structure and volume of the active site are unaffected, and may therefore explain why there is little change in the strength of binding between the protease and saquinavir. Analysis of the second and third eigenvectors (data not shown) also reveals that these correlated motions do not impact the active site.

Analysis of the principle components of the HM system show very similar results. Although the exact locations and extent of the correlated motions on the structure do not concur, the extension of the simulation to 50ns resulted in an increased range of the correlated motions observed within 10ns (Figure 6.11(b)). Sampling of a conformation representing a different potential energy well would be observed as an occupied region with few conformational data points linking it to the other conformational region. This would be because the space connecting the two regions would represent the 'higher energy' region connecting the two wells that are energetically-unfavourable and therefore would not be sampled frequently. As the conformational landscapes in Figures 6.11(a) and 6.11(b) do not show this, it suggests that the structures are confined to a single conformational region which it thoroughly samples. These PCA results, in conjunction with the calculated binding affinity values, show that running the simulations for 50ns does not improve upon the results after 10ns, and does not significantly improve the configurational sampling undertaken by the protease. It is likely therefore that alternative methods must be employed to address this.

6.4 Ensemble simulations

In addition to the research carried out in Section 6.3, improvements on the results of Section 6.2 were investigated by extending the HM and WT systems to ensemble MD simulations. As with the simulation length extension in Section 6.3, only these two systems were considered for replication because they represent the extreme ends of the chain in terms of drug-sensitivty. Therefore, both WT and HM systems were simulated another 9 independent times to attain 10 systems, each 10ns long. The total simulation time for these two ensembles was 216 nanoseconds (HM and WT both run another 9

times for 2 nanoseconds equilibration and 10 nanoseconds production-phase), which is over double the length of time of the simulation extension in Section 6.3. In order to achieve a turnaround of results in a comparable timeframe to the simulation extensions, these simulations were run concurrently on the Ranger machine at the Texas Advanced Compute Center (TACC). This machine has 62,976 processor cores [35], so utilising 576 of these to concurrently run the 18 replicate systems across 32 processors required only a fraction of the total number of processor cores available. Ranger was able to perform the required NAMD force calculations at a rate of approximately 8 hours/simulated-ns for the 50,000 atom protease system when parallelised across 32 of its processors. Therefore the time required to run all 18 systems through 12 nanoseconds was 4 days, and required 55,296 computer-hours. Comparing this to that of Perryman *et al.*, a 4,243 fold increase in turnaround time was achieved by utilising the power of concurrently running simulation across a Grid network [77]. Even in comparison to the simulation extension performed on the Lonestar machine, the ensemble method was able to achieve a 7.88 fold increase in turnaround time. This shows that, without considering the accuracy of the methodology, utilising Grid computers and supercomputer facilities for ensemble MD results in a much faster turnaround of results than long-timescale MD simulations. The development of automated scripts such as the Binding Affinity Calculator (BAC), which is integrated into the Application Hosting Environment (AHE) to automate the task of generating a system of any genotype ready for simulation; farming the simulations to Grid networks and supercomputers; performing equilibration and simulation of the system; bringing the outputted files back; and analysing them through MMPBSA and normal mode analysis [95, 97], makes the aim of 'using molecular dynamics simulations to aid clinicians in their decision of drug therapy to administer' more tractable when the necessary simulations can be run in 4 days with the clinician only needing to know the genotype of the HIV protease.

Structural analysis of the simulations was performed through global RMSD for the WT (Figure 6.13(a)) and HM (Figure 6.13(b)) ensembles. The results showed that all repetitions for both ensembles were structurally stable across their simulation; their high-frequency and low-frequency fluctuations show significant overlap. RMSF analysis of the repetitions with respect to the time-averaged structure also showed that the

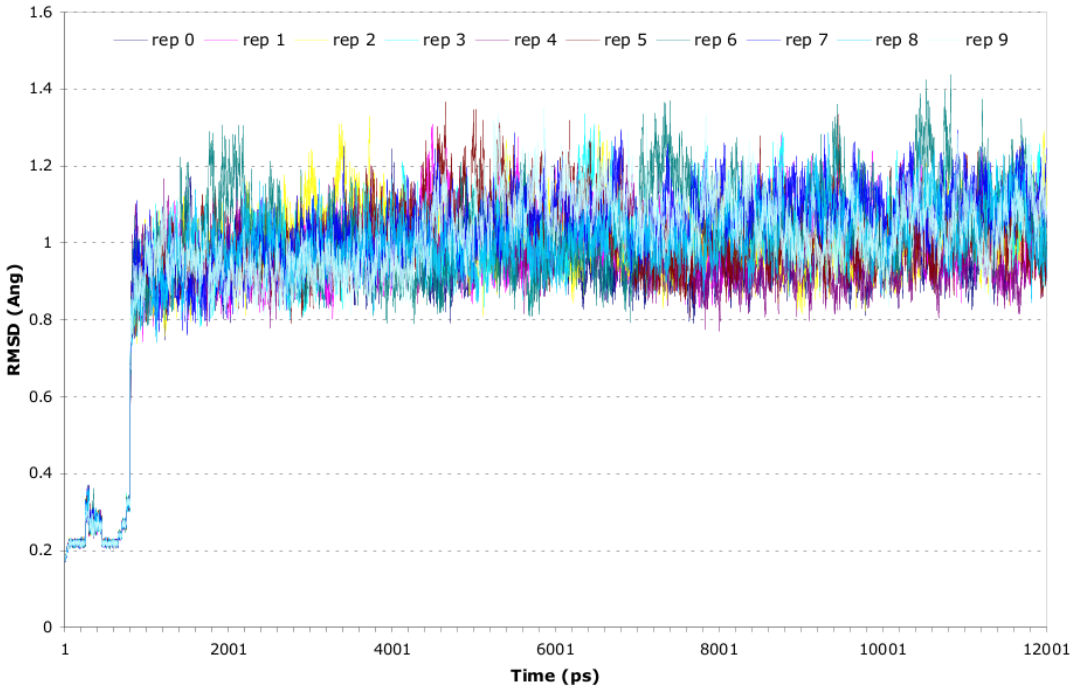
structures fluctuate stably around their average structure across the simulation (data not shown). Having determined the structural stability of the 20 simulations, the binding affinity values for each system could be calculated and the result averaged over the 10 repetitions.

Table 6.5: Comparison of ΔH, TΔS and ΔG values for the WT and HM 10-repetition 10ns ensembles

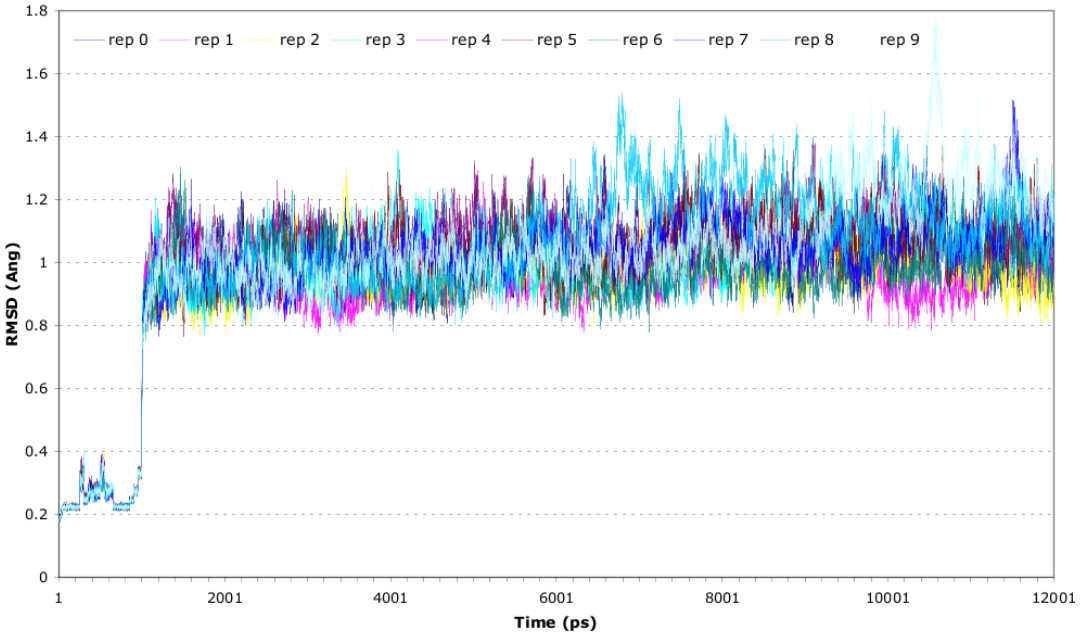
Repetition	**WT (kcal/mol)**			**HM (kcal/mol)**		
	ΔH	TΔS	ΔG	ΔH	TΔS	ΔG
0[a]	-41.90	-31.74	-10.16	-57.99	-33.23	-24.76
1	-43.46	-28.47	-14.99	-47.74	-30.35	-17.38
2	-45.55	-30.84	-14.71	-46.15	-29.53	-16.62
3	-43.85	-30.14	-13.71	-40.55	-29.93	-10.62
4	-39.10	-30.98	-8.12	-53.98	-29.59	-24.39
5	-54.39	-31.26	-23.12	-39.40	-31.95	-7.45
6	-50.84	-30.60	-20.24	-47.61	-28.14	-19.47
7	-44.55	-29.69	-14.86	-44.17	-30.78	-13.39
8	-43.19	-30.85	-12.34	-41.14	-29.62	-11.53
9	-39.70	-32.53	-7.17	-42.64	-31.85	-10.79
MEAN	-44.65	-30.71	**-13.94**	-46.14	-30.50	**-15.64**
STDEV	4.73	1.11	4.98	6.00	1.49	5.90

[a] Repetition 0 refers to the original 10ns simulation run in Section 6.3.

The 'MEAN' row in Table 6.5 gives the mean ΔG for the WT ensemble as -13.94 ±4.98 kcal/mol, and for the HM ensemble as -15.64 ±5.90 kcal/mol. This still makes the HM protease system more attracted to saquinavir than the WT system, but this is a marked improvement on both the results of the single 10ns simulation in Section 6.2 and the extended 50ns simulation in Section 6.3. A comparison of the ΔG values by each method is given in Table 6.6. This shows that while all three simulation methods gave WT results comparable to the experimental, the ensemble average gave the best result, with only a 0.94kcal/mol deviation between the experimental results and the computed ensemble mean. Furthermore, while the HM results were still more negative than the WT results, the HM ensemble average was only 7.14kcal/mol more negative than experimental compared to the 19.08kcal/mol difference for the 50ns simulation. The error bounds of the HM and WT ensembles overlap (Table 6.5) so it is possible that the ensemble methodology has reversed the correlation to a positive one, but that



(a)



(b)

Figure 6.13: Evolution of WT (a) and HM (b) ensembles' global RMSDs with respect to their post-mutation pre-simulation structure. Each ensemble contains 10 repetitions, and run for 2ns equilibration followed by 10ns data-collection. The data shows that all the repetitions within an ensemble show stability across the simulation, and show strong correlation to other repetitions in the ensemble. All repetitions therefore show structural equilibration by 2ns.

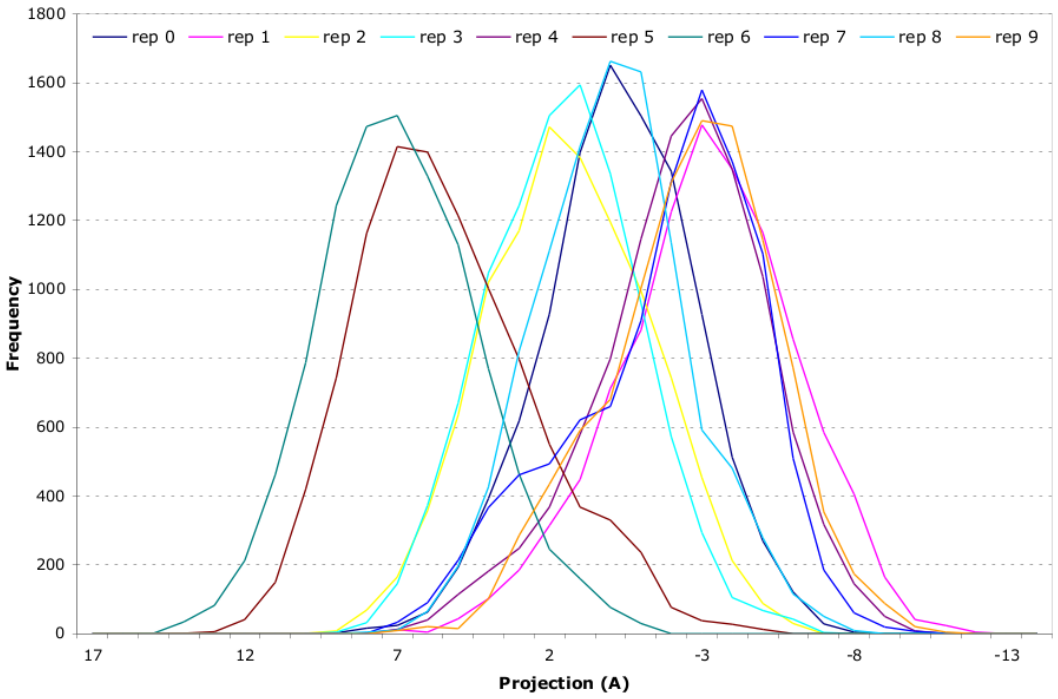
more repetitions are required to make the distinctions more apparent.

