UNIVERSITY COLLEGE LONDON

Molecular Dynamics Simulation of Drug Resistance in HIV-1 Protease and Reverse Transcriptase

by

David William Wright

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Department of Chemistry
&
Centre for Mathematics and Physics in the Life Sciences and EXperimental
Biology (CoMPLEX)

September 2011

Declaration of Authorship

I, David William Wright, declare that this thesis titled, 'Molecular Dynamics Simulation of Drug Resistance in HIV-1 Protease and Reverse Transcriptase' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: ____

“By convention hot, by convention cold, but in reality atoms and void”

Democritus

UNIVERSITY COLLEGE LONDON

Abstract

Department of Chemistry
&
CoMPLEX

Doctor of Philosophy

by David William Wright

The emergence of drug resistant strains of HIV represents a major challenge in the treatment of patients who contract the virus. We investigate the use of classical molecular dynamics to give quantitative and qualitative molecular insight into the causes of resistance in the two main drug targets in HIV, protease and reverse transcriptase.

We initially establish a simulation and free energy analysis protocol for the study of resistance in protease. Focussing on the binding of the inhibitor lopinavir to a series of six mutants with increasing resistance we demonstrate that ensemble simulations exhibit significantly enhanced thermodynamic sampling over single long simulations. We achieve accurate and converged relative binding free energies, reproducible to within 0.5 kcal mol^{-1}. The experimentally derived ranking of the systems is reproduced with a correlation coefficient of 0.89 and a mean relative deviation from experiment of 0.9 kcal mol^{-1}.

Our protocol is then applied to investigate a patient derived viral sequence for which contradictory resistance assessments for lopinavir were obtained from existing clinical decision support systems (CDSS). Mutations at only three locations (L10I, A71I/V and L90M) influenced the ranking. Free energies were computed for HXB2 wildtype sequences incorporating each mutation individually and all possible combinations, along with the full patient sequence. Only in the case of the patient sequence was any resistance observed. This observation suggests an explanation for the discordance found using the CDSS. The effects on drug binding of the mutations at positions 10, 71 and 90 appear to be highly dependent on the background mutations present in the remainder of the sequence.

In preparation for the extension of our simulation and free energy protocol to reverse transcriptase the impact of binding both natural DNA substrates and two non nucleoside reverse transcriptase inhibitor (NNRTI) class drugs on the dynamics of reverse transcriptase are investigated. Free energies of both inhibitors (efavirenz and neviripine) are determined which are seen to be independent of the subdomain motions of the protein observed during simulation. Preliminary calculations of the free energies for a set of NNRTI resistant mutants bound to efavirenz are also presented.

Acknowledgements

This thesis would not have been possible without the inspiration, wisdom and assistance provided by some many people and it is a pleasure to show my appreciation here. Firstly I would like to thank my family and especially my parents Clifford and Brenda Wright and my sister Helen whose love and encouragement have been constant and much appreciated. I also wish to express my particular gratitude to my late Nana Beck for her support for my education.

Further thanks are due to the friends who have kept me sane and made my time in London so enjoyable. In no particular order I offer my heartfelt gratitude to Ben Scheck, Sibylle Duell, Matt and Steph Sinclair, Gautham Bhatnager, Ellie Murch and James Baker. Special thanks goes to Chris and Gemma Hook for their kindness, generosity and friendship.

I offer my sincerest gratitude to my supervisor, Peter Coveney, who has supported me throughout my thesis whilst allowing me the room to work in my own way. I am also indebted to many of my colleagues in the Centre for Computational Science for their support, guidance and willingness to listen to be rant. In particular I am grateful to Kashif Sadiq, Hugh Martin, Owain Kenway and Tom Hewer. I would also like to thank Shunzhou Wan, Ileana Stoica and Simon Clifford. I would also like to express my gratitude to all at CoMPLEX who have made my Masters and PhD experience as enjoyable and rewarding as it has been. In particular I would like to thank Alex Stewart, Lisa Willis and Samantha Lee for their friendship, insight and understanding.

I would like to thank our collaborators Paul Kellam, Simon Watson and Deenan Pillay at the Centre for Virology, UCL. Thanks also to Mattias Assel, Annemie Vandamme, Gokhan Ertaylan and David van der Vijver from the Virolab project.

I am indebted to CoMPLEX for paying my stipend and the EU funded ViroLab project for additional assistance.

Contents

List of Figures

List of Tables

Abbreviations

DNA	**D**eoxyribo**n**ucleic **A**cid
RNA	**R**ibo**n**ucleic **A**cid
PDB	**P**rotein **D**ata **B**ank
MD	**M**olecular **D**ynamics
HIV	**H**uman **I**mmunodeficiency **V**irus
AIDS	**A**cquired **I**mmuno**d**eficiency **S**yndrome
NMR	**N**uclear **M**agnetic **R**esonance
PME	**P**article **M**esh **E**wald
SMD	**S**teered **M**olecular **D**ynamics
CPU	**C**entral **P**rocessing **U**nit
GPU	**G**raphical **P**rocessing **U**nit
ITC	**I**sothermal **T**itration **C**alorimetry
EGFR	**E**pidermal **G**rowth **F**actor **R**eceptor
MMPBSA	**M**olecular **M**echanics **P**oisson-**B**oltzmann **S**urface **A**rea
SASA	**S**olvent **A**ccessible **S**urface **A**rea
DDM	**D**ifference **D**istance **M**atrices
DM	**D**ifference **M**atrices
PR	**P**rotease
RT	**R**everse **T**ranscriptase
PI	**P**rotease **I**nhibitor
NRTI	**N**ucleoside/nucleotide analogue **RT** **I**nhibitor
NNRTI	**N**on-**N**ucleoside/nucleotide analogue **RT** **I**nhibitor
NNRTIBP	**NNRTI** **B**inding **P**ocket

Physical Constants

Boltzmann Constant	k_B	$=$	3.2976268×10^{21} kcal K^{-1}
Plank Constant	h	$=$	1.582611×10^{-37} kcal s
Gas Constant	R	$=$	1.9858775×10^{-3} kcal K^{-1} mol^{-1}
Speed of Light	c	$=$	299792458 m s^{-1}
Permittivity of Free Space	ϵ_0	$=$	$1/(4\pi c^2) \times 10^7$ F m^{-1}

Symbols

T	Temperature	K
U	Internal Energy	kcal mol^{-1}
H	Enthalpy	kcal mol^{-1}
S	Entropy	kcal mol^{-1} K^{-1}
G	Gibbs Free Energy	kcal mol^{-1}
A	Helmholtz Free Energy	kcal mol^{-1}
p	Pressure	bar
N	Number	
$\mathscr{L}$	Lagrangian	kcal mol^{-1}
$\mathscr{H}$	Hamiltonian	kcal mol^{-1}
$\mathscr{K}$	Kinetic Energy	kcal mol^{-1}
$\mathscr{V}$	Potential Energy	kcal mol^{-1}
K_a	Association Constant / Binding Affinity	M^{-1}
K_d	Dissociation Constant	M
K_d	Dissociation Constant (Protein - Inhibitor)	M
K_M	Michaelis Constant	M

Dedicated to my parents and sister and to the memory of my godson, Sam

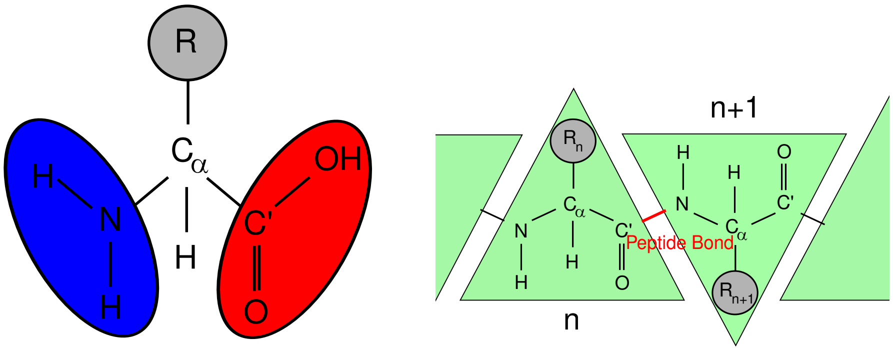
Chapter 1

Proteins

Proteins are a diverse class of macromolecules which form the majority of the dry mass of all cells and are responsible for the structure and functioning of all biological systems. The functions performed by proteins span every level of cellular processes and include the maintenance of cell shape, selective transport of small molecules, intra and intercellular messaging and the catalysis of chemical reactions. Despite their phenotypic variation all proteins share a common underlying construction, they are all polypeptide chains formed from a set of only 20 basic units. From these simple building blocks a huge diversity of structures can be formed, ranging in size from a few units to many thousands [1]. This variety of structure allows proteins to form the complex networks of interactions necessary for life.

Many biological processes involve the catalysis of chemical reactions. The class of proteins that performs the role of a catalyst are called enzymes. Nearly all processes in cells require enzymatic activity in order to occur at appreciable rates. Key to the ability of enzymes to perform this function in the crowded cellular environment is their ability to selectivity bind only their target substances. The study of the strength and specificity of enzyme binding has been central to the way we understand the operation of biological systems. The pharmaceutical industry exploits this property to allow the creation of inhibitory drugs, which are designed to interfere with the natural operation of target enzymes without causing harm by perturbing other cellular processes.

Detailed overviews of protein structure and function can be found in a range of standard molecular biology textbooks [2–4]. In this chapter the basic structure and function of proteins will be explored, although the description will be confined to the features which are observed and briefly noting how this may impact upon function. A number of the experimental techniques used to elucidate the structures are described in Chapter 2 and



(a) (b)

Figure 1.1: a) shows the structure of an amino acid. b) shows two amino acids linked by a peptide bond.

further details of protein-ligand interactions and experiments used to probe them are given in Chapter 3.

1.1 Peptide Chains and Primary Structure

The basic units which provide the monomers used to construct proteinaceous polypeptide chains are called amino acids. All amino acids share the same basic structure (see Figure 1.1a) in which a central carbon atom (designated C_α) is joined to a hydrogen atom (H), a carboxyl group (COOH) and an amino group (NH_2) as well as a side chain (R). It is the side chain which distinguishes one amino acid from another. In all amino acids (with the exception of glycine) all four groups attached to the C_α are different. This asymmetry means that amino acids are chiral molecules (i.e. it is not possible to superimpose one onto its mirror image, see Figure 1.2) and so can exist in one of two distinct forms whose properties are almost identical. The two forms are known as L- and D- form amino acids. Biological systems have evolved to almost exclusively use the L-form.

In the process of protein synthesis individual amino acids are joined together by the formation of peptide bonds (Figure 1.1b). A peptide bond is formed when the carboxyl group of one amino acid is involved in a condensation reaction with the amino group of another. The repetition of this process results in a chain with a repeated backbone joined together by peptide bonds from which the side chains project. The termini of this chain are the same as those in an individual amino acid. The chain is said to run from its amino (or N) terminus to its carboxyl (or C) terminus [3].

In living organisms, the sequence of amino acids which form a specific protein is provided by an RNA template (which is transcribed from the DNA genome). The RNA template

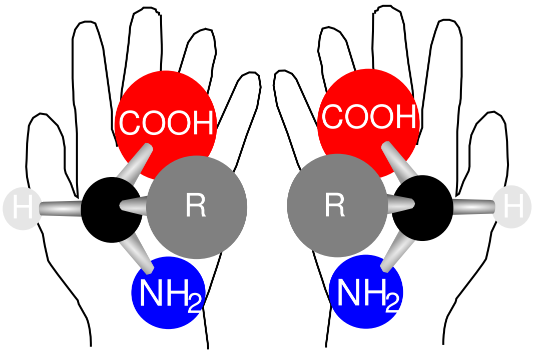
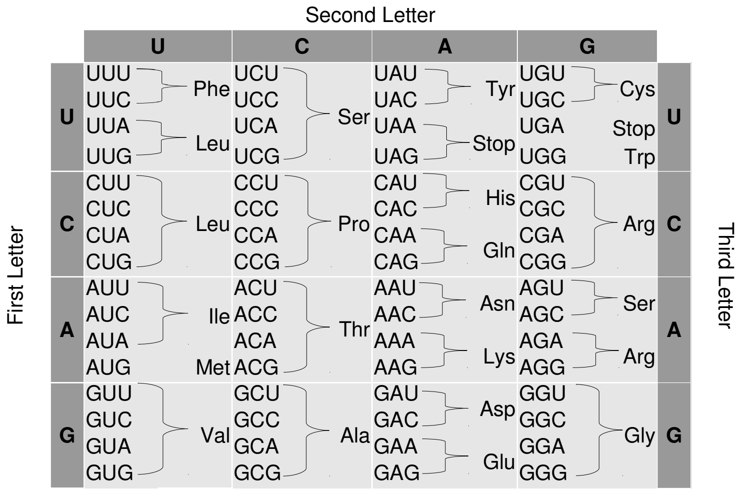


Figure 1.2: The stereo isomers of amino acids. The the L-form is shown on the left and the D-form on the right.



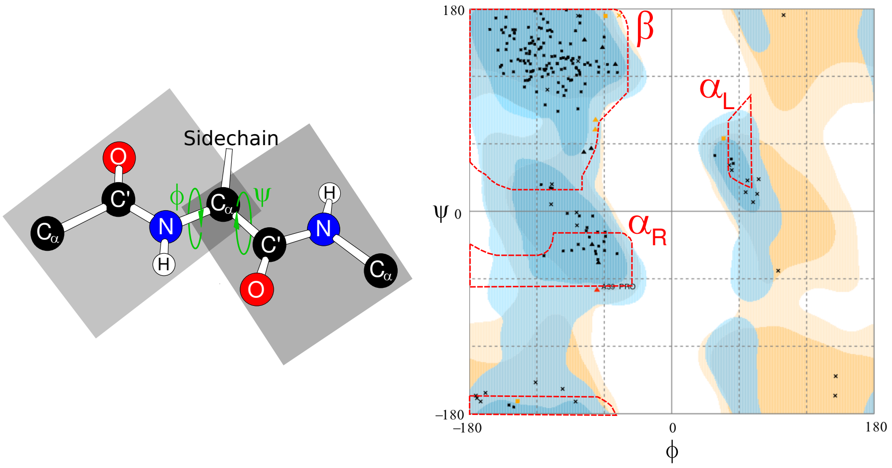
	Second Letter								
First Letter	U		C		A		G		Third Letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	
	UUG		UCG		UAG		UGG	Trp	
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	C
	CUC		CCC		CAC		CGC		
	CUA		CCA		CAA	Gln	CGA		
	CUG		CCG		CAG		CGG		
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	A
	AUC		ACC		AAC		AGC		
	AUA		ACA		AAA	Lys	AGA	Arg	
	AUG	Met	ACG		AAG		AGG		
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	G
	GUC		GCC		GAC		GGC		
	GUA		GCA		GAA	Glu	GGA		
	GUG		GCG		GAG		GGG		

Figure 1.3: The RNA code which specifies which amino acids to be included in a protein uses four bases (U, A, C and G) to make up one codon. There are thus 64 possible codons which code for 20 amino acids, leading to a high level of redundancy. The direction of the mRNA is 5′ to 3′. The codon AUG not only codes for methionine but also acts as an initiation site: protein translation begins at the first AUG in an RNA coding region[5].

codes for 20 different amino acids using a three base code (each base can be one of four different options (A)denine, (C)ytosine, (G)uanine or (U)racil) [2, 5]. The code is degenerate and is shown in Figure 1.3. The 20 naturally occurring amino acids have been given both single and three letter codes. Figure 1.4 shows all of these amino acids (with both forms of code) categorised according to the chemical properties of their side chains. Proteins do not, in general, take the form of simple extended chains but form complex structures. The process by which the polypeptide chain adopts this conformation is known as ‘folding’. The sequence in which the amino acids occur plays a decisive role in determining the eventual structure of the protein and for this reason it is known as the primary structure.

When thinking of the three dimensional structure of proteins a more helpful grouping

Figure 1.4: The 20 naturally occurring amino acids, grouped according to the properties of their side chains. This figure has been removed due to copyright restrictions but is available in Mat [6]



(a) (b)

Figure 1.5: The shaded area in a) indicates a peptide unit. The rotational angles ϕ and ψ are labelled within the unit. b) A Ramachandran plot showing the "allowed" regions of ψ/ϕ space. The regions labelled α_R and α_R correspond to those angles permitted in right and left handed α helices respectively. The region labelled β contains the angles associated with β sheets.

than the amino acid is the peptide unit. A peptide unit is defined as running from one C_α to the next (see Figure 1.5a). This means that all C_αs, except the first and last, belong to two such groups. These units exclude the amino acid sidechains and can be well described as rigid groups. As such they enjoy two degrees of freedom: they can rotate around the N-C_α or the C_α-C$'$ bonds. The rotational angles around these bonds are conventionally known as phi (ϕ) and psi (ψ) respectively.