Table 6.6: Comparison of the outputted ΔG value by the 3 methods researched in this chapter

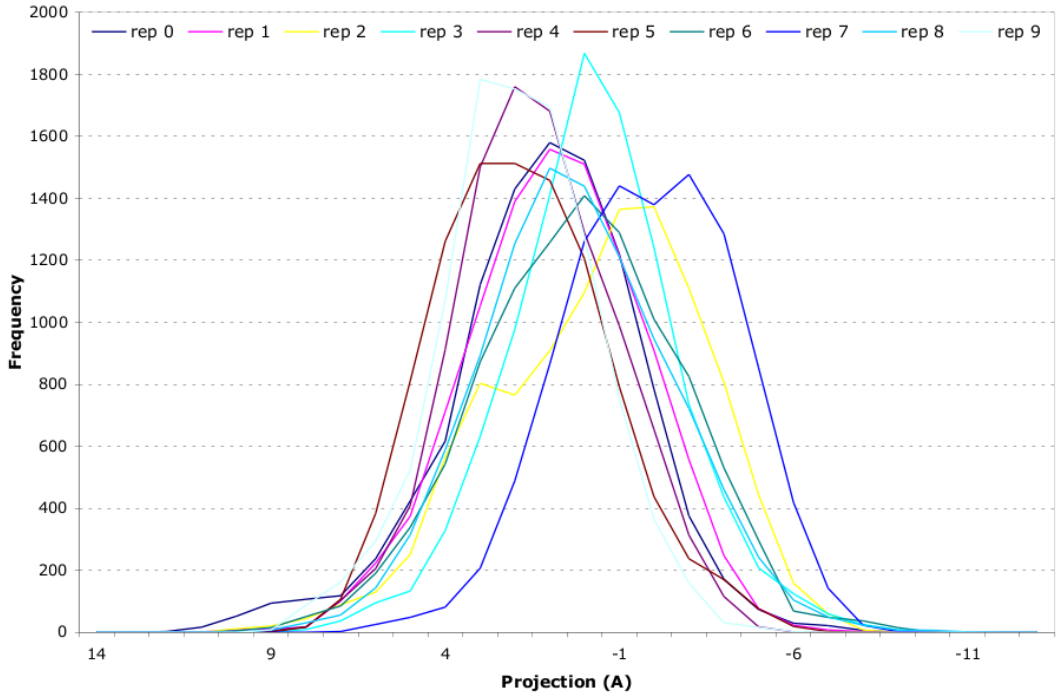
System	**ΔG (kcal/mol)**			
	Experimental	**10ns single**	**50ns single**	**10-rep ensemble**
WT	-13.00	-10.16	-8.20	-13.94
HM	-8.50	-24.76	-27.58	-15.64

The results also show that both the WT and HM ensembles show large ranges in the ΔG values between repetitions, with WT showing a 15.95kcal/mol range between simulations, and HM showing a 17.31kcal/mol range. These ranges are larger than the average ΔG values, which highlights the caution of relying on a single simulation to generate a result; had WT repetition 7 been run in conjunction with HM repetition 5, the computational results would have concurred with experimental, and the conclusion would have been made that the methodology was sufficient for reproducing the trend in saquinavir resistance. It is also important to note that the variation in ΔG values between simulations was far greater than any variation previously observed across a simulation, which indicates that a much greater chance of reaching novel conformations is achieved through the random assignment of velocities at the first MD time-step combined with the minimisation protocol than is achieved through natural conformational sampling in unrestrained simulation over longer time periods. The conformational sampling of the ensemble methodology was investigated through PCA of the WT and HM ensembles.

The frequency distribution of the primary eigenvector (Figure 6.14(a)) was encouraging because it showed that all repetitions followed a normal-distribution. This means that within each repetition the protease was stably sampling a particular conformation over the length of its simulation, which gives further indication that the structures were equilibrated. It also reinforces the observations made in Section 6.3 that once the structure has minimised to a particular energy minima on the conformational landscape, it

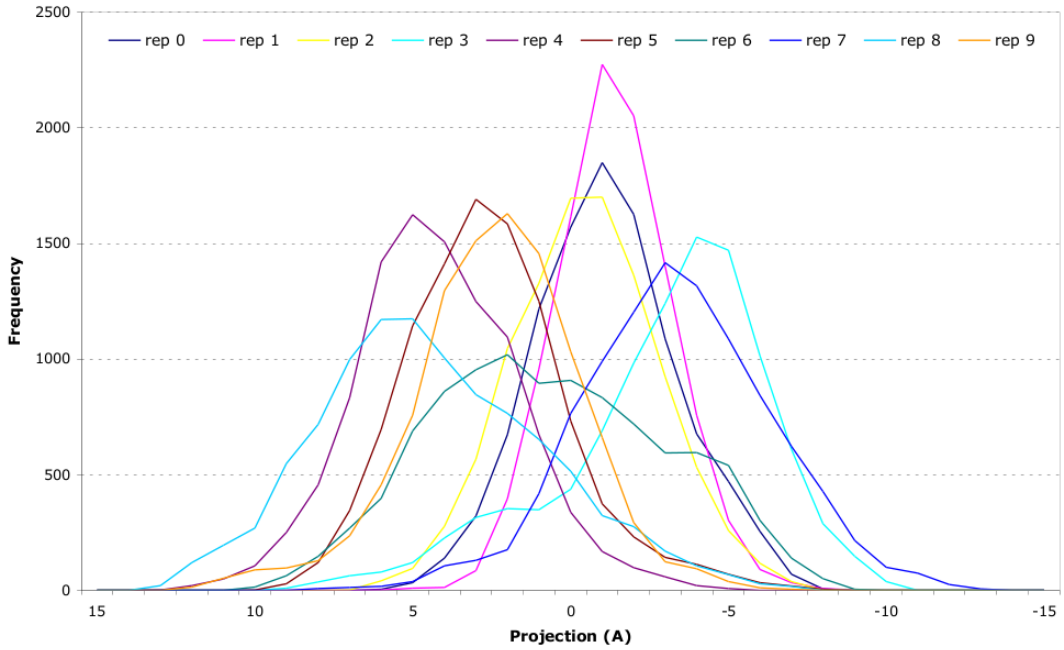


(a)

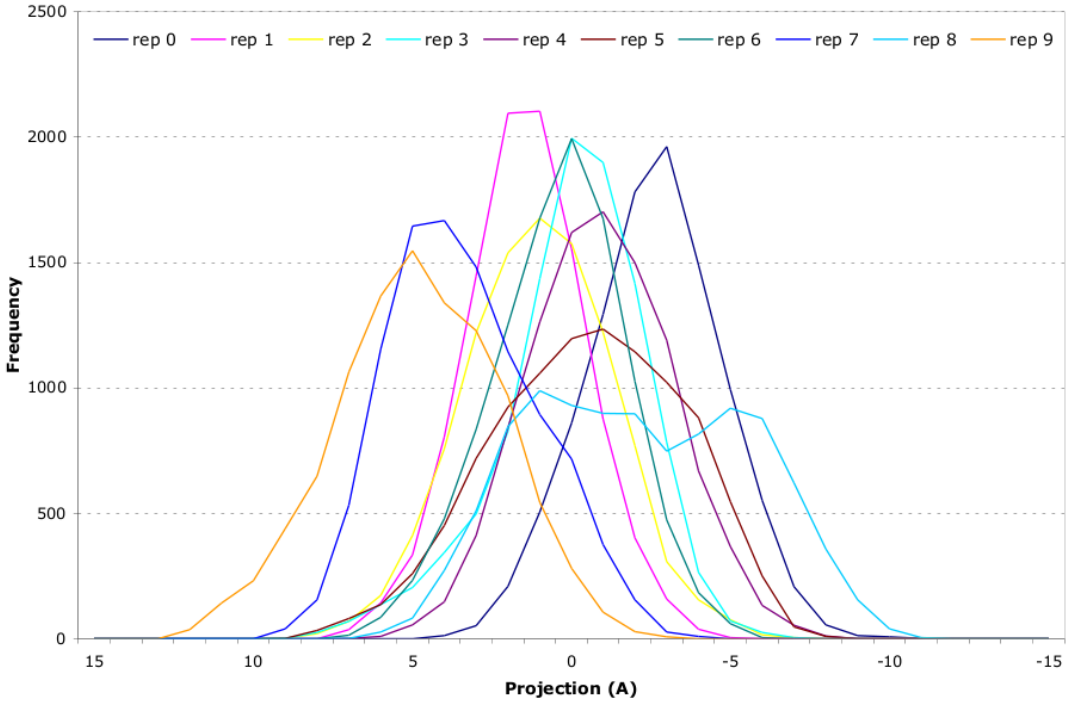


(b)

Figure 6.14: PCA of the 10-repetition WT ensemble. (a) Frequency distribution of the primary mode's projections for each of the 10 repetitions. Results show a high degree of overlap in sampling, and all show normal distribution in their projections. (b) Frequency distribution of the secondary mode's projections for each of the 10 repetitions. Results show almost complete overlap in the sampling of this mode, which means there is little difference in these correlated motions between repetitions.



(a)



(b)

Figure 6.15: PCA of the 10-repetition HM ensemble. (a) Freqency distribution of the primary mode's projections for each of the 10 repetitions. The results show reasonable overlap between repetitions, but with a greater overall range of configurations than WT. (b) Frequency distribution of the secondary mode's projections for each of the 10 repetitions. The results show a normal distribution of sampling for all repetitions, and significant overlap between repetitions.

does not leave this minima over the course of its simulation. While there was significant overlap between the projection of the primary eigenvector for different repetitions, they fell into distinct regional distributions suggesting they sampled different regions of the same mode. Correlation of the modal projection value for each repetition to its corresponding ΔH value gives a correlation coefficient of $r = -0.89$, which is a significant anti-correlation. Therefore, if you consider the entropy as a constant value for the system, the calculated binding affinity value gets more negative as the primary mode's projection gets larger. This means that the correlated motions which comprise the primary mode directly affect the strength of binding between the protease and inhibitor, and subsequently the regions of these motions that are sampled across a simulation determine its accuracy to experimental. To determine whether this is unique to the motions of the primary eigenvector, the correlation between the modal values of the secondary eigenvector (Figure 6.14(b))to the ΔH value was calculated. The resultant correlation coefficient was $r = 0.29$, which was not significant.

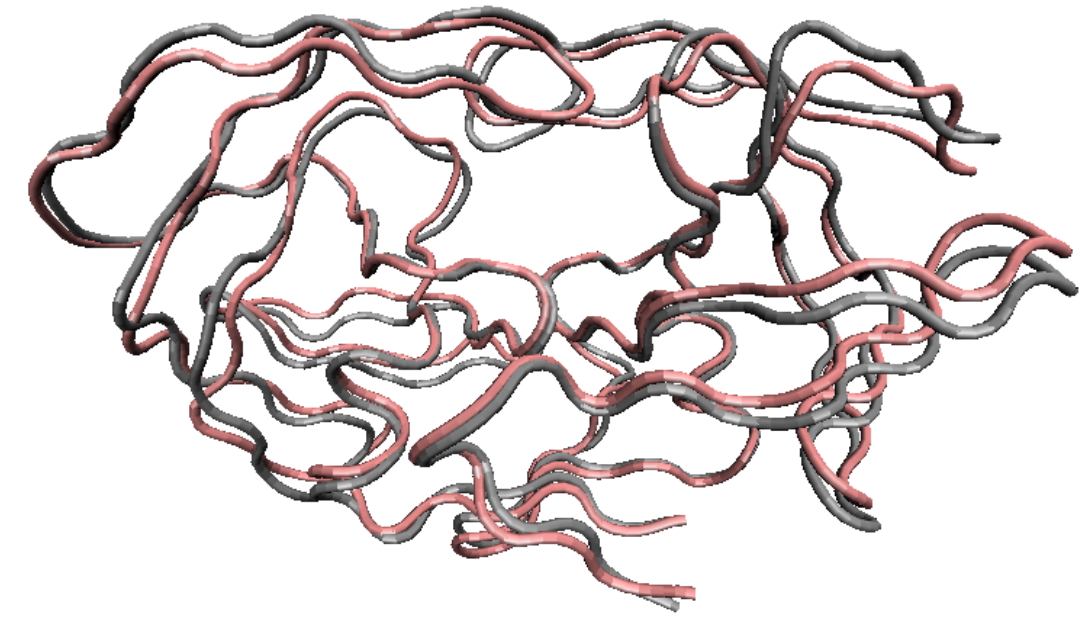
The frequency distribution of the projections of the primary eigenvector for the HM ensemble were calculated to determine their correlation to the ΔH value (Figure 6.15(a)). The resultant correlation coefficient was $r = -0.05$, which represents no correlation between the two sets of values. Therefore the strong correlation between the correlated motions of the WT system and the binding affinity value do not follow through to the HM system. It is important to note that the motions comprising the primary eigenvector of the WT system are not necessarily the same as the motions comprising HM's primary eigenvector, so these results do not indicate a disassociation between these motions and the ability to determine drug sensitivity. The correlation between the modal values of the frequency distribution of the second eigenvector's projections (Figure 6.15(b)) and the ΔH value was calculated to be $r = 0.52$. Values between 0.40 and 0.69 indicate a modest correlation [20], so therefore although the HM system does not show correlation to the most significant correlated motions, it does show moderate correlation to the second-most significant motions. Further analysis showed that there was no significant correlation to subsequent eigenvectors (data not shown).