Due to the steric constraints imposed upon them by the attached sidechains the areas of ϕ/ψ conformational space which are accessible by the protein backbone is limited. For most units (except those involving glycine, whose short sidechain is much less restrictive than any other) this limits the possible conformations to those areas shown as shaded in Figure 1.5b. This type of plot is known as a Ramachandran plot [7] after the biophysicist who first calculated the sterically allowed regions.

In common with the backbone, the side chains of the amino acids can, in general, adopt a variety of conformations known as rotamers. The sidechains will preferentially adopt the rotamers with the lowest energy. Which rotamer of a given amino acid has the lowest energy will depend upon its environment and in particular the conformation of the backbone at the position where it is attached as well as the position and rotameric state of other side chains which interact with it.

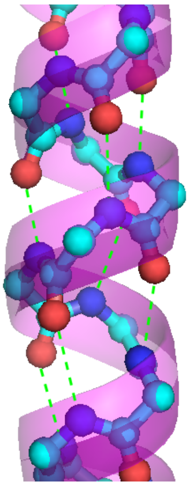


Figure 1.6: Hydrogen bonding in α helices. The protein backbone is shown in CPK representation with hydrogen bonds as dashed green lines.

1.2 Secondary Structure

The creation of peptide bonds in the polypeptide chain results in partial electron delocalisation, leaving each peptide unit capable of forming two hydrogen bonds. The networks of bonds formed can organise regions of the peptide chain into structural elements in proteins, the two most common are called α helices and β sheets [3, 8]. These structural elements correspond to the two "allowed" regions on the left of the Ramachandran plot in Figure 1.5b and are said to form the secondary structure of proteins.

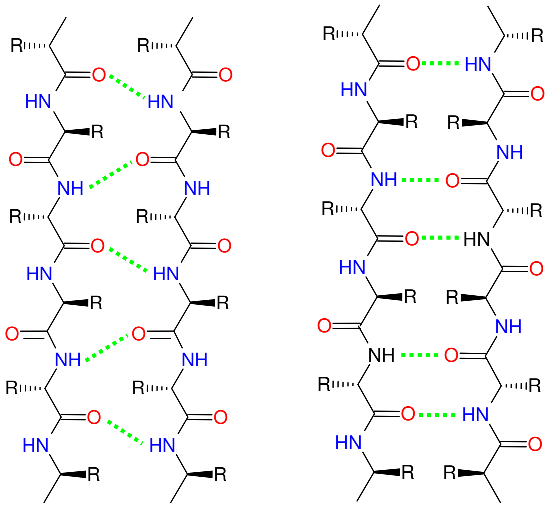
1.2.1 α Helices

α helices form when the ψ, ϕ angle pair of consecutive residues are approximately -60° and -50°. This creates a helical structure with approximately 3.5 residues per turn, with the nth residues C$'$=O hydrogen bonding to the amino group of the (n+4)th (see Figure 1.6). Hence, all but the terminal NH and C$'$O groups are involved in hydrogen bond formation. As a result, the ends of α helices are polar and consequently they are most frequently found on the surface of proteins.

α helices contain between 4 and 40 residues (averaging approximately 10) with each additional residue extending the helix by 1.5 Å along the helical axis.

1.2.2 β Sheets

Unlike the α helix, which is formed of one continuous region, β sheets form from a series of adjacent strands separated by turn regions. These strands are usually around five



(a) (b)

Figure 1.7: Hydrogen bonding in (a) parallel and (b) antiparallel β sheets.

Figure 1.8: Secondary structural elements combine together to form the tertiary structure. This figure has been removed due to copyright restrictions but is available from http://www.press.uillinois.edu/epub/books/brown/ch6.html

residues long and can either run parallel or anti-parallel to one another (Figure 1.7). The residues in the strands adopt an extended conformation which allows the adjacent C$'$O and NH groups to hydrogen bond. The ψ, ϕ values of the constituent amino acids are contained within the large range shown in the top left of Figure 1.5b.

1.3 Tertiary Structure

These secondary structure elements are joined together by regions called loops. Loop regions rarely contain hydrogen bonds between residues but often hydrogen bond with surrounding water molecules. The lack of internal bonding results in these regions being much less well ordered than the structural elements and consequently they exhibit greater flexibility.

The three dimensional arrangement of the various secondary structure elements and loops is known as the tertiary structure (Figure 1.8). The process by which the protein arrives at its final conformation is known as folding. The process of folding occurs on a timescale ranging from microseconds to milliseconds.

The folding of the protein is driven by the dispersion of hydrophilic and hydrophobic residues in water, the consequent packing of the hydrophobic residues within the molecule and non-covalent interactions between residue side chains. Some sections of a

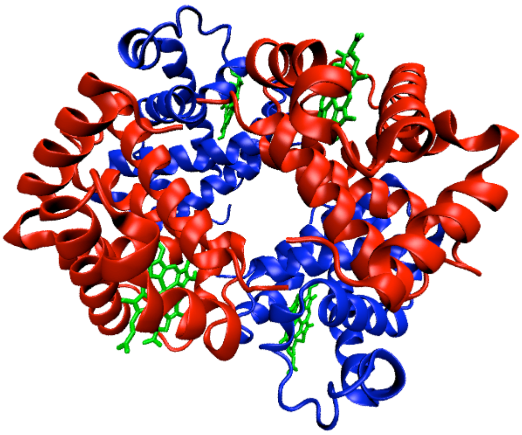


Figure 1.9: Structure of human haemoglobin (from the 1GZX PDB structure [10]) in cartoon representation, α and β subunits are shown in red and blue, respectively. The iron containing haem groups are in green.

polypeptide chain can independently fold into stable tertiary structures and are known as 'domains'. The tertiary structure of many proteins can be subdivided into functionally important domains linked by loop regions.

1.4 Quaternary Structure

Proteins frequently do not act alone and often form part of larger biological structures and complexes [3]. In this context, the individual protein chains are often referred to as 'subunits'. The conformation adopted by the subunits within the overall oligomeric protein is known as its quaternary structure. This higher level organisation is what allows proteins to perform their specific functions, providing sites for substrates to bind and creating the precise geometries required to catalyse specific reactions. One well known example of quaternary structure is that of haemoglobin, an enzyme key to oxygen transport in vertebrates [9], in which two α and two β chains combine to form a roughly tetrahedral assembly as shown in Figure 1.9.

The binding of either another protein or a small molecule in a location other than any active site can alter either the tertiary or quaternary structure of a protein (or complex of proteins). These changes underly the phenomenon of allosteric regulation in which they act to either increase or decrease activity [4]. Small molecule binding to regions other than the active sites of a complex can also impact upon protein function, for example in haemoglobin carbon dioxide binding to the α subunits alters the conformation of the complex and decreases its affinity for oxygen [11].

1.5 Sequence-Structure Relationships

It is often postulated that the native structure of a protein is solely dependent on its amino acid sequence, a proposal often attributed to Christan B. Anfinsen [12] and hence known as Anfinsen's dogma. This implies that for given conditions (temperature, pH, etc.) a unique minimum of the free energy exists for every protein and that it must be both stable and kinetically accessible. The process of folding and the ability to predict structure from protein sequence is one of considerable interest. If we assume that each bond connecting amino acids can have three possible states, a protein of, for example, 100 amino acids could exist in $3^{100} = 5 \times 10^{47}$ configurations. Even if new configurations are sampled at a rate of 10^{13} per second it would take 10^{27} years to try them all. This would mean that a sequential search would take longer that the age of the universe to arrive at the native structure of a protein, whereas protein folding generally occurs on a microsecond timescale. This problem was first considered by Levinthal [13] in 1968 and has become known as Levinthal's paradox. It is thus clear that proteins do not use this approach and that they follow specific pathways defined by their composition and the environment in which they fold.