Having determined that there is very strong correlation between WT's primary eigen-

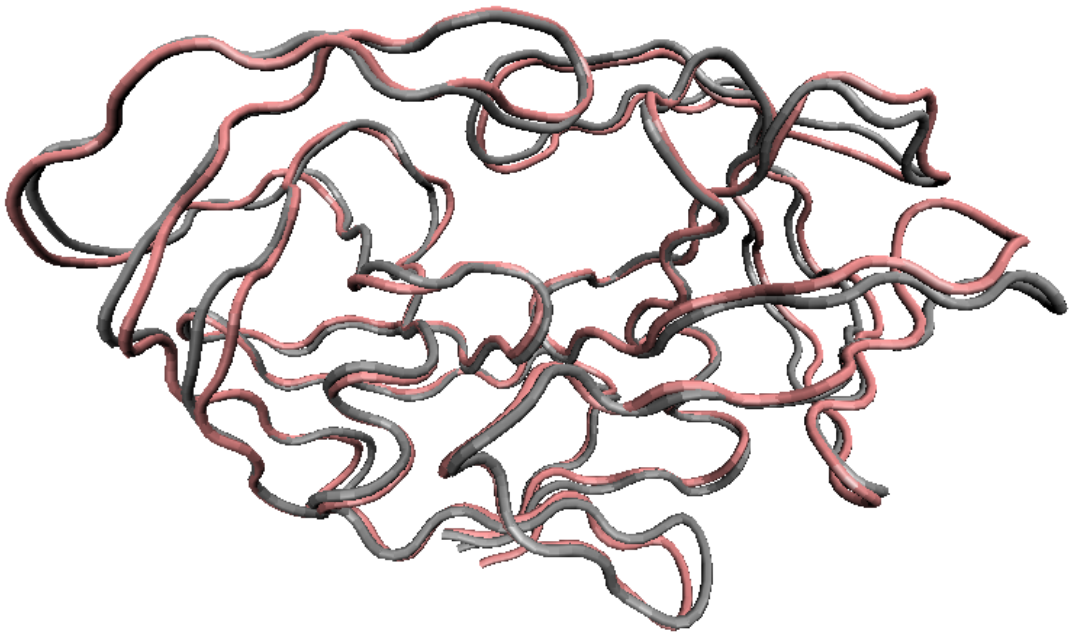
vector's correlation motions and the binding affinity value, and moderate correlation between HM's secondary eigenvector's correlated motion and the binding affinity value, the location and magnitude of these motions needed to be visualised to see if there was any similarity between the two. The results showed that the correlated motions of WT's primary eigenvector (Figure 6.16(a)) and HM's secondary eigenvector (Figure 6.16(b)) were very similar. The correlated motions were concentrated around the extremities of the structure, particularly the fulcrum and cantilever regions of the right-hand monomers in the figures. The active site remains relatively static except for the flaps and loop regions that enclose the top. In contrast, the correlated motions of HM's primary eigenvector, which resulted in no correlation to calculated ΔH, showed the structure stretching and contracting across its horizontal axis (data not shown). The similar location of the correlated motions was encouraging, as it narrowed the cause of the differences between replicate simulations, and between WT and HM systems, down to a single eigenvector. However, as the results have shown that the ΔG results across a simulation do not significantly change, and that the adopted configuration is much more dependant on the initial minimisation and equilibration protocol rather than the sampled configurations during the simulation, only running 10 replicates for each system is not sufficient for the protease to sample all the possible configurations in the eigenvector that correlates with the ΔH.

6.5 Extended ensemble simulations

Following these encouraging results, the HM ensemble was extended to 100 repetitions and the WT ensemble extended to 20 repetitions. The reason for extending the HM ensemble further than the WT was because the 10-repetition WT ensemble already gave results that concurred with experimental data whereas the 10-repetition HM ensemble showed improvement upon the single-trajectory simulation but still did not completely concur. Therefore, due to the difficulty of analysing such a large amount of data, both in terms of technical limitations of the software used, and the time required to analyse 100 repetitions, this was only performed on the HM ensemble. However, the WT ensemble was extended to 20 repetitions to ensure that the results still agreed with experimental.



(a)



(b)

Figure 6.16: Location and magnitude of the correlated motions comprising WT's primary eigenvector (a) and HM's secondary eigenvector (b). PCA performed for each system across their 10-repetition 10ns ensemble. Results show strong similarities between the proteases in the location of the correlated motions.

Due to the amount of computational resource made available at this time, running the ensemble simulations was no longer the rate-limitation in turnaround of results. Instead file transfer from the supercomputers back to the central storage, and performing the MMPBSA and normal mode analysis became more time-prohibitive than running the simulations. In order to extend the HM ensemble to 100 repetitions and the WT ensemble to 20 repetitions, 100 extra 12 nanosecond (2ns equilibration and 10ns data collection) simulations were run. These were run concurrently on the Lonestar and Ranger machines at TACC, the Leibniz-Rechenzentrum (LRZ) at München, and the High End Computing Terascale Resources (HECToR) supercomputer of UK National Supercomputing Service. Collectively, these resources contain 82,120 processor cores [35, 53, 54, 3], so it was possible to comfortably run all 100 simulations simultaneously across 3,200 processors. All simulations were therefore completed in 4 days. This generated 600Gb of data which then needed to be transferred back from the various supercomputers around the world to the centralised location where they were analysed. This transfer was non-trivial, and required the development of in-house scripts to help speed up the transfer-rate. The subsequent normal mode analysis is not able to be parallelised, and so because the method performs its own minimisation of the structure prior to configurational energy calculation, this took approximately 1.5 days to analyse 1 nanosecond.

Global RMSD analysis of the structures showed that all simulations for both WT and HM rapidly relax to a structure about which they stably fluctuate for the course of the simulation (data not shown). Due to the difficulty in presenting such large amounts of data, only select simulations, and the statistics of the ensemble as a whole will be presented. The average of the extended ensembles, and their comparisons to the original ensembles and experimental data, are shown in Table 6.7. The average ΔG of 20-repetition WT ensemble retains its concordance to Ohtaka *et al.*'s experimental value, with both the component enthalpies and configurational entropies retaining a stable average as the number of repetitions increased. However, the HM ensemble actually became more negative, moving back in the direction of the single trajectory results (see Section 6.2). The cause of this increased negativity was the enthalpic component (ΔH) of the free energy, which became 1.09kcal/mol more negative when averaged

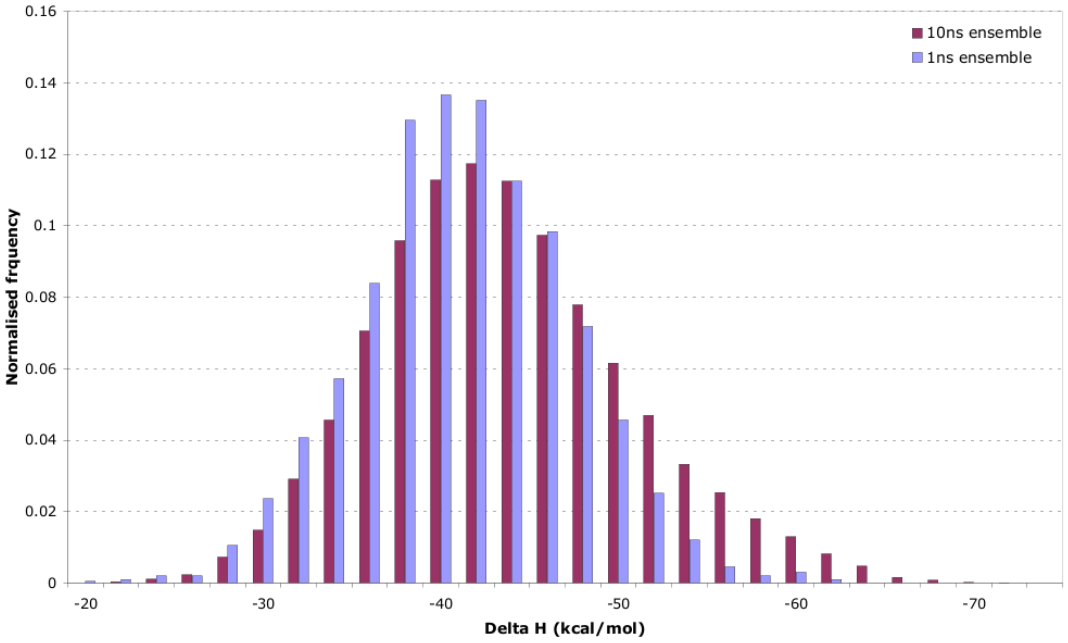
over 100 repetitions instead of the original 10. Conversely, the entropic component remained stable, which indicates that the cause of the increased negativity in the HM binding affinity value is stronger non-bonded forces between the protease and inhibitor averaged across repetitions 10-99 compared to 0-9. Decomposition of the ΔH and $T\Delta S$ into their component energies showed that the distribution of the ΔH_{ele} values becomes more negative across 100 repetitions than across 10, whilst the other components remain stable (data not shown). Therefore the non-bonded electrostatic forces caused the ΔG 100-repetition average to become more negative in the HM system.

Table 6.7: Thermodynamics of the extended WT and HM ensembles

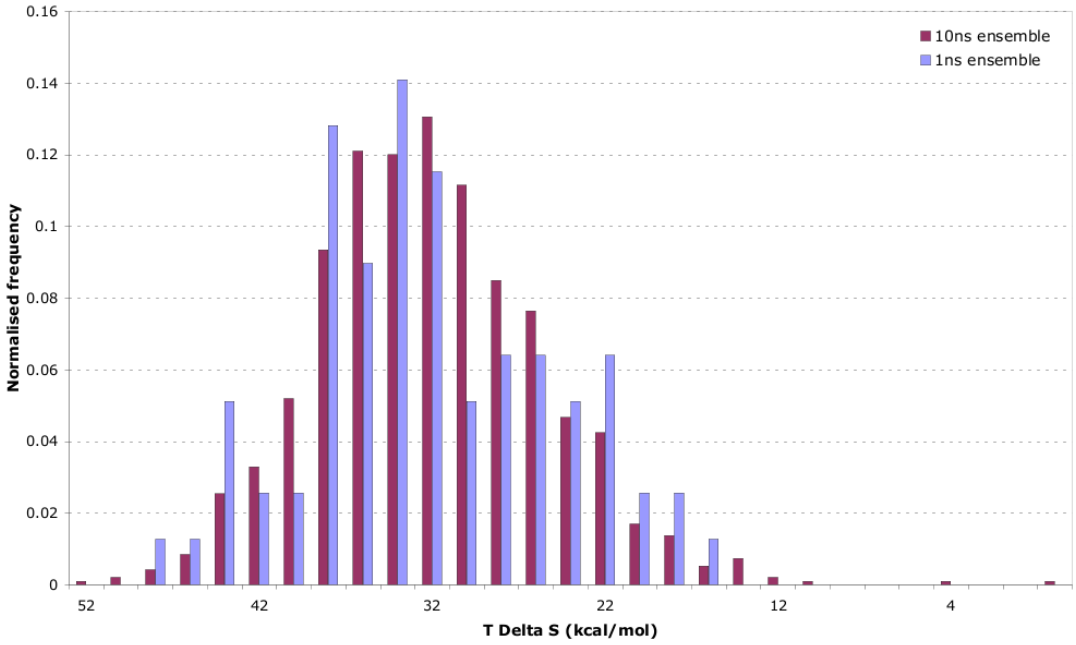
Pr	**Thermodynamics (kcal/mol)**								
	Rep 0-9			**Rep 10-99**			**Rep 0-99**		
	ΔH	$T\Delta S$	ΔG	ΔH	$T\Delta S$	ΔG	ΔH	$T\Delta S$	ΔG
WT	-44.65	-30.71	-13.94	-45.02	-31.67	-13.36	-44.85	-31.21	-13.63
HM	-46.14	-30.50	-15.64	-47.37	-30.68	-17.54	-47.23	-30.64	-17.11