Protein sequence data is much easier to generate experimentally than the related structure. Levinthal's paradox highlights the difficulty of the theoretical challenge of modelling the process of protein folding. This has resulted in the adoption of modelling techniques which use similar and possibly evolutionarily related sequences to help predict unknown protein structures. Some of the experimental techniques used to determine protein structures and methods of predicting those which are not amenable to these approaches are described in Chapter 2.

1.6 Dynamics and Function

In addition to the structure another factor which defines the way in which proteins function is their dynamics. Even at equilibrium proteins are not static structures and in many instances their inherent flexibility is necessary for them to perform their biological function, it is for example frequently important in the recognition of substrates. At the extreme end, some proteins are thought to be largely disordered until they bind a ligand that stabilises their structure [14]. Even small changes which result in only subtle alterations in structure can still have a marked effect on the dynamics of the system and have significant effects on the functioning of a protein. Such changes are of particular interest for the class of proteins responsible for catalysing biochemical reactions, known as enzymes. By enhancing the rates of chemical reactions between

10^6 and 10^{14} times, enzymes make possible the vast array of processes which are crucial in sustaining biological life [2, 3]. Enzymes are frequently classified by the types of reaction they catalyse. Common enzyme groups include oxidoreductases, which catalyse redox reactions, transferases, which catalyse reactions in which one chemical group is transferred between substrates, and hydrolases, which catalyse hydrolysis reactions.

In order to function in the densely packed environment of biological cells, enzymes must be highly selective about the substrates with which they interact. The specificity of enzymes is enhanced by the strong dependence of their efficiency on the pH and temperature of the local environment, which can allow them to be efficacious in some compartments or organelles of a cell and not others. The relationship between protein dynamics and their interactions with other chemicals will be further examined in Chapter 3.

The key role that enzymes play in the life cycles of biological entities has made them major targets for pharmaceutical treatments. In this context, there is major interest in the impact of subtle sequence changes on dynamic behaviour, and consequently substrate specificity, due to the emergence of drug resistant variants of proteins within etiological organisms. The changes in these proteins are often as small as the substitution of single amino acids, which result in almost undetectable alterations in the structure but which can dramatically alter the specificity of action of the protein. In many cases the impact of these changes not just on the binding of drugs, but also on how the target interacts with its natural substrate must be considered. The evaluation and prediction of selectivity hence represents a major challenge in molecular biology. In Chapter 4 drug resistance and the effect of mutations on selectivity will be discussed in detail for the case of two HIV viral proteins which are major targets for antiviral drugs.

Chapter 2

Protein Modelling and Molecular Dynamics

2.1 Atomistic Modelling Of Proteins

The importance of protein structure and function was discussed in Chapter 1 but how do we gain knowledge of these properties? Experimental techniques such as x-ray crystallography and NMR spectroscopy allow us to “see” protein structures and in the latter case even provide information on protein motions. However, in many cases protein structure and dynamics are not amenable to investigation using experimental techniques. In such cases *in silico* techniques such as homology modelling and molecular dynamics can provide insight which cannot be achieved without computational modelling. In this chapter we describe a variety of such modelling techniques, focusing on molecular dynamics.

All simulation methodologies require some experimental data to ground them in biological reality and so we begin our discussion with a brief review of the most commonly used techniques for determination of protein structure.

2.2 Structural Models From Experiments

The structures of proteins can be determined experimentally in two main ways, X-ray crystallography and nuclear magnetic resonance [3]. It is not, however, always possible to gain an experimental structure. In that eventuality it is sometimes possible to use a combination of the existing structures and knowledge of the protein sequence to construct a model of the unknown structure in a process known as homology modelling. This chapter gives a brief introduction to these three techniques.

2.2.1 Crystal Structures

The resolution of any image is determined by the wavelength of radiation used to produce it. In the case of protein structures we require atomic resolution, meaning that we wish to distinguish objects of approximately 1 Å (10^{-10} m), radiation of this wavelength is known as X-ray radiation. In practice X-rays with wavelengths between 0.4 Å to 1.6 Å are used to image proteins. The production of high quality images requires a regular array of objects to scatter the incoming x-rays. Hence, rather than illuminating proteins in a biologically relevant context it is necessary to crystallise them first.

2.2.2 X-ray Scattering

When X-rays are targeted at a protein crystal most will travel straight through, however some will encounter the electrons and nuclei of the target. When an X-ray photon encounters an electron it may be absorbed, increasing the vibrational energy of the electron. This vibrating electron then emits an X-ray photon of the same wavelength in a random direction. This process is known as coherent scattering, and is key to the X-ray crystallographic technique. More often, though, the X-ray will cause the electron to make orbital transitions which result in the emission of a photon of a different wavelength. This is known as incoherent scattering and can lead to radiation damage of the crystal. Luckily there are so many atoms ($\sim 10^{15}$) in a protein crystal that this is not too serious a problem. X-rays may also interact with nuclei in the sample, however their greater mass means that the scattering is negligible, resulting in X-ray imaging techniques only being able to “see” electrons [15].

Most of the rays scattered by a crystal sample will destructively interfere but some will constructively interfere and form a diffraction pattern which can be detected on a film or image plate. Analysis of this pattern using Bragg’s law allows the spacing of the diffraction peaks to be related to the spacing of the atoms within the illuminated crystal. Techniques which involve the addition of heavy atoms to the crystal structure are used to gain estimates of the phase of the scattered waves. The amplitudes and phases of the diffraction pattern are then input to computer software used to reconstruct a map of the electron density of the repeating unit of the crystal [3, 15].

2.2.3 Model Production

A model of the protein structure is produced by fitting the known sequence of the protein to the electron density map. The production of the initial model is a process of trial and error as there will be experimental uncertainties and errors in the electron map. In most cases there will be discontinuities in the map to which the polypeptide chain is being fitted. In general, a 5 Å resolution map can be used to obtain the shape of the protein, at 3 Å it is usually possible to trace the polypeptide chain and most amino acid sidechains. At 1 Å each atom is resolvable as a discrete ball of density [3].

The initial model is bound to contain some inaccuracies; these can be reduced by a process known as refinement. In this process the model is altered to minimise the difference between the experimental diffraction data and the equivalent information produced from a simulation using the hypothetical model structure. A measure called the R factor [16] (the R stands for residual disagreement) is used to express the quality of agreement. An R factor of 0.0 indicates perfect agreement and 0.6 comparison of a model with random reflections. For a large molecule (anything containing more that around 300 atoms) 0.2 would represent a well refined macro-molecular model at a resolution of 2.5 Å.

2.2.4 Model Quality

X-ray crystallography is highly accurate but presents many challenges when applied to proteins. These are mainly connected to the requirement of well ordered crystals. The better ordered the crystal the higher the resolution of the diffraction data and consequently the more accurate the model of protein structure. High quality protein crystals are hard to produce as proteins are often large, near spherical objects with irregular surfaces, which makes packing hard without leaving gaps and channels that become filled with disordered solvent. Further to this, proteins may exist in multiple conformations. Another difficulty is the fact that protein crystals may take months to grow and are highly sensitive to factors such as temperature, pH and enzyme concentration [15]. It should be noted that the extremely high protein concentrations and non-physiological conditions required to form crystals represent a significant potential source of error, which can produce considerable distortion of the target structure.

A frequently quoted measure of the positional error in a crystal structure is the B factor (also known as the temperature or Debeye-Waller factor). It is calculated for each atom during model refinement. Unfortunately, the calculation of this factor does not allow the discrimination between thermal motion, genuine structural flexibility or modelling error [3, 15].

2.2.5 Nuclear Magnetic Resonance

Nuclear magnetic resonance (NMR) is a technique in which the intrinsic magnetic moment of the ^{1}H, ^{13}C, ^{15}N or ^{31}P nuclei are used to probe their chemical surroundings [3, 17]. Large magnetic fields are used to align the nuclear spins of the atoms in a sample. Then the atoms are exposed to radio frequency pulses which move them into an excited state. When the atom reverts to its equilibrium position it emits radio frequency radiation. The precise frequency of the emitted signal depends both on the particular atom type and its environment. This resonant frequency is compared to a reference signal, the shift in the signal is called the chemical shift and is measured in parts per million (ppm). By varying the frequency of radiation to which the sample is exposed different properties can be probed.