The range of ΔG and ΔH values does not change upon extension of the WT ensemble (Tables 6.8 and 6.9), indicating that the range of configurations required to calculate the binding affinity value is captured within the first ten repetitions. However, this is not the case for the HM system, where the standard deviations and increased range of ΔG and ΔH values indicates that the protease has many more configurations that it needs to sample, and which cannot be captured within 10 repetitions. The extended ensemble for HM also highlighted the benefit of running more repetitions in an ensemble over extending a simulation for sampling configurations. The average standard deviation of ΔH values across a 10ns simulation was 2.96kcal/mol, with some simulations giving standard deviations as low as 1.12kcal/mol. In contrast, the average standard deviation of ΔH values across the first nanosecond of all 100 simulations was 7.23 kcal/mol. In fact, taking the enthalpic and entropic values from the first nanosecond of each simulation in the ensembles gives ΔG values of -13.18kcal/mol for WT and -16.09kcal/mol for HM. These differ by 0.45kcal/mol for WT and 1.02kcal/mol for HM, which fall within the error bounds of the 10ns-ensemble. This is further emphasised by comparing the distribution of ΔH values for both WT and HM systems across 1ns

ensembles and 10ns ensembles (Figures 6.17(a) and 6.18(a)). The comparison shows that the ΔH values are normally-distributed within 1ns for both WT and HM, and that extending each repetition to 10ns does not significantly change the magnitude or the distribution. In the HM system, the ΔH distribution (Figure 6.18(a)) does show a slight over-representation of values < -54.0kcal/mol which would contribute to the 1.02kcal/mol more negative ΔG mean. In comparing the TΔS distributions for the 1ns and 10ns ensembles of both WT and HM (Figures 6.17(b) and 6.18(b)) the results show that the 1ns ensemble does not follow the normal-distribution adopted by the full 10ns ensemble. However, this does not necessarily mean that 1ns is not a sufficient length of time to calculate a configurational entropy value; it may indicate that an insufficient number of snapshots were used to calculate the entropy. The analysis protocol performed by the BAC averages the entropy from 1 snapshot every 200ps, which means that only 5 values are calculated for a 1ns simulation. The results of Table 6.10 showed that the configurational entropy remains almost a constant value across simulations, and the average TΔS value across the 1ns ensembles only differed by 0.38kcal/mol for WT and 0.76kcal/mol for HM. Therefore, by calculating the configurational entropy from more snapshots per nanosecond, a normal distribution similar to the 10ns ensemble may be attained. Although an increased number of snapshots per nanosecond was not implemented, the minimum required number of nanoseconds to attain the normal distribution observed at the 10ns ensemble was investigated (Figure 6.19). The results showed that by 5ns the distribution was almost exactly the same as at 10ns. To calculate the configurational entropy over 5ns utilises 25 snapshots, so this suggests that either the binding affinities can be averaged over 5ns to attain a normal distribution of entropies, or they can be averaged over just the first nanosecond where the configurational entropy is averaged over 25 snapshots, which corresponds to 1 every 40ps. This would allow the binding affinity of an ensemble to be calculated from simulations with 2ns equilibration and 1ns data-collection; ensuring that the structures simulated over that time sample configurations resulting in a normal distribution of enthalpies and entropies.

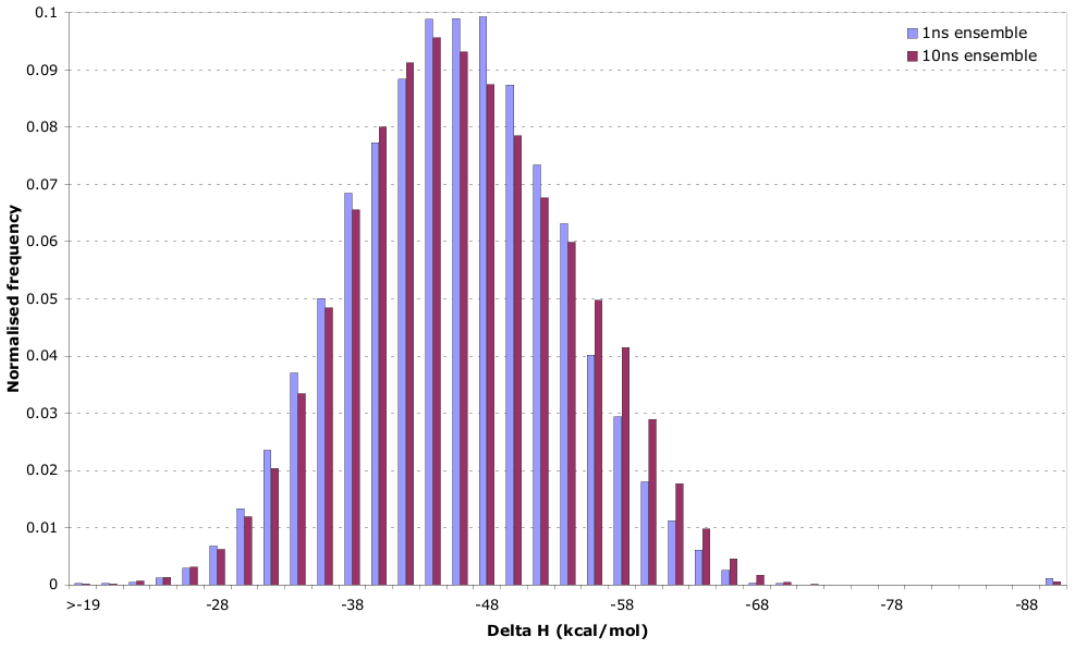


(a)

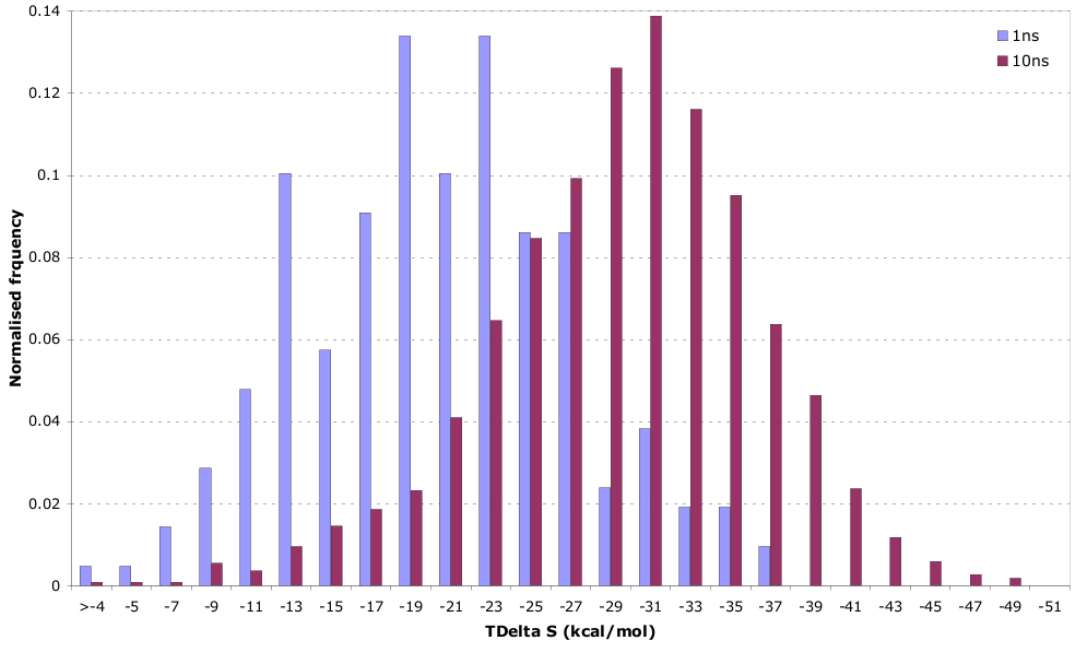


(b)

Figure 6.17: Comparison of frequency distributions of the ΔH component (a) and TΔS component (b) of the binding free energy across the WT ensemble. The blue bars represent the consideration of only the first nanosecond of production-phase for each repetition; the red bars consider the whole 10ns simulation. The graphs were normalised by dividing the frequency by the total number of values to allow direct comparison between 1ns and 10ns.



(a)



(b)

Figure 6.18: Comparison of frequency distributions of the ΔH component (a) and TΔS component (b) of the binding free energy across the HM ensemble. The blue bars represent the consideration of only the first nanosecond of production-phase for each repetition; the red bars consider the whole 10ns simulation. The graphs were normalised by dividing the frequency by the total number of values to allow direct comparison between 1ns and 10ns.

Table 6.8: ΔG statistics of the extended WT and HM ensembles

Pr ensemble		**ΔG statistics (kcal/mol)**			
		Lower range	**Upper range**	**Mean**	**STDEV**
WT	0-9	-20.24	-7.17	-13.94	6.10
	10-19	-19.79	-8.49	-13.36	6.49
	0-19	-20.24	-7.17	-13.63	6.30
HM	0-9	-24.76	-7.56	-15.64	6.90
	10-99	-27.20	-5.59	-17.54	7.74
	0-99	-27.20	-5.59	-17.11	7.60

Table 6.9: ΔH statistics of the extended WT and HM ensembles

Pr ensemble		**ΔH statistics (kcal/mol)**			
		Lower range	**Upper range**	**Mean**	**STDEV**
WT	0-9	-54.39	-39.10	-44.65	5.14
	10-19	-50.20	-39.93	-45.02	5.42
	0-19	-54.39	-39.10	-44.85	5.28
HM	0-9	-57.99	-39.40	-46.14	6.00
	10-99	-58.34	-35.99	-47.24	6.42
	0-99	-58.34	-35.99	-47.12	6.41

Table 6.10: TΔS statistics of the extended WT and HM ensembles

Pr ensemble		**TΔS statistics (kcal/mol)**			
		Lower range	**Upper range**	**Mean**	**STDEV**
WT	0-9	-32.53	-28.47	-30.71	1.11
	10-19	-35.18	-27.15	-31.67	2.01
	0-19	-35.18	-27.15	-31.21	1.68
HM	0-9	-33.23	-28.14	-30.52	1.48
	10-99	-33.73	-27.31	-30.71	1.48
	0-99	-33.73	-27.31	-30.64	1.48

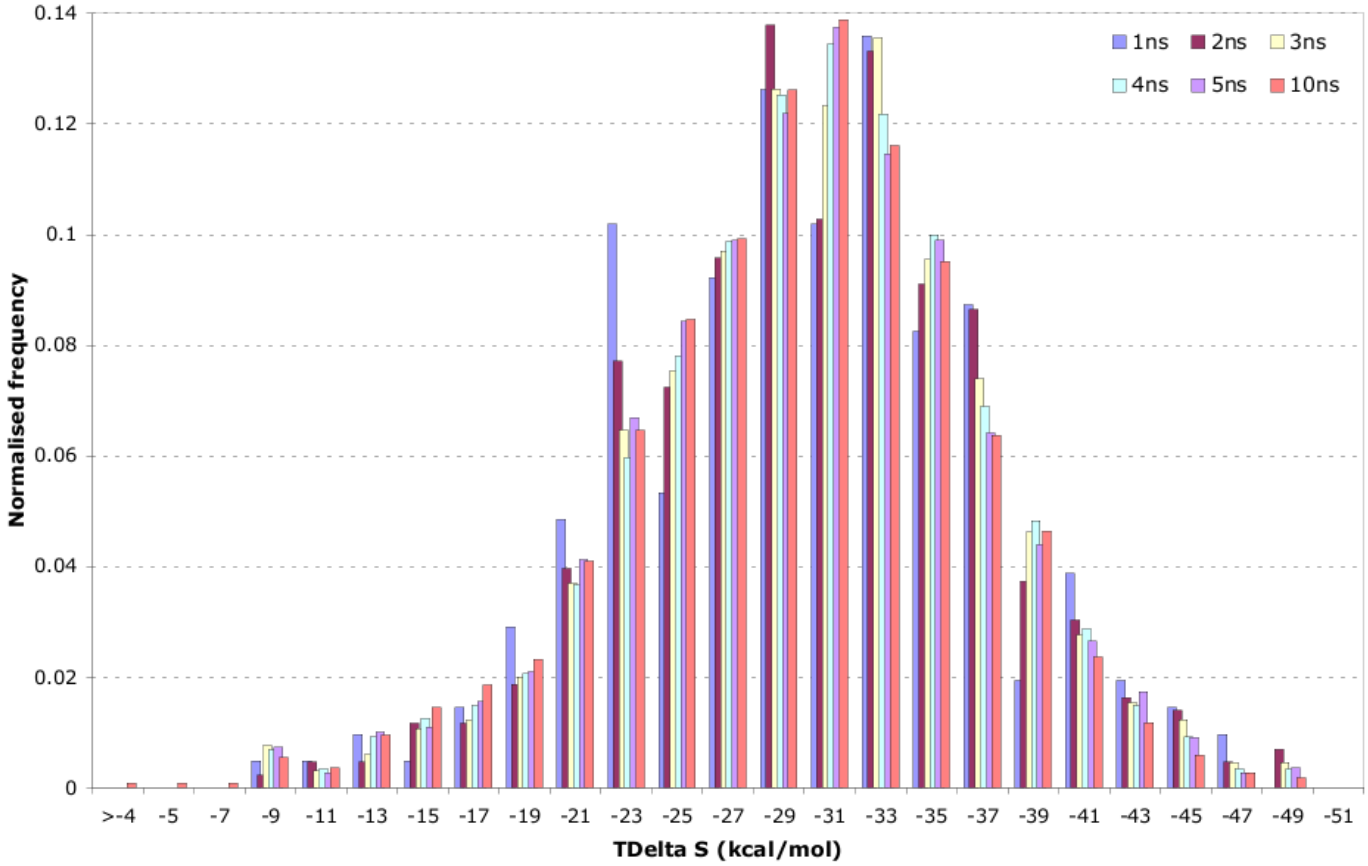


Figure 6.19: Comparison of entropic frequency distributions across varying lengths of production-phases for the HM system. The results show that by 5ns the distribution becomes equivalent to the distribution over the full 10ns.