In terms of three dimensional protein structures, the most important types of probes are called correlation spectroscopy (COSY) and nuclear Overhauser effect (NOE) experiments [3]. These give information on ^{1}H atoms which are covalently connected through one or two other atoms (i.e. they are very close in the protein sequence) and atoms which are close in space irrelevant of where they occur in the sequence respectively. Combining information from these two protocols with knowledge of the protein sequence allows distances between atoms to be computed. The distances between atoms can be used to create constraints on the atomic positions, which can then be used to compute the three dimensional dimensional protein structure. Usually this process results not in a single structure but a selection of structures all of which equally well satisfy the constraints, as the problem is under determined. This means that it is hard to quantify the accuracy of protein structures determined by NMR experiments but has the advantage that ensembles of structures representative of genuine conformational flexibility of the protein can be produced [18].

Samples for NMR are usually dissolved in 0.5 ml of water in a setup which allows the temperature and pH to be much closer to physiological conditions than is possible with crystallography. Another advantage of NMR is that there are no crystal packing effects. However, the concentration of protein needed for good result is of the order of 5 mM or higher, which is much greater than that of most proteins *in vivo* although it is comparable with the total enzymatic concentration in many bodily fluids. The biggest drawback to NMR is that it can only be used on small proteins (in general only up to 32 kDa) [3, 17].

2.2.6 Homology Modelling

While the number of proteins with a structure in the PDB[1] continues to grow, it remains the case that the structure of the vast majority of proteins in existence remains unknown. Many of these proteins may never have an experimental structure as they are too large for NMR analysis and cannot be crystallised. Although the sequence of a protein plays a key role in defining its structure, as discussed in Section 1.5, we cannot simply explore all of the astronomical number of possible conformations it could adopt. Attempts have been made to use direct simulation (such as molecular dynamics) to perform a biased search but this approach remains too computationally expensive.

One approach that has been applied to help gain insight into unknown structures is homology modelling. This type of modelling based on the observation that whilst the structure of a protein is determined by it's amino acid sequence, the structure is more stable, changing more slowly than the associated sequence. The assumption made in this type of modelling is that similar sequences will fold almost identically and that even more distantly related sequences will retain a high level of structural similarity. A comparison between the sequence of a protein of unknown structure and those of known structures can thus be used to predict the unknown structure. The implementation of this approach has been described as a seven step process [19]:

1. Template identification and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modelling
5. Sidechain modelling
6. Model optimisation
7. Model validation

What follows is a brief description of what is involved in each of these stages.

2.2.7 Template Identification

The first stage of homology modelling is to identify a suitable template structure upon which to base the prediction of the protein structure of interest. Templates are usually

[1] PDB: `http://ww.pdb.org`

acquired by performing a search using standard sequence comparison tools (such as BLAST [20] or FASTA [21]) against all of the structures in a database such as the PDB. These methods apply a scoring system for differences between sequences, with the substitution of chemically similar residues incurring a small penalty and insertions, deletions and substitutions of non-similar residues having larger penalties associated with them. The sequences with the lowest penalty scores are identified as possible templates. If areas of the template structure are poorly defined in one template then it is sometimes possible to use multiple structures and to take the most well defined areas from each [19].

2.2.8 Sequence Alignment and Correction

Sequence alignment is the process of matching the order of amino acids in one protein sequence against that of another [22]. Allowance has to be made for the fact that some mutations are conservative (i.e. the change in amino acid side chain only slightly alters the biochemical properties) but some represent a more significant alteration. Attempts to align a pair of sequences can be hard in regions of low identity. The alignment can in some cases be improved by using a third intermediate sequence which is more similar to the target in the low identity region. Programs such as CLUSTALW [23] can perform these multiple sequence alignments. This can be particularly useful when aligning areas where there are insertions or deletions between the sequences.

2.2.9 Backbone Generation

Once an alignment has been made, model building can begin. The first stage is to model the backbone (N, C_α, C and O) atoms. For most of the model the coordinates can simply be copied directly from the template [19]. Rigid side chains are also often copied at this stage.

Modelling with multiple templates can easily be achieved by using servers such as Swiss-Model [24] which perform the alignment between the two structures as well as performing the modelling stages described here. An experienced modeller may, however, find that automatic alignments can be improved by hand using software such as Deep View (previously known as Swiss-PdbViewer) [25] which allows the user to visualise and refine structural alignments.

2.2.10 Loop Modelling

Loop sections in models often contain insertions, deletions and mutations all of which can alter the backbone conformation. The effects of these changes are notoriously hard to predict [19]. One widely used approach is to search for loops within structures from the PDB which have endpoints that match the sequence being modelled and then copy the loop conformation. An alternative approach is to use fold prediction where an energy function is used to judge the quality of a given loop conformation. The energy function is then minimised using Monte Carlo or molecular dynamics techniques (see Section 2.3.3 and Section 2.4 respectively for more details on these techniques).

2.2.11 Sidechain Modelling

In areas of high sequence identity it is usually safe to simply copy the side chain conformation from the template [19]. In regions of lower sequence identity it is necessary to use a library of possible rotamers (low energy conformations are generated by rotating the sidechain around the bond to the backbone) [26, 27]. These libraries are built by identifying the rotamers which occur most often for particular backbone conformations. The backbone of the model is compared to those in the library and the best matches used to select which rotamer is incorporated into the model.

2.2.12 Model Optimisation

From the description above, it is clear that the predicted backbone position and sidechain rotamers are interdependent. Thus an iterative process is often adopted using the predicted side chains to update the backbone and this updated backbone to re-predict the sidechains and so on. One frequently used method of model optimisation is to run a molecular dynamics minimisation (see Section 2.4 for details of this approach) hoping that this will mimic the true folding process [19].

2.2.13 Model Validation

There are two main sources of error in the modelling process which are the quality of the template, and the level of similarity between the modelled sequence and that of the template. When there is a 90% or greater level of identity the model accuracy can reasonably be compared to crystallography (with a few exceptional sidechains), but once the identity drops below this level model quality is highly variable with large local deviations often being introduced [19, 28].

Checking the quality of models is usually done by examining the energy of the system using a molecular dynamics forcefield and examining the bond lengths, bond and torsion angles and distribution of polar and apolar residues to ensure that they are within normal bounds. A detailed verification of any model is an essential part of the modelling process.

2.3 Enhancing Structural Understanding Using Computational Modelling

Computational models are key parts of the experimental techniques used to interpret experimental data on protein structure but they can also be used to explore beyond this. Models derived from fundamental physical considerations can be used to refine structures, investigate the effects of protein environment, explore dynamics and enzyme chemistry in order to provide insights which cannot be obtained from experiments. Simulation techniques are available that range in detail level from those that describe the electrons which govern chemistry to those which focus on the long range interactions of proteins and ligands. Here we give a brief overview of a variety of techniques which have been used to investigate the proteins of HIV. Later, in Section 2.4, a more detailed account of the molecular dynamics methodology is provided, which will be used in the studies presented in this thesis.

2.3.1 Quantum Mechanics

The most fundamental description of the world available to us at the atomic level is quantum mechanics and it provides the only way we can realistically model chemical reactions in atomic detail. In principle protein structure and dynamics can be understood by using quantum mechanics. To take this approach would require the solution of the time dependent Schrödinger equation for the entire protein. In practice, even for very small systems, all that can be achieved is an approximation to the true solution.

The approximate methods used to solve Schrödinger equation are generally divided into two categories; those which include empirical parameters, known as semi-empirical methods, and those derived directly from theoretical principles, with no inclusion of experimental data, known as *ab initio* methods.

One, almost ubiquitously used, simplification employed is the Born-Oppenheimer approximation, which assumes that owing to their relatively large mass the nuclei of atoms are fixed with respect to the electrons. This reduces the problem to calculation of the, approximate, wavefunction of the electrons in the field of the fixed nuclei. This wavefunction can be used to calculate the forces on the nuclei, whose positions are updated using classical mechanics. The process can then be iterated with the electrons assumed to move instantaneously with the nuclei to continue the evolution of the system. Other approaches, such as Car-Parrinello method, are available which can calculate the coupled nuclear and electron motions at further computational cost.

The calculation of the wavefunction remains exceptionally computationally expensive even when further approximations such as Hartree-Fock or Density Functional Theory (in which the basic quantity calculated is the electronic density not the wavefunction) are applied, as is done in commonly used packages such as Gaussian 03 [29]. This means that it is not feasible to use this method to study the dynamics of entire proteins. Detailed descriptions of quantum mechanical simulation techniques can be found in the literature [30–32]

Despite this limitation, quantum mechanical simulation still plays an important role in biological simulations. It can be used to optimise or minimise small structures, investigate enzymatic reactions and, perhaps more importantly, quantum level simulations are often used, in conjunction with experimental data, to provide parameters (such as atomic charges derived from the electron density distribution) used in more coarse grained approaches such as molecular dynamics. Furthermore, it is also increasingly used in conjunction with molecular dynamics in what are regularly called quantum mechanical/molecular mechanical (QM/MM) hybrid models (see Section 2.3.4) [32, 33].