From these results, the conclusion was made that the variation of binding affinities within an ensemble is due almost solely to differences in the enthalpic component, with the entropic component essentially a constant value at $\sim$ 31kcal/mol for WT and $\sim$ 30.5kcal/mol for HM. The small variations observed in the entropic component were due to limitations of the normal mode analysis methodology, whereas the variations observed in the enthalpic component were representative of the different configurations that must be considered collectively to calculate a binding affinity comparable to experimental. That the range of ΔG encompasses the experimental value for both WT and HM further reinforces the conclusion that individual binding affinity values are unimportant, and that the population as a whole, and the population's ΔH distribution must be analysed. Although it was determined that the number of configurations sampled by an ensemble does not increase as the simulation length is extended beyond 1ns, it was also determined that while a 10-repetition ensemble is enough to sample sufficient configurations to calculate a WT binding affinity value comparable to experimental, 100 is not enough to sample sufficient configurations to calculate an HM binding affinity comparable to experimental.

PCA of both the WT and HM systems was restricted to the first nanosecond of data-collection because the amount of data created by each ensemble was too great for either PCAZIP or PTRAJ programs to handle. As the results in this chapter showed that analysis of the first nanosecond was a sufficient representation of the 10ns simulation as a whole, the configurational sampling of the two extended ensembles were analysed from their first nanosecond. The primary eigenvector of 20-repetition 1ns WT ensemble showed a correlation coefficient of $r = -0.63$ to each repetition's ΔH value. While the correlation is not as strong as with the primary eigenvector of the 10ns simulation, this is still a significant correlation. Visual inspection of the location of the correlated motions comprising the primary eigenvector show that they are almost identical to those in Figure 6.16(a) (data not shown). Plotting the projections of primary principle component against the secondary principle component show that there is almost no additional sampling of these eigenvectors by doubling the number of repetitions (Figure 6.20). This suggests that the WT conformational landscape is effectively sampled within 10 repetitions.

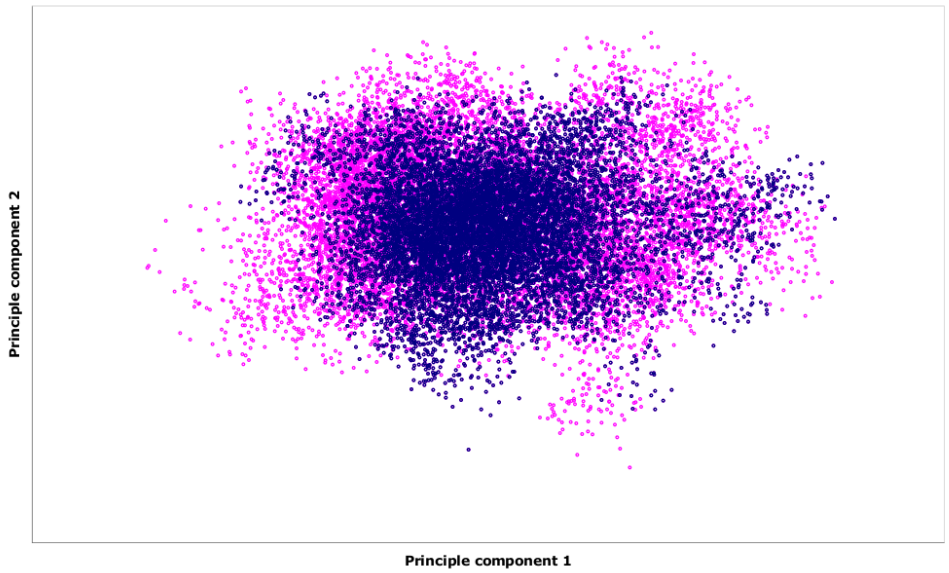
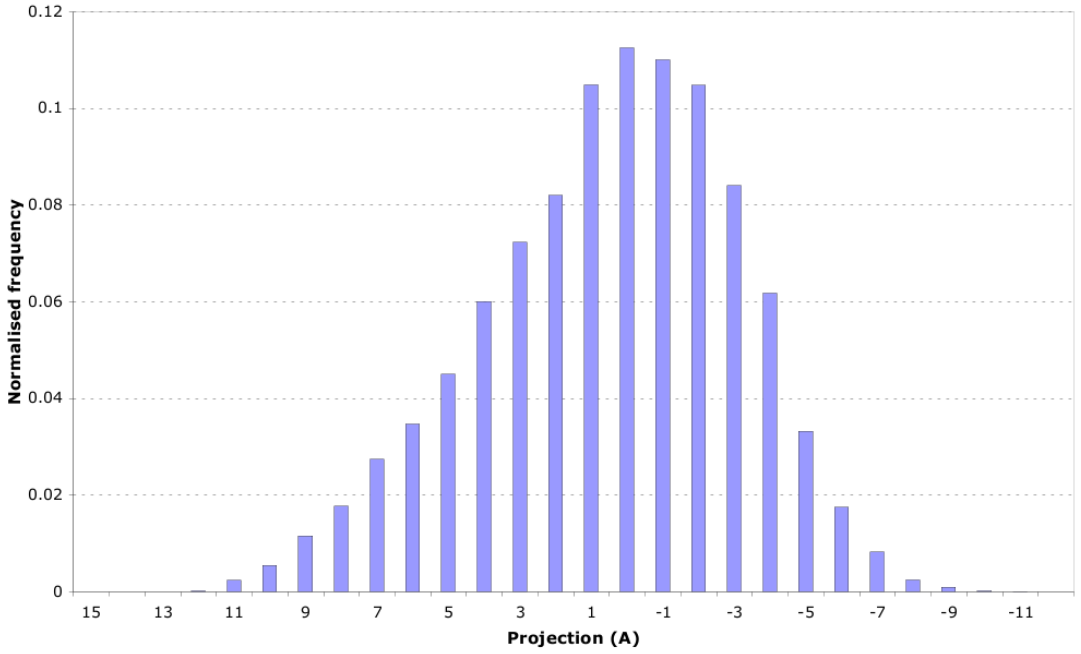
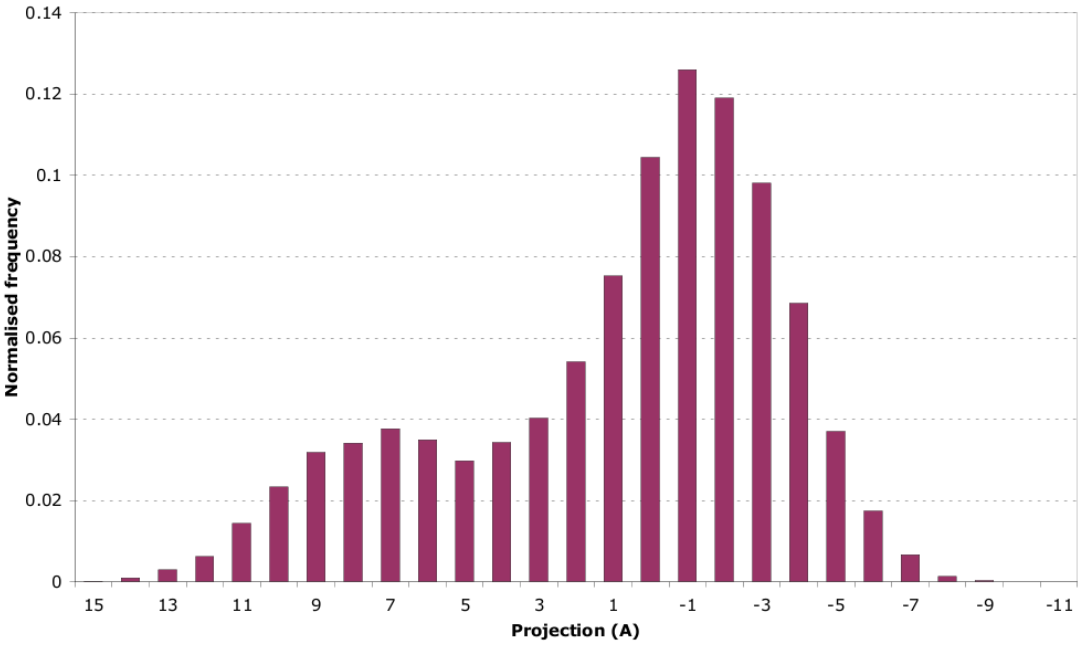


Figure 6.20: Projections of the first 2 principle components of the WT 20-repetition 1ns ensemble plotted against each other. In pink are the configurations adopted across the whole ensemble, and in dark-blue are the configurations adopted by the first 10-repetitions. The results show that little additional sampling is achieved upon extension to 20 repetitions for the WT system.

Analysis of the frequency distribution of the principle component for each of the 100 1ns HM repetitions revealed a very strong correlation between a repetition's modal projection and its average ΔH (data not shown). The correlation coefficient between the two was $r = 0.90$, which means that the regions of the correlation motions captured by the primary eigenvector that are sampled by the repetition are directly correlated to the strength of binding between the protease and the inhibitor. Visualisation of the location of these correlated motions showed that they are almost identical to those motions of the secondary eigenvector for the 10-repetition 10ns simulation (Figure 6.16(b)). Subsequent eigenvectors analysed showed no correlation to the ΔH, indicating that these motions alone are responsible for the calculated binding affinity value. The frequency distribution of the primary eigenvector's projections across the whole ensemble highlighted the cause of the erroneous calculated binding affinity value for this system (Figure 6.21(b)).



(a)



(b)

Figure 6.21: Frequency distribution of the primary eigenvector's projections for the 20-repetition 1ns WT ensemble (a) and the 100-repetition 1ns HM ensemble (b). (a) Results show a normal distribution about the projection which shows moderate correlation to a ΔH of ~ -44kcal/mol. (b) Results show a bimodal distribution; the large peak corresponds to simulations where the ΔH is ~ -50kcal/mol; and the small peak corresponds to simulations where the ΔH is ~ -38kcal/mol, which corresponds to a decreased sensitivity to the inhibitor.

The bimodal distribution has a dominant peak at a projection of -1 and a much smaller peak at +7. The average ΔH for the 1ns simulations that fall in the larger peak is ~ -50kcal/mol, while those that fall in the smaller peak have an average ΔH of ~ -38kcal/mol. Taking the entropy of the system to be constant at -30kcal/mol (taken from Table 6.5), these peaks represent binding affinities of ~ -20kcal/mol and ~ -8kcal/mol for the larger and smaller peaks respectively. In contrast, the frequency distribution of the primary eigenvector's projections across the 20-repetition 1ns ensemble show a normal distribution. The modal projection shows moderate correlation to a ΔH value of ~ -44kcal/mol which, when considered with a TΔS value of -31kcal/mol (value taken from Table 6.5) gives a binding affinity of ~ -13kcal/mol.

6.6 Conclusions

The results presented in this chapter show that MMPBSA and normal mode analysis of a single molecular dynamics simulation of HIV-1 protease complexed with saquinavir is insufficient for calculating a binding affinity value. The single trajectories of the 6 protease systems simulated in Section 6.2 resulted in a strong negative correlation to experimental data published by Ohtaka *et al.* (2003). However, subsequent expansion of the WT and HM systems into ensembles showed considerable overlap in the ranges of ΔG values exhibited by their constituent simulations. Therefore, all types of correlation; from strongly-positive, to no-correlation, to the strongly-negative correlation observed, are possible from combinations of the individual simulations contained within the ensembles. While energy analysis of single trajectories of single simulations has been successfully applied to HIV-1 protease complexed with a protease-inhibitor [120, 50], these findings strongly suggest that the results would not be consistently reproducible, and therefore this methodology is not suitable as a diagnostic tool for predicting drug-resistance phenotype from the genotype.

Although expansion of the WT and HM systems into ensembles was shown through PCA to improve the conformational sampling, it was still not able to replicate the experimental drug-resistance trend. While the extended WT ensemble gave an average ΔG only 0.63kcal/mol more negative than the experimental value, the average ΔG

across the HM ensemble was 8.61kcal/mol more negative than the experimental value and 3.48kcal/mol more negative than the computational WT value. Therefore, while a mean of the ΔG values within an ensemble was sufficient for the WT system, it was insufficient for the HM ensemble. The spread of ΔG values within each ensemble was shown to be moderately-to-highly correlated to a set of concerted motions in the protease; the WT ensemble showed a normal distribution of the frequency about which it sampled these motions, with the modal configuration's binding affinity equivalent to the experimental binding affinity. The HM ensemble, conversely, showed an uneven bimodal distribution of the frequency at which it sampled these motions. The dominant peak's modal configuration had a binding affinity of approximately -20kcal/mol, whilst the minor peak's modal configuration had a binding affinity of -8kcal/mol. These results therefore suggest a hitherto unseen dynamic in the HM protease where it transitions between a configuration with a stronger affinity to saquinavir than the WT protease, and a configuration with a weaker affinity to saquinavir than the WT protease. These transitions could act as a mechanism by which the HM protease reduces its sensitivity to saquinavir; the infrequent transitions to configurations with a reduced binding affinity to saquinavir could cause the expulsion of the inhibitor. Meanwhile, these configurational dynamics may have less of an effect on the substrate because the ES complex only need to transiently form in order for the catalysis to occur, whereas the competitive inhibitor fulfills its role by staying as an EI complex for as long a possible, thus preventing the enzyme from being able to catalyse any reactions.

However, this is an interpretation of the results, and does not address the fact that the HM ensemble's minor peak corresponds to the experimental binding affinity value of -8.5kcal/mol, and therefore the larger peak at -20kcal/mol brought the average value down to -17.11kcal/mol in the 100-repetition ensemble. It may be due to the pre-simulation protocol that was followed for all simulations; the mutation and subsequent minimisation protocol may result in the minimisation of the structure towards one of two configurational minima. The majority of the time the structure is directed into a non-physiological configurational well which should not be considered, and only occasionally does the structure minimise down the gradient towards the minimum representing physiological configurations. A third interpretation of these results is that this

ensemble is not sufficiently large enough to correctly sample the required landscape. The configurations relating to the minor peak only became apparent upon extension of the ensemble to 100 repetitions, so it may be necessary to extend the ensemble even further to see how these frequency distribution of these correlated motions changes. It may be that the smaller peak becomes more significant, and with a better sampling of those configurations, the modal value of the peak shifts towards a less negative ΔG configuration that subsequently brings the ensemble average to a less negative value than the WT system. Extension of this work should therefore concentrate on expansion of the HM ensemble to determine whether these low-frequency configurations become more significant.

Chapter 7

Final discussion and future work

7.1 Final discussion

The aim of this thesis was to investigate whether explicit-solvent, nanosecond-timescale molecular dynamic simulations of HIV-1 protease complexed with saquinavir could be analysed through the heuristic MMPBSA methodology to calculate an accurate binding affinity value, which would allow prediction of the effects of drug-resistance mutations. This would have clinical applications, where it could be used for the prediction of drug-resistance phenotype straight from the viral genotype much faster than current clinical methods. In this thesis a local SQL database was generated to address the lack of a collective resource for genotypic, structural and biochemical data on HIV-1 protease. This database connected the various data according to the sequence of the protease on which the experiment was performed, which allowed for the rapid collation of the data for a particular sequence using a simple SQL query. In Chapter 5 it was shown that up to 13 mutations in each monomer (26 residues in the homodimer) could be computationally alchemically-mutated in the protease without adversely affecting the structure or flexibility of the protease. This novel study is important as it showed that HIV-1 protease can be computationally converted into a wide array of genotypes, thereby greatly increasing the possible numbers of genotypes that can be simulated through molecular dynamics. This has a significant impact on its practical applications, as simulations of drug-resistant protease genotypes observed *in vivo* will not be restricted to those with previously-determined crystal structures.

In Chapter 6, it was shown that single trajectory molecular dynamics simulations of HIV-1 protease complexed with saquinavir were insufficient for accurate calculation of a binding affinity, and are therefore insufficient for calculation of a change in drug resistance between genotypes. Molecular dynamics ensembles of the wild-type (WT) and multi-drug resistant (MDR) proteases were shown to improve conformational sampling, and for the WT system was able to calculate an accurate binding affinity, but was unable to do so for the MDR system. This was suggested to be due to an insufficient number of repetitions in the MDR ensemble, which meant that the configurational landscape was not adequately sampled. This indicates that the number of replicates required to adequately sample a system's configurational landscape is specific to that system, making automation of the simulations, which is necessary for clinical application, difficult because the required number of replicates in an ensemble is unknown. This can be addressed by consistently running a set number of replicates for each simulation that will guarantee configurational sampling for all systems, but this is inefficient and not conducive for a rapid turnaround of results. Furthermore, this negates the benefit of calculating the binding affinity through the heuristic MMPBSA method; if the rapid calculation of a binding affinity for a single simulation is offset by the requirement to analyse hundreds of simulations in order to attain an accurate value, then the method is not time-effective or resource-efficient. Thermodynamic Integration (TI) is a highly-accurate method of calculating the change in binding affinity between two systems. Its accuracy comes at the cost of a prohibitively-high computational demand - requiring approximately 40 simulations, each several nanoseconds in length [21]. However, the findings of this thesis show that the computational requirements for analysing an ensemble through MMPBSA is of the same order as TI. With the availability of Grid networks and supercomputers at an unprecedented level, application of TI to the systems studied in this thesis should be considered.

A requirement for practical application of MMPBSA analysed MD simulations is a rapid turnaround of results, in the order of days instead of weeks for current diagnostic methods. The production of new supercomputers such as HECToR at Edinburgh, and Legion at University College London, along with the ever-increasing network of

computers available for scientific research means that multiple simulations can be run simultaneously, thereby greatly increasing the throughput. The number of available processors has meant that the limitation in the turnaround of results has become the speed at which parallel processors can communicate with each other. It was shown in this thesis that it is possible to run a 3ns molecular dynamics simulation in 18-24 hours. However, the greater bottleneck was the post-simulation processing - transferring the data back from the various supercomputers for analysis, and the non-parallelised normal mode analysis and PCA. Together these require another day per 2 or 3 simulations to generate a result. Therefore, with the current availability of computational resources, it is possible to generate a binding affinity value for a single simulation in approximately 2 days, though transferring the data for a 100-repetition ensemble takes approximately 1 week with standard 'SCP' or 'SFTP'. The ability to perform the post-processing analyses on supercomputers will greatly enhance this, and allow the turnaround of results in approximately 2 days.

Another finding in this thesis is a previously un-postulated mechanism of drug resistance in the HM system of Ohtaka *et al.* (2003), based on principle components analysis of the HM ensemble. This system was shown to transition between configurations with a strong affinity for saquinavir, to configurations with an affinity lower than the WT system at an approximately 30% frequency. This transitioning would result in a higher chance of dissociation of the enzyme-inhibitor complex. However, while a concerted set of motions has been shown to be correlated to the calculated ΔG, it is unlikely that all residues included in these correlated motions are necessary, so the PCA should be re-calculated for subsets of the WT and HM backbones to determine the key residues involved. This would lead to an understanding of *how* their configurations result in a lower binding affinity. This may be due to the a transient reduction in the interaction energy between protease and saquinavir, similar to the mechanism postulated for the D30N mutation, or it may be due to a transient reduction in the hydrogen bonding pattern, as therosied for the N88S mutation[122]. Similar dynamic mechanisms have been theorised for the V82F/I84V mutant, which was postulated to more frequently adopt a 'semi-open flaps' configuration that results in the inhibitor having a greater enthalpic cost upon binding [77].

7.2 Future work

In light of the results presented in this thesis, it is apparent that further work is necessary to understand why the HM ensemble is unable to calculate an accurate binding affinity through the MMPBSA methodology. It was shown that particular correlated motions in each system were strongly associated with the strength of binding between the protease and saquinavir, but the mechanism by which these motions affect the binding was not ascertained. Therefore, future work should focus, in the first instance, on determining the key residues whose motions correlate with the calculated binding affinity. Recognising which residues affect the binding affinity will enable the mechanism of drug resistance in the HM system to be theorised. However, this does not address the inability to calculate an accurate binding affinity for the HM ensemble. In order to determine whether this was due to a lack of configurational sampling, the HM ensemble should be expanded to observe whether an improvement on the average ΔG with respect to the experimental value is attained. Performing PCA on the expanded HM ensemble will determine whether an improvement in configurational sampling is achieved. Analysis of the other 4 protease genotypes in the paper by Ohtaka *et al.* is necessary to determine whether the inability to calculate a binding affinity value for the HM system occurs in this system alone, or whether the anti-correlation still persists. Therefore, the 4 remaining systems should be expanded to ensembles and analysed through PCA in the same manner as presented in this thesis. Due to the findings in this thesis, the number of snapshots for normal mode calculations should be increased to approximately 25 per nanosecond. The increase in computational resources available at present facilitates the subsequent increased computational requirements.

Another consideration, that was not addressed in this thesis, is the change in free energy associated with conformational change upon complex formation. For each system a single simulation of the complex was run, and the conformations of the receptor and ligand necessary for calculation of the change in free energy upon ligand binding (Equation 2.14) attained by removing the unnecessary atoms in the complex file. Therefore there are no values representing the change in internal energy upon ligand binding through the simulations. This was to reduce the computational requirements

for each simulation as otherwise 3 separate simulations would have to be run for the ligand, receptor, and ligand-receptor complex, which would nearly triple the required computational resources. However, with the increase in available computational resources since conception of this research, addressing this issue should be addressed to give a more realistic free energy calculation. However, while this may improve the results, this should be considered against the substantial increase in computational resource required which goes against the aim of rapidly generating binding energies using a heuristic methodology.

Future work beyond these HIV-1 protease systems should be considered with caution; the findings in this thesis are applicable only to this system. Therefore, these findings cannot apply to, for example, the reverse transcriptase enzyme of HIV. There are also other clinically-relevant viral protease enzymes against which inhibitors have been designed; for example hepatitis C virus (HCV), human cytomegalovirus (HCMV), herpes simplex virus type 1 (HSV-1) and picornavirus [75]. The findings in this thesis may have some application to these enzymes, but as they have no structural homology to the HIV protease, the proposed mechanism of resistance is not cross-applicable. Nevertheless, the mutation and equilibration protocol employed by the BAC, and utilised in this research, should be robust enough to mutate and structurally-equilibrate other crystal structures. This is reinforced by the utilisation of this protocol by collaborators for MD simulations of HIV reverse-transcriptase, which is considerably larger and less globular than HIV protease. The findings of this thesis also suggest that simulations on other systems should be run as short-timescale ensembles, but that the number of replicates required to attain a binding energy value will be dependant on its configurational landscape.

Bibliography

[1] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1994) *Molecular Biology of the Cell, Third Edition*. New York: Garland Publishing, Inc.

[2] Amber Advance Tutorials: Tutorial 3 MM/PBSA (2006) <http://amber.scripps.edu/tutorials/advanced/tutorial3/index.htm>.

[3] Architecture Overview (2008) Viewed 6th April 2009 <http://www.hector.ac.uk/support/documentation/userguide/hectoruser/Architecture_Overview.html>.

[4] Balsera, M. A., Wriggers, W., Oono, Y. and Schulten, K. (1996) "Principal Component Analysis and Long Time Protein Dynamics" *The Journal of Physical Chemistry* 100 **7** 2567-572.

[5] Basic Chemistry: Atoms and Ions (2002) Viewed 10th April 2009 <http://www.personal.psu.edu/staff/m/b/mbt102/bisci4online/chemistry/chemistry8.htm>.

[6] Beerenwinkel, N., Däumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J. and Walter, H. (2003) "Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes" *Nucleic Acids Research* 31 **13** 3850-3855.

[7] Berg, J. M., Tymoczko, J. L. and Stryer, L. (2002) *Biochemistry, fifth edition (International edition)*. New York: W. H. Freeman and Company.

[8] The Binding Database (2007) Viewed 3rd November 2005, <http://www.bindingdb.org>.

[9] Boden, D. and Markowitz, M. (1998) "Resistance to Human Immunodeficiency Virus Type 1 Protease Inhibitors" *Antimicrobial Agents and Chemotherapy* 42 **11** 2775-2783.

[10] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. and Karplus, M. (1983) "CHARMM: A program for macromolecular energy, minmimization, and dynamics calculations" *Journal of Computational Chemistry* **4** 187-217.

[11] Carvajal-Rodríguez, A. (2007) "The Importance of Bio-Computational Tools for Predicting HIV Drug Resistance" *Recent Patents on DNA and Gene Sequences* **1** 63-68.

[12] Campbell, M. K. (1999) *Biochemistry, Third Edition.* Florida: Harcourt Brace & Company.

[13] Ceccherini-Silberstein, F., Erba, F., Gago, F., Bertoli, A., Forbici, F., Bellocchi, M. C., Gori, C., d'Arrigo, R., Marcon, L., Balotta, C., Antinori, A., d'Arminio Monforte, A., Perno, C-F. (2004) "Identification of the minimal conserved structure of HIV-1 protease in the presence and absence of drug pressure." *AIDS* **18** 11-19.

[14] Copeland, R. A. (2000) *Enzymes: A Practical Introduction to Structure, Mechanism and Data Analysis, Second Edition.* New Jersey: Wiley-Blackwell.

[15] Cornish-Bowden, A. (2004) *Fundamentals of Enzyme Kinetics, Third Edition.* London: Portland Press Ltd.

[16] Domingo, E. (2006) *Quasispecies: Concepts and Implications for Virology.* Birkhäuser Publications.