2.3.2 Molecular Dynamics

A very common method of atomistic simulation is molecular dynamics (MD), sometimes also known as molecular mechanics (MM). In this formalism the atoms of the system are modelled as points with a given mass and charge. The charges are used to calculate an electrostatic forcefield from which the force on each atom in the system can be calculated. The force is then used to update the positions of each atom using classical mechanics. This process is then iterated to evolve the system configuration. This approach allows us to gain information not only of the conformations explored by protein systems but also their dynamics [32, 33].

The reason for the widespread use of molecular dynamics is that it is much less computationally expensive than quantum mechanics whilst still allowing all of the atoms of a protein to be simulated. It has been found that neglecting the electrons does not, in general, prevent the method from generating realistic protein dynamics. The focus of this thesis is upon the use of classical MD and consequently the method is described in more detail in Section 2.4.

A variant of molecular dynamics in which small groups of atoms are represented by single interaction sites, known as coarse grained (CG) models, are becoming increasingly popular. In particular this methodology is often applied to study systems containing lipids [34, 35]. This allows the exploration of the dynamics of bigger systems and longer timescales using the same computational resources but at the expense of atomic resolution and hence physical fidelity.

2.3.3 Monte Carlo Simulation

Monte Carlo methods are a class of algorithms based upon the repeated sampling of random numbers [33]. In the context of protein simulations they provide a stochastic approach to explore the molecular level configurational space available to a protein at equilibrium. Unlike MD this class of simulation cannot provide dynamic information about the system of interest.

The underlying concept is to take a three dimensional protein structure and use this as the starting point for a random walk in conformational space [33, 36]. At each step along this walk the probability of a given change in conformation is dependent on the change in energy required from the previous state. In order to ensure thermodynamically correct sampling the probability of visiting a particular state $\mathbf{r}$ is proportional to the Boltzmann factor, $e^{\frac{-U(\mathbf{r})}{k_B T}}$.

The most commonly used method for choosing the next state is the Metropolis algorithm [36] where a move from state $\mathbf{r}$ to $\mathbf{r}'$ happens with probability:

$$P(\mathbf{r}'|\mathbf{r}) = min\left(1, e^{\frac{-1}{k_B T}}\left(U(\mathbf{r}') - U(\mathbf{r})\right)\right) \tag{2.1}$$

Monte Carlo simulations are frequently used in energy minimisation problems as well as to calculate thermodynamic properties of systems. In general, the difficulty in efficiently choosing conformational steps for the random walk results in high rejection rates, making Monte Carlo simulations less attractive for biomolecular systems than molecular dynamics.

2.3.4 Quantum Mechanical/Molecular Mechanical Hybrid Models

While molecular dynamics can probe many of the properties of biological systems, the inability to treat important chemical events such as bond breaking (lysis) and the move into transition states during reactions is a considerable limitation. A full QM treatment is, as has been mentioned, usually impractical. This has led to the development of the hybrid QM/MM methodology in which most of the system is treated by molecular dynamics but a critical segment is described at the quantum level. The main challenge faced in this form of modelling is how to handle the transition between the two scales of model [37]. Considerable progress has been made using several methods to bridge the divide including link atoms and local self-consistent field formulations [38, 39].

2.3.5 Brownian Dynamics

Protein-protein and protein-ligand association involves processes on different length and time scales. At large distances, only the relative motion of the two centres of mass is important. In this regime, the system evolution can be sampled using Brownian dynamics. The Brownian dynamics simulation technique is a mesoscopic method in which explicit solvent molecules are replaced by a stochastic force [40]. The technique takes advantage of the large separation in time scales between the rapid motion of solvent molecules and the much slower motion of polymers or colloids. The elimination of the fast modes of the solvent the simulation of much longer time scales than can be explored in a molecular dynamics simulation. Brownian dynamics simulation is based on the integration forward in time of a stochastic differential equation in order to create molecular trajectories. The incorporation of time in the governing equation, coupled with the reduced computational demands compared to fully atomistic simulations, allows for the study of the temporal evolution and dynamics of complex fluids (such as polymers, large proteins, DNA molecules and colloidal solutions) [41]. Brownian dynamics can be viewed as a limit of Langevin dynamics, which will be discussed in Section 2.5.1 in the context of temperature control in MD. The main disadvantage of the approximations made in this approach is that the use of a random force independent of particle positions means that momentum is not conserved locally. A related technique known as dissipative particle dynamics (DPD) has been developed which solves this problem by incorporating a dissipative force [42, 43] but which has not been widely applied to HIV protein systems.

2.3.6 Network Models

Network models provide a minimalist, coarse grained, method for understanding protein motions. The prototypical form is the Gaussian network model (GNM) where the basic model is to take the C_α of the proteins as the nodes of a network and to model the connections between the nodes as harmonic springs (with spring constant γ). With two nodes being connected if they fall within a cut off distance r_c (usually between 7 and 10 Å) of one another in the three dimensional protein structure [44, 45]. Due to the simplistic nature of the model it is a computationally very inexpensive method. the protein structure is described as an elastic network of N nodes. This description is then encoded in an $N \times N$ matrix (where N is the number of residues in the protein), known as the Kirchoff matrix, Γ, with elements given by

$$\Gamma_{ij} = \begin{cases} -1 & \text{if } i \neq j \text{ and } R_{ij} \leq r_c \\ 0 & \text{if } i \neq j \text{ and } R_{ij} > r_c \\ -\sum_{i,i\neq j} \Gamma_{ij} & \text{if } i = j \end{cases}$$

where R_{ij} is the equilibrium distance between two atoms and r_c a cut off distance beyond which two atoms are deemed unconnected. The motion of the protein can also be decomposed using normal modes by diagonalising Γ. The slowest modes contribute most to the fluctuations of the protein and are often seen to represent collective motions of domains or other structural elements [46–48]. It is these modes which are though to be most likely to relate to protein function. It is also possible to calculate the expectation values for the fluctuations of individual residues from Γ.

2.4 Details of the Molecular Dynamics Methodology

The behaviour of proteins is inherently dynamic, so in order to understand them better we need to create models which capture their motion. Perhaps, the most popular method for modelling the interactions of the large numbers of atoms involved in such systems is molecular dynamics. Essentially the methodology of molecular dynamics is very simple. Initially all of the atoms in a system of interest are assigned coordinates, velocities and charges. The positions and charges are then used to calculate a potential. This potential is used to compute the force felt by each of the atoms in the simulation. By integrating Newton's laws of motion over a short time step a new set of positions and velocities is produced for each of the atoms. The updated values can now be fed back into the first step of the calculation and the process repeated, creating a trajectory of atomic locations and velocities through time. What follows is a short description of the steps described above; a more thorough treatment can be found in standard texts (such as Leach [32] or Frenkel & Smit [33]).

Despite the conceptual simplicity of molecular dynamics, the computational load required to achieve numerical stability of the integration schemes used remains high. The main factors in this are the need to use very small time steps (typically of the order of 1 fs) in order to capture the fastest dynamic processes in the system (in protein systems this is usually the vibration of hydrogen atoms) and the calculation of the force.

2.4.1 Equations of Motion

The mechanical state of an N particle system can be completely described by $3N$ generalised coordinates q_i (where i = 1,2,3…$3N$), $3N$ generalised velocities $\dot{q}_i$ and a potential energy function $\mathscr{V}(q_i)$. The potential energy function describes the interactions between the particles and is dependent only on the coordinates, q_i. Both the Lagrangian and Hamiltonian formalisms can be used to derive equations of motion from these constituents. The former is naturally associated with configuration space, extended by time, while the latter is the natural description for working in phase space. Here we present condensed derivations of the equations of motion in both formalisms; more detailed ones can be found in standard texts such as Landau & Lifshitz [49].