[17] European Molecular Biology Network (2007) Viewed 30th April 2007, <http://www.ch.embnet.org/MD_tutorial>

[18] Feinberg, M. B., Baltimore, D. and Frankel, A. D. (1991) "The role of Tat in the human immunodeficiency virus life cycle indicates a primary effect on transcriptional elongation" *P.N.A.S.* **88** 4045-4049.

[19] Ferdinand, W. N. M., van Leeuwen, R., Weverling, G. J., Jurriaans, S., Nauta, K., Steingrover, R., Schuijtemaker, J., Eyssen, X., Fortuin, D., Weeda, M., de Wolf, F., Reiss, P., Danner, S. A. and Lange, J. M. A. (1999) "Outcome and Predictors of Failure of Highly Active Antiretroviral Therapy: One Year Follow-Up of a Cohort of Human Immunodeficiency Virus Type 1-Infected Persons" *Journal of Infectious Diseases* **179** 790-798.

[20] Fowler, J., Cohen, L., Jarvis, P. (2001) *Practical Statistics for Field Biology, Second Edition.* Chichester: John Wiley and Sons Ltd.

[21] Fowler, P. W., Jha, S. and Coveney, P. V. (2005) "Grid-based steered thermodynamic integration accelerates the calculation of binding free energies" *Philosophical Transactions of the Royal Society***363** 1999-2015.

[22] Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A. and Schulten, K. (2006) "Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus" *Structure* **14** 437-449.

[23] Frutos, S., RodriguezMias, R. A., Madurga, S., Collinet, B. Reboud-Ravaux, M., Ludevid, D. and Giralt, E. (2007) "Disruption of the HIV-1 Protease Dimer with Interface Peptides: Structural Studies Using NMR Spectroscopy Combined with [2- ^{13}C]-Trp Selective Labeling" *Peptide Science* 88 **2** 164-173.

[24] Ganser-Pornillos, B. K., Yeager, M. and Sundquist, W. I. (2008) "The structural biology of HIV assembly" *Current Opinion in structural biology* **18** 203-217.

[25] Geretti, A. M., Harrison, L., Green, H., Sabin, C., Hill, T., Fearnhill, E., Pillay, D. and Dunn, D. (2009) "Effect of HIV-1 Subtype on Virologic and Immunologic Response to Starting Highly Active Antiretroviral Therapy" *Clinical Infectious Diseases* **48** 1296-1305.

[26] Gohlke, H. and Case, D. A. (2003) "Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein-Protein Complex Ras-Raf" *Journal of Computational Chemistry* 25 **2** 238-250.

[27] Gulnik, S. V., Suvorov, L. I., Liu, B., Yu, B., Anderson, B., Mitsuya, H. and Erickson, J. W. (1995) "Kinetic characterization and Cross-Resistance Patterns

of HIV-1 Protease Mutants Selected under Drug Pressure" *Biochemistry* **34** 9282-9287.

[28] Hanes, J. W. and Johnson, K. A. (2008) "Exonuclease Removal of Dideoxycytidine(Zalcitabine) by the Human Mitochondrial DNA Polymerase" *Antimicrobial Agents and Chemotherapy* 52 **1** 253-258.

[29] Haynie, D. T. (2008) *Biological Thermodynamics, Second Edition.* Cambridge: University Press.

[30] Hilser, V. J., García-Moreno, B. E., Oas, T. G., Kapp, G. and Whitten, S. T. (2006) "A Statistical Thermodynamic Model of the Protein Ensemble" *Chemical Reviews* **106** 1545-1558.

[31] HIV Drug Resistance Database (1999) Viewed 10th April 2009, <http://hivdb.stanford.edu/pages/documentPage/primer.html>.

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

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

[34] Hornak, V. and Simmerling, C. (2007) "Targeting structural flexibility in HIV-1 protease inhibitor binding" *Drug Discovery Today* 12 **3/4** 132-138.

[35] HPC Systems (2008) Viewed 5th April 2009, <http://www.tacc.utexas.edu/resources/hpcsystems/>.

[36] Ishima, R., Freedberg, D. I., Wang, Y. X., Louis, J. M. and Torchia, D. A. (1999) "Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function." *Structure* 7 **9** 1047-1055.

[37] Isothermal titration calorimetry to measure protein/peptide/lipid interactions. Viewed 10th April 2009, <http://www2.mrc-lmb.cam.ac.uk/groups/hmm/techniqs/ITC.html>.

[38] Ives, K. J., Jacobsen, H., Galpin, S. A., Garaev, M. M., Dorrell, L., Mous, J., Bragman, K. and Weber, J. N. (1997) "Emergence of resistant variants of HIV *in vivo* during monotherapy with the proteinase inhibitor saquinavir" *Journal of Antimicrobial Chemotherapy* **39** 771-779.

[39] Jayaram, B., Sprous, D. and Beveridge, D. L. (1998) "Solvation Free Energy of Biomacromolecules: Parameters for a Modified Generalized Born Model Consistent with the AMBER Force Field" *Journal of Physical Chemistry B* **102** 9571-9576.

[40] Ji, J. and Loeb, L. A. (1992) "Fidelity of HIV-1 Reverse Transcriptase Copying RNA in Vitro" *Biochemistry* **31** 954-958.

[41] Kagan, R. M., Shenderovich, M. D., Heseltine, P. N. R., Ramnarayan, K. (2007) "Structural analysis of an HIV-1 protease I47A mutant resistant to the protease inhibitor lopinavir" *Protein Science* **14** 1870-1878.

[42] Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K. and Schulten, K. (1999) "NAMD2: Greater Scalability for Parallel Molecular Dynamics" *Journal of Computational Physics* **151** 283-312.

[43] Kalra, P., Das, A. and Jayaram, B. (2001) "Free-energy analysis of enzyme-inhibitor binding" *Applied Biochemistry and Biotechnology* **96** 93-108.

[44] Kantor, R., Fessel, W. J., Zolopa, A. R., Israelski, D., Schulman, N., Montoya, J. G., Harbour, M., Schapiro, J. M. and Shafer, R. W. (2002) "Evolution of Primary Protease Inhibitor Resistance Mutations during Protease Inhibitor Salvage Therapy" *Antimicrobial Agents and Chemotherapy* 46 **4** 1086-1092.

[45] Kantor, R., Katzenstein, D. A., Efron, B., Carvalho, A. P., Wynhoven, B., Cane, P., Clarke, J., Sirivichayakul, S., Soares, M. A., Snoeck, J., Pillay, C., Rudich, H., Rodrigues, R., Holguin, A., Ariyoshi, K., Bouzas, M. B., Cahn, P., Sugiura, W., Soriano, V., Brigido, L. F., Grossman, Z., Morris, L., Vandamme, A-M., Tanuri, A., Phanuphak, P., Weber, J. N., Pillay, D., Harrigan, P. R., Camacho,

R., Schapiro, J. M., Shafer, R. W. (2005) "Impact of HIV-1 Subtype and Antiretroviral Therapy on Protease and Reverse Transcriptase Genotype: Results of a Global Collaboration" *Public Library of Science: Medicine* 2 **4** 0325-0337.

[46] Katoh, E., Lousid, J. M., Yamazaki, T., Gronenborn, A. M., Torchia, D. A. and Ishima, R. (2003) "A solution NMR study of the binding kinetics and the internal dynamics of an HIV-1 protease-substrate complex" *Protein Science* **12** 1376-1385.

[47] Knipe, D. M. and Howley, P. M. (2007) *Fields Virology, Fifth Edition*. Philadelphia: Lippincott, Williams and Wilkins.

[48] Kurt, N., Scott, W. R. P., Schiffer, C. A. and Haliloglu, T. (2003) "Cooperative Fluctuations of Unliganded and Substrate-Bound HIV-1 Protease: A Structure-Based Analysis on a Variety of Conformations from Crystallography and Molecular Dynamics Simulations" *PROTEINS: Structure, Function, and Bioinformatics* **51** 409-422.

[49] Leach, A. R. (2001) *Molecular Modelling: Principles and Applications, Second Edition*. Harlow: Pearson Education Ltd.

[50] Lepsik, M., Kriz, Z. and Havlas, Z. (2004) "Efficiency of a second-generation HIV-1 protease inhibitor studied by molecular dynamics and absolute binding free energy calculations" *PROTEINS: Structure, Function, and Bioinformatics* **57** 279-293.

[51] Levy, J. A. (1992) *The Retroviriae Volume 1* New York: Plenum Press.

[52] Lin, H. and Truhlar, D. G. (2007) "QM/MM: what have we learned, where are we, and where do we go from here?" *Theoretical Chemistry Accounts* **117** 185-199.

[53] Lonestar User Guide (2009) Viewed on 6th April 2009 <http://services.tacc.utexas.edu/index.php/lonestar-user-guide>.

[54] LRZ: Overview of the Cluster Configuration (2008) Viewed on 6th April 2009 <http://www.hector.ac.uk/support/documentation/userguide/hectoruser/Architecture_Overview.html>.

[55] Macromolecular Crystallography Facility: Crystallography 101 (2007) Viewed 9th September 2007, <http://xray0.princeton.edu/ phil/Facility/x101/x101.html>

[56] Mager, P. P. (2001) "The Active Site of HIV-1 Protease" *Medicinal Research Reviews* 21 **4** 348-353.

[57] Mallolas, J., Blanco, J. L., Labarga;, P., Vergara, A., Ocampo, A., Sarasa, M., Arnedo, M., Lpez-Pa, Y., García, J., Juega, J., Guelar, A., Terrón, A., Dalmau, D., García, I., Zárraga, M., Martínez, E., Carné, X., Pumarola, T., Escayola, R. and Gatell, J. M. (2007) "Inhibitory Quotient as a Prognostic Factor of Response to a Salvage Antiretroviral Therapy Containing Ritonavir-boosted Saquinavir. The CIVSA Study" *HIV Medicine* 8 **4** 226-233.

[58] Martinez-Picado, J., Savara, A. V., Sutton, L. and D'Aquila, R. (1999) "Replicative Fitness of Protease Inhibitor-Resistant Mutants of Human Immunodeciency Virus Type 1" *Journal of Virology* 73 **5** 3744-3752.

[59] Maschera, B., Darby, G., Palú, G., Wright, L. L., Tisdale, M., Myers, R., Blair, E. D. and Furfine, E. S. (1996) "Human Immunodeficiency Virus. Mutations in the viral protease that confer resistance to saquinavir increase the dissociation rate constant of the protease-saquinavir complex" *Journal of Biological Chemistry* 271 **52** 33231-33235

[60] Meagher, K. L. and Carlson, H. A. (2005) "Solvation Influences Flap Collapse in HIV-1 Protease" *Proteins: Structure, Function and Bioinformatics* **58** 119-125.

[61] Meyer, T., Ferrer-Costa, C., Pérez, A., Rueda, M., Bidon-Chanal, A., Luque, F. J., Laughton, C. A. and Orozco, M. (2006) "Essential Dynamics: A Tool for Efficient Trajectory Compression and Management" *Journal of Chemical Theory and Computation* 2 **2** 251-258.

[62] Moumen, A., Polomack, L., Roques, B., Buc, H. and Negroni, M. (2001) "The HIV-1 repeated sequence R as a robust hot-spot for copy-choice recombination" *Nucleic Acids Research* 29 **18** 3814-3821.

[63] Moumen, A., Polomack, L., Unge, T., Véron, M., Buc, H. and Negroni, M. (2003) "Recombination during reverse transcription: evidence for a mechanism

dependent on the structure of the acceptor RNA" *Journal of Biological Chemistry* **278** 15973-15982.

[64] NAMD and molecualr dynamics simulations (2002) Viewed 11th February 2009 <http://www.ks.uiuc.edu/Research/namd/2.7b1/ug/node5.html>

[65] National Institutes for Health: Center for Molecular Modeling (2007) Viewed 9th September 2007, <http://cmm.cit.nih.gov/>

[66] Negroni, M. and Buc, H. (2000) "Copy-choice recombination by reverse transcriptases: Reshuffling of genetic markers mediated by RNA chaperones" *Proceedings of the National Academy of Sciences of the United States of America* 97 **12** 6385-6390.

[67] Ode, H., Neya, S., Hata, M., Sugiura, W. and Hoshino, T. (2006) "Computational Simulations of HIV-1 Proteases - Multi-drug Resistance Due to Nonactive Site Mutation L90M" *Journal of the American Chemical Society* 128 **24** 7887-7895.

[68] Ode, H., Ota, M., Neya, S., Hata, M., Sugiura, W. and Hoshino, T. (2004) "Resistant Mechanism against Nelfinavir of Human Immunodeficiency Virus Type 1 Proteases" *Journal of Physical Chemistry B* **109** (1) 565-574.

[69] Ohtaka, H., Schön, A. and Freire, E. (2003) "Multidrug Resistance to HIV-1 Protease Inhibition Requires Cooperative Coupling between Distal Mutations" *Biochemistry* 42 **46** 13659-13666.

[70] Ohtaka, H., Muzammil, S., Schön, A., Velazquez-Campoy, A., Vega, S. and Freire, E. (2004) "Thermodynamic rules for the design of high affinity HIV-1 protease inhibitors with adaptability to mutations and high selectivity towads unwanted targets" *The International Journal fo Biochemistry and Cell Biology* **36** 1787-1799.