The derivation in the Lagrangian framework proceeds from Lagrange's equations,

$$\frac{d}{dt}\left(\frac{\partial \mathscr{L}}{\partial \dot{q}_i}\right) - \frac{\partial \mathscr{L}}{\partial q_i} = 0, \tag{2.2}$$

where the Lagrangian function $\mathscr{L}(q_i, \dot{q}_i)$ is defined as:

$$\mathscr{L}(q_i, \dot{q}_i) = \mathscr{K}(\dot{q}_i) - \mathscr{V}(q_i), \tag{2.3}$$

with $\mathscr{K}(\dot{q}_i)$ representing the kinetic and $\mathscr{V}(q_i)$ the potential energy of the system. In Cartesian coordinates, representing the position, velocity and acceleration vectors as $\mathbf{r}_i$, $\dot{\mathbf{r}}_i$ and $\ddot{\mathbf{r}}_i$ respectively, the kinetic energy is defined as:

$$\mathscr{K} = \frac{1}{2}\sum_{i=1}^{N} m_i \dot{\mathbf{r}}_i^2. \tag{2.4}$$

Substituting from equations Equation 2.3 and Equation 2.4 into Lagrange's equations allows the derivation of Newton's equations of motion in the form of $3N$ second order differential equations:

$$\mathbf{f}_j = m_j \ddot{\mathbf{r}}_j, \tag{2.5}$$

where $\mathbf{f}_j$ is the force on the j^{th} particle and is given by the spatial derivative of the potential function:

$$\mathbf{f}_j = -\nabla_{\mathbf{r}_j} \mathscr{V}. \tag{2.6}$$

We now proceed to give a simplified account of the steps involved in the derivation using the Hamiltonian formalism. The Hamiltonian, representing the total energy, of a system is defined as

$$\mathscr{H}(p_i, q_i) = \sum_{i=1}^{3N} p_i \dot{q}_i - \mathscr{L}, \tag{2.7}$$

where the generalised momentum, p_i, conjugate to q_i is defined as $p_i = \partial \mathscr{L} / \partial \dot{q}_i$. The total differential can be expressed in terms of the Hamiltonian such that:

$$d\mathscr{H} = -\sum_{i=1}^{3N} \dot{p}_i dq_i + \sum_{i=1}^{3N} \dot{q}_i dp_i. \tag{2.8}$$

From this we can derive Hamilton's equations of motion in the independent variables p_i and q_i:

$$\dot{q}_i = \frac{\partial \mathscr{H}}{\partial p_i}, \tag{2.9a}$$

$$\dot{p}_i = \frac{\partial \mathscr{H}}{\partial q_i}. \tag{2.9b}$$

Whereas the Lagrangian derivation produced $3N$ second order equations, here we have the equations of motion expressed as $6N$ first order equations. Both sets of equations describe the evolution of a many-body system in time and form the basis of the theory of classical molecular dynamics.

The classical equations of motion are deterministic and time-reversible and, consequently, a mechanical system is completely described by the positions and momenta of its N atoms. This means that we can conceive of the position and momentum as the coordinates of a single point in a $(6N + 1)$-dimensional phase space which represents the state of the system. The point moves through phase space according to the equations of motion derived in this section. The concept of phase space is very useful when relating the microscopic mechanics of a system to thermodynamic properties, a process we will discuss in Chapter 3.

2.4.2 Force Fields and the Potential Energy Function

In order to calculate the force felt by each atom it is first necessary to compute the potential energy function, $\mathscr{V}$. Although a precise calculation of the potential energy of a N atom system would have to consider the contribution of each individual atom, pair, triplet and so forth, most molecular dynamics programs (including the NAMD [50] package used to perform all simulations reported in this thesis, more details of available packages for performing MD are provided in Section 2.5.3) describe the potential energy using a more simplistic five component picture. In this scheme the potential energy has the following basic form:

$$\mathscr{V}_{total} = \mathscr{V}_{bonded} + \mathscr{V}_{non-bonded} \tag{2.10a}$$

$$\mathscr{V}_{bonded} = \mathscr{V}_{bond} + \mathscr{V}_{angle} + \mathscr{V}_{dihedral} \tag{2.10b}$$

$$\mathscr{V}_{non-bonded} = \mathscr{V}_{vdW} + \mathscr{V}_{Coulomb} \tag{2.10c}$$

The first three components can be described as representing the stretching, bending and torsional bonded interactions. These are usually represented in terms of the deviation of the bond length r, angle θ and dihedral angle ψ (see Figure 2.1) from a reference, or equilibrium value, see Equation 2.11.

$$\mathscr{V}_{bond} = \sum_{bonds} k_r(r - r_{eq}) \tag{2.11a}$$

$$\mathscr{V}_{angle} = \sum_{angles} k_\theta(\theta - \theta_{eq}) \tag{2.11b}$$

$$\mathscr{V}_{dihedral} = \sum_{dihedrals} \frac{\mathscr{V}_n}{2}(1 + \cos(n\phi - \gamma)) \tag{2.11c}$$

The last two terms represent the van der Waal's forces and the electrostatic interactions between the non-bonded atoms. The former is approximated as a Lennard-Jones 6-12 potential (Equation 2.12a), the later is given by the Coulomb potential (Equation 2.12b).

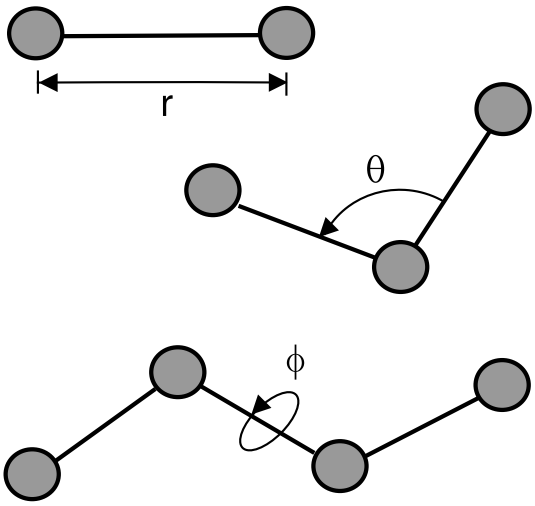


Figure 2.1: The coordinates used to describe bonded interactions in the interatomic interaction potential of an MD forcefield: r governs bond stretching, θ the bond angle and ϕ the dihedral angle between two atoms connected by three covalent bonds.

$$\mathscr{V}_{vdW} = \sum_{i}\sum_{j>i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right) - \left(\frac{\sigma_{ij}}{R_{ij}}\right)\right] \tag{2.12a}$$

$$\mathscr{V}_{Coulomb} = \sum_{i}\sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.12b}$$

The constants k_r, r_{eq}, k_θ, θ_{eq}, etc. are then taken from standard parameterisation schemes such as CHARMM [51] and AMBER [52]. The required parameters are determined by grouping combinations of atoms of varying types and fitting to either experimental or *ab initio* quantum mechanical calculations. This approach assumes that the parameters derived from these small subsets of atoms can provide a sufficiently accurate approximation of the properties of the same groupings of atoms embedded in a larger molecular structure. Forcefields may differ in their functional form and in the systems and physical conditions (such as temperature and pressure) for which they are parameterised. The simulations performed in this thesis use the AMBER values, which are parameterised to be suitable for studying proteins, nucleotides and lipid bilayers.

There are a number of limitations which are inherent in the use of any forcefield. The parameters in commonly used forcefields, such as AMBER, have been validated for equilibrium structures over short timescales but inaccuracies may arise when systems move away from equilibrium or for pressure or temperature conditions far from those used in their parameterisation. Furthermore, the number of atomic combinations used to create the parameter set are limited and whilst the configurations of all amino acids have been parameterised this is not true for many molecules of interest. An example, which is relevant to the simulations described in this thesis, is that of novel compounds used as drugs. The General AMBER forcefield (GAFF) attempts to extend the coverage to encompass the majority of organic and pharmaceutical compounds [53]. The additional parameters were primarily obtained through fitting results from *ab initio* calculations.

2.4.3 Updating the Atomic Positions

Once the force on, and consequent acceleration of, each particle are calculated (from Equation 2.6 and Equation 2.5) they can be used in conjunction with standard finite difference numerical integration methods to advance the atomic positions over a small time step. In molecular dynamics the main criteria which govern the choice of integration scheme are the requirement that energy be conserved and the need for computational efficiency. The Verlet class of integrators are the most commonly used methods employed by molecular dynamics codes. Other approaches are available but are used less frequently and we shall not discuss them here.