[71] Okimoto, N., Hata, M., Hoshino, T. and Tsuda, M. (2000) "Protein hydrolysis mechanism of HIV-1 protease: Investigation by the *ab initio* MO calculations" *RIKEN Review* **29**

[72] de Oliveira, T., Engelbrecht, S., van Rensburg, E. J., Gordon, M., Bishop, K., zur Megede, J., Barnett, S. W. and Cassol, S. (2003) "Variability at Human Immunodeficiency Virus Type 1 Subtype C Protease Cleavage Sites: an Indication of Viral Fitness?" *Journal of Virology* 77 **17** 9422-9430.

[73] Onufriev, A., Case, D. A. and Bashford, D. (2002) "Effective Born Radii in the Generalized Born Approximation: The Importance of being Perfect" *Journal of Computational Chemistry* 23 **14** 1297-1304.

[74] Page, C. S. and Bates, P. A. (2006) "Can MM-PBSA Calculations Predict the Specificities of Protein Kinase Inhibitors?" *Journal of Computational Chemistry* 27 **16** 1990-2007.

[75] Patick, A. K. and Potts, K. E. (1998) "Protease Inhibitors as Antiviral Agents" *Clinical Microbiology Reviews* 11 **4** 614-627.

[76] Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E. III, DeBolt, S., Ferguson, D., Seibel, G. and Kollman, P. (1995) "AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules" *Computer Physics Communications* **91** 1-41.

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

[78] Perryman, A. L., Lin, J-H. and McCammon, J. A. (2006) "Restrained Molecular Dynamics Simulations of HIV-1 Protease: The First Step in Validation a New Target for Drug Design" *Biopolymers* **82** 272-284.

[79] Pettit, S. C., Moody, M. D., Wehbie, R. S., Kaplan, A. H., Nantermet, P. V., Klein, C. A. and Swanstrom, R. (1994) "The p2 Domain of Human Immunodeficiency Virus Type 1 Gag Regulates Sequential Proteolytic Processing and Is Required To Produce Fully Infectious Virions" *Journal of Virology* 68 **12** 8017-8027.

[80] Piana, S., Carloni, P. and Parrinello, M. (2002) "Role of Conformational Fluctuations in the Enzymatic Reaction of HIV-1 Protease" *Journal of Molecular Biology* **319** 567-583.

[81] Pillay, D. (1995) "HIV-1 Protease Inhibitors: Their Development, Mechanism, Action and Clinical Potential" *Reviews in Medicinal Virology* 5 **1** 23-33.

[82] Plimpton, S. J. (1995) "Fast parallel algorithms for short-range molecular dynamics" *Journal of Computational Physics* **117** 1-19.

[83] Posch, H. A. and Hoover, W.G. (2006) "Lyapunov Instability of Classical many-Body Systems" *Journal of Physics: Conference Series* **31** 9-17.

[84] Preston, B. D., Poiesz, B. J. and Loeb L. A. (1988) "Fidelity of HIV-1 reverse transcriptase" *Science* 242 **4882** 1168-1171.

[85] Purves, W. K., Sadava, D., Orians, G. H., Heller, H. C. (2001) *Life: The Science of Biology, Sixth Edition.* Vancouver: W. H. Freeman and Company.

[86] Quan, Y., Brenner, B. G., Dascal, A. and Wainberg, M. A. (2008) "Highly diversified multiply drug-resistant HIV-1 quasispecies in PBMCs : a case report" *Retrovirology* 5 **43** 1-6.

[87] RCSB Protein Data Bank (2009) Viewed 11th February 2009, <http://www.rcsb.org/pdb/home/home.do>.

[88] Reid, D. G., MacLachlan, L. K., Edwards, A. J., Hubbard, J. A. and Sweeney, P. J. (1998) "Introduction to the NMR of Proteins" *Methods in Molecular Biology*, Volume 60 methods in Molecular Biology.

[89] Resch, W., Parkin, N., Watkins, T., Harris, J. and Swanstrom, R. (2005) "Evolution of Human Immunodeficiency Virus Type 1 Protease Genotypes and Phenotypes In Vivo under Selective Pressure of the Protease Inhibitor Ritonavir" *Journal of Virology* 79 **16** 10638-10649.

[90] Roberts, J. D., Bebenek, K. and Kunkel, T. A. (1988) "The Accuracy of Reverse Transcriptase from HIV-1" *Science* 242 **4882** 1171-1173.

[91] Roitt,I., Brostoff, J., Male, D. (2002) *Immunology, Sixth edition.* London: Mosby.

[92] Rosso, L., Abrams, J. B. and Tuckerman, M. E. (2005) "Mapping the Backbone Dihedral Free-Energy Surfaces in Small Peptides in Solution Using Adiabatic Free-Energy Dynamics" *Journal of Physical Chemistry B* 109 **9** 4162-4167.

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

[94] Sadiq, S. K. (2008) "Molecular dynamics simulation studies of drug resistance in HIV-1 protease" *PhD thesis* U.C.L.

[95] Sadiq, S. K., Mazzeo, M. D., Zasada, S. J., Manos, S., Stoica, I., Gale, C. V., Watson, S. J., Kellam, P., Brew, S. and Coveney, P. V. (2008) "Patient-specific simulation as a basis for clinical decision-making" *Philosophical Transactions of the Royal Society* **366** 3199-3219.

[96] Sadiq, S. K., Wan, S. and Coveney, P. V. (2007) "Insights into a Mutation-Assisted Lateral Drug Escape Mechanism from the HIV-1 Protease Active Site" *Biochemistry* 46 **51** 14865-14877.

[97] Sadiq, S. K., Wright, D., Watson, S. J., Zasada, S. J., Stoica, I. and Coveney, P. (2008) "Automated molecular simulation based binding affinity calculator for ligand-bound HIV-1 protease." *Journal of Chemical Information and Modeling* 48 **9** 1909-1919.

[98] Schwartzl, S. M., Tschopp, T. B., Smith, J. C. and Fischer, S. (2002) "Can the Calculation of Ligand Binding Free Energies Be Improved with Continuum Solvent Electrostatics and an Ideal-Gas Entropy Correction?" *Journal of Computational Chemistry* 23 **12** 1143-1149.

[99] Scott, W. R. and Schiffer, C. A. (2000) "Curling of flap tips in HIV-1 protease as a mechanism for substrate entry and tolerance of drug resistance." *Structure* 8 **12** 1259-1265.

[100] Seibold, S. A. and Cukier, R. I. (2007) "A molecular dynamics study comparing a wild-type with a multiple drug resistant HIV protease: Differences in flap and aspartate 25 cavity dimensions" *Proteins: Structure, Function and Bioinformatics* **69** 551-565.

[101] Serdyuk, I. N., Zaccai, N. R., Zaccai, J. and Zaccai G. (2007) *Methods in Molecular Biophysics* Cambridge: Cambridge University Press.

[102] Shenderovich, M. D., Kagan, R. M., Heseltine, P. N. R. and Ramnarayan, K. (2003) "Structure-based phenotyping predicts HIV-1 protease inhibitor resistance" *Protein Science* **12** 1706-1718.

[103] Shuman, C. F., Hämäläinen and Danielson, U. H. (2004) "Kinetic and thermodynamic characterization of HIV-1 protease inhibitors" *Journal of Molecular Recognition* **17** 106-119.

[104] Stalking the AIDS Virus: A Killer With Many Faces (2008) Viewed 10th April 2009, <http://www.lanl.gov/quarterly/q_fall03/stalking_aids.shtml>.

[105] Stanford University HIV Drug Resistance Database (2007) Viewed 6th September 2007, <http://hivdb.stanford.edu/>

[106] Stoica, I., Sadiq, S. K. and Coveney, P. V. (2008) "Rapid and Accurate Prediction of Binding Free Energies for Saquinavir-Bound HIV-1 Proteases" *Journal of the American Chemical Society* 130 **8** 2639-2648.

[107] Stryer, L. (2000) *Biochemistry, Fourth Edition.* New York: W. H. Freeman and Company.

[108] Suñé, C., Brennan, L., Stover, D. R. and Klimkait, T. (2003) "Effect of polymorphisms on the replicative capacity of protease inhibitor-resistant HIV-1 variants under drug pressure" *Clinical Microbiology and Infection* 10 **2** 119-126.

[109] Tang, C., Louis, J. M., Aniana, A., Suh, J-Y. and Clore, G. M. (2008) "Visualizing transient events in amino-terminal autoprocessing of HIV-1 protease" *Nature Letters* **455** 693-696.

[110] Taubenberger, J. K., Reid, A. H. and Fanning, T. G. (2000) "The 1918 Influenza Virus: A Killer Comes into View" *Virology* **274** 241-245.

[111] Theoretical and Computational Biophysics Group: NAMD - Scalable Molecular Dynamics (2007) Viewed 10th April 2007, <http://www.ks.uiuc.edu/Research/namd/>

[112] Tie, Y., Boross, P. I., Wang, Y-F., Gaddis, L., Liu, F., Chen, X., Tozser, J., Harrison, R. W. and Weber, I. T. (2005) "Molecular basis for substrate recognition and drug resistance from 1.1 to 1.6Åresolution crystal structures of HIV-1 protease mutants with substrate analogs" *Federation of the Societies of Biochemistry and Molecular Biology Journal* 272 **20** 5265-5277.

[113] Tóth, G. and Borics, A. (2006) "Flap opening mechanism of HIV-1 protease" *Journal of Molecular Graphics and Modelling* **24** 465-474.

[114] Trylska, J., Grochowski, P. and McCammon, J. A. (2004) "The role of hydrogen bonding in the enzymatic reaction catalyzed by HIV-1 protease" *Protein Science* **13** 513-528.

[115] United Nations 2008 Report on the Global AIDS Epidemic.

[116] U.S. Food and Drug Administration (2007) Viewed 4h September 2007, <http://www.fda.gov>

[117] Vega, S., kang, L-W., Velazquez-Campoy, A., Kiso, Y., Amzel, L. M. and Freire, E. (2004) "A Structural and Thermodynamic Escape Mechanism from a Drug Resistant Mutation of the HIV-1 Protease" *PROTEINS: Structure, Function, and Bioinformatics* **55** 594-602.

[118] Wagner, E. K. and Hewlett, M. J. (1999) *Basic Virology.* Massachusetts: Blackwell Science, Inc.

[119] Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M. and Shafer, R. W. (2009) "Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance" *Genome Research* **17** 1195-1201.

[120] Wang, W. and Kollman, P. A. (2001) "Computational study of protein specificity: The molecular basis of HIV-1 protease drug resistance" *Biophysics* 98 **26** 14937-14942.

[121] Wang, W., Donini, O., Reyes, C. M. and Kollman, P. A. (2001) "BIOMOLECULAR SYSTEMS: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions" *Annual Review of Biophysics and Biomolecular Structure* **30** 211-243.

[122] Wartha, F., Horn, A. H. C., Meiselbach, H. and Heinrich, S. (2005) "Molecular Dynamics Simulations of HIV-1 Protease Suggest Different Mechanisms Contributing to Drug Resistance" *Journal of Chemical Theory and Computation* 1 **2** 315-324.

[123] Westhead, D. R., Parish, J. H. and Twyman, R. M. (2002) *Instant Notes in Bioinformatics*. Oxford: BIOS Scientific Publishers Ltd.

[124] Wittayanarakul, K., Aruksakunwong, O., Saen-oon, S., Chantratita, W., Parasuk, V., Sompornpisut, P. and Hannongbua S. (2005) "Insights into Saquinavir Resistance in the G48V HIV-1 Protease: Quantum Calculations and Molecular Dynamic Simulations" *Biophysical Journal* 88 **2** 867-879.

[125] Wittayanarakul, K., Hannongbua, S. and Feig, M. (2007) "Accurate Prediction of Protonation State as a Prerequisite for Reliable MM-PB(GB)SA Binding Free Energy Calculations of HIV-1 Protease Inhibitors" *Journal of Computational Chemistry* 29 **5** 673-685.

[126] Wlodawer, A. (2002) "Structure-based design of AIDS drugs and the development of resistance" *Vox Sanguinis Proceedings Paper* 83 **1** 023-026.

[127] Wlodawer, A. and Vondrasek, J. (1998) "Inhibitors of HIV-1 protease: A Major success of structure-assisted drug design" *Annual Review of Biophysics and Biomolecular Structure* **27** 249-284.

[128] Wolfenden, R. (1999) "Conformational Aspects of Inhibitor Design: Enzyme-Substrate Interactions in the Transition State" *Bioorganic & Medicinal Chemistry* **7** 647-652.

[129] Wood, A. J. J. (2009) "HIV-Protease Inhibitors: Review Article" *New England Journal of Medicine* 338 **18** 1281-1292.

[130] Wüthrich, K. (2002) "NMR Studies of Structure and Function of Biological Macromolecules" *Nobel Lecture, December 8 2002.*

[131] Zhang, Z. and Wriggers, W. (2008) "Coarse-Grain Protein Structures With Local Multivariate Features From Molecular Dynamics" *Journal of Physical Chemistry* **112** 14026-4035.

Appendix A

Published works