The Verlet method[54] begins by assuming that the positions $\mathbf{r}(t+\delta t)$ and velocities $\mathbf{v}(t+\delta t)$ can be approximated by Taylor expansions

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}\delta t + \frac{\mathbf{a}(t)}{2}\delta t^2 + \frac{\mathbf{b}(t)}{6}\delta t^3 + \dots, \tag{2.13a}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \mathbf{a}\delta t + \frac{\mathbf{b}(t)}{2}\delta t^2 + \dots \tag{2.13b}$$

where $\mathbf{a}(t)$ is the acceleration and $\mathbf{b}(t) = \dot{\mathbf{a}}(t)$. The equivalent expansion of $\mathbf{r}(t-\delta t)$ is

$$\mathbf{r}(t-\delta t) = \mathbf{r}(t) - \mathbf{v}\delta t + \frac{\mathbf{a}(t)}{2}\delta t^2 - \frac{\mathbf{b}(t)}{6}\delta t^3 + \dots \tag{2.14}$$

Adding or subtracting Equation 2.14 from Equation 2.13a and substituting for the acceleration using $\mathbf{F} = m\mathbf{a}$ then yields the following expressions for $\mathbf{r}(t+\delta t)$ and $\mathbf{v}(t)$

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + m^{-1}\mathbf{F}(t)\cdot\delta t^2, \tag{2.15a}$$

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\delta t) - \mathbf{r}(t-\delta t)}{2\delta t} \tag{2.15b}$$

This is the Verlet algorithm. It has many attractive properties as it is simple, time reversible and conserves energy. However, the expressions for position and velocities involve differences between large, similar numbers which leads to numerical inaccuracies. A variant of the method known as the velocity Verlet algorithm [55] removes these disadvantages (by replacing the subtractions of quantities by sums) and is consequently numerically preferable when using computers of finite precision. The expressions for the positions and velocities using this algorithm are

$$\mathbf{v}(t+\frac{\delta t}{2}) = \mathbf{v}(t) + m^{-1}\mathbf{F}(t)\cdot\frac{\delta t}{2}, \tag{2.16a}$$

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}(t+\frac{\delta t}{2})\delta t, \tag{2.16b}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t+\frac{\delta t}{2}) + m^{-1}\mathbf{F}(t+\delta t)\cdot\frac{\delta t}{2} \tag{2.16c}$$

It is this method that is employed by the NAMD code used to perform the simulations presented in this thesis.

The size of timestep required in order to describe the fastest motions in biochemical systems and maintain the stability of numerical integration is a major factor in the high computational expense of MD. Typically step sizes of 0.1-1 femtosecond are required. In Section 2.5.2.4 and Section 2.5.2.5 two methods used to reduce computational workload and increase the size of the timestep are described.

2.5 Interpreting Trajectories

In principle the trajectories of atomic positions (and velocities) can be used as the basis for a wide range of qualitative and quantitative assessments of protein dynamics and functions. However, an important issue that needs to be addressed when dealing with computer simulations is the robustness of results to perturbations in the initial conditions [32, 33]. Consider the time evolution of a system trajectory Γ, where $\Gamma = (\mathbf{p}, \mathbf{q})$, in a system defined by a Hamiltonian, $\mathscr{H}$, giving us Equation 2.17.

$$\frac{d\Gamma}{dt} = \nabla_{\Gamma}\mathscr{H}(\Gamma). \tag{2.17}$$

If the trajectory is slightly perturbed we then obtain Equation 2.18.

$$\frac{d(\Gamma+\delta\Gamma)}{dt} = \nabla_{\Gamma}\mathscr{H}(\Gamma+\delta\Gamma). \tag{2.18}$$

If the right hand side of Equation 2.18 is expanded using the first term of the Taylor expansion and then comparing terms we get Equation 2.19b.

$$\frac{d(\Gamma + \delta\Gamma)}{dt} \simeq \nabla_\Gamma \mathscr{H}(\Gamma) + \nabla_\Gamma \left(\nabla_\Gamma \mathscr{H}(\Gamma)\right) \delta\Gamma \tag{2.19a}$$

$$\frac{d\Gamma}{dt} = \nabla^2_\Gamma \mathscr{H}(\Gamma)\delta\Gamma. \tag{2.19b}$$

It is found from this analysis that a small disturbance develops exponentially with an exponent characteristic of the particular system, known as the Lyapunov exponent [56, 57]. At first glance this result may seem to indicate that accurate simulation is impossible. The way out of this problem is to consider the simulation not as producing a time course but as exploring the allowed areas of phase space. If the sampling is well enough performed then we should be able to calculate thermodynamic properties.

Here we define phase space as the $6N$ dimensional space described by the position and momentum vectors, $\mathbf{q}$ and $\mathbf{p}$, of the system. Each point defined in this space is known as a microstate of the system. Functions of the position and momentum, such as free energies, can now be thought of as forming a complex topological landscape in phase space. In this conception we can think of a molecular dynamics simulation exploring this landscape and sampling the properties of interest at each microstate it visits. Correct statistical treatment of this sampling can then be used in order to evaluate macroscopic thermodynamic properties of the system (such as temperature, pressure and volume).

Statistical mechanics uses the concept of an ensemble of systems, with different microstates but the same macrostate, exploring phase space independently [58]. The average of a property across this ensemble then provides the measurement at the macroscopic level. This conception leads to the construction of a density function, ρ, representing the distribution of the ensemble members over all possible microstates of the system. Liouville's theorem (see Equation 2.20) states that ρ is invariant over time (this applies to both equilibrium and non-equilibrium situations). If this is also true of the integrator used to perform molecular dynamics then we can make use of the ergodic hypothesis which suggests that if the sampling interval used is greater than the correlation interval and the simulation is of sufficient length then the distribution of any thermodynamic variable measured should converge on that of the theoretical ensemble average used to define the macroscopic quantity.

$$\frac{d\rho}{dt} = \frac{\partial\rho}{\partial t} + \sum_{i=1}^{3N} \left(\frac{\partial\rho}{\partial q_i}\dot{q}_i + \frac{\partial\rho}{\partial p_i}\dot{p}_i \right) = 0 \tag{2.20}$$

The functional form of ρ is dependent on which quantities are held constant when averaging is performed. Properties which can be kept constant over a MD simulation are the number of particles N, the energy E, the pressure P and the chemical potential μ. Classical MD can sample in so called NVE ("microcanonical"), NPE, μVT ("grand canonical"), NVT ("canonical") and NPT ensembles. Modifications to the MD algorithm as described so far are required in order to sample ensembles other than NVE, alterations which maintain temperature and pressure are known as thermostats and barostats respectively and are described in the next section.

2.5.1 Thermostats and Barostats

Most experiments occur in conditions approximating the NPT ensemble. In order to sample system states from such an ensemble requires methods to keep the temperature and pressure constant. A method which maintains the temperature at a set value is known as a thermostat, one which maintains a constant pressure a barostat. A wide variety of both thermostats and barostats are available [32, 33] but we will confine ourselves to a discussion to those implemented in NAMD [50].

A measure of temperature may be obtained by making use of the equipartition theorem

$$T = \frac{2\mathscr{K}}{Nk_B} \tag{2.21}$$

where $\mathscr{K}$ is the total kinetic energy of the system, N the number of degree of freedom and k_B is the Boltzmann constant. As this may suggest, the temperature can be controlled by altering the velocities of the components of the system. The approach adopted by Berendsen *et al.* [59] couples the simulation unit cell to a large heat bath, with the velocities scaled such that the change in temperature is proportional to the difference in temperature between unit cell and heat bath. The rate of change of the temperature is then given by

$$\frac{dT(t)}{dt} = \frac{1}{\tau_T}(T_{bath} - T(t)) \tag{2.22}$$

where T is the temperature calculated from the simulation, T_{bath} that of the heat bath and τ_T is the coupling parameter. As this approach simply scales the velocities already found in the system it will maintain discrepancies in the velocity distribution. The consequent "hot solvent, cold solute" problem can be overcome by instead using a model in which heat is transferred from heat bath to unit cell via collisions between particles from the bath and those of the main simulation. In NAMD this is implemented via the Langevin equation [60, 61]:

$$m_i \frac{d^2\mathbf{r}_i(t)}{dt^2} = \mathbf{F}_i(\mathbf{r}_i(t)) - \gamma_i \frac{d\mathbf{r}_i(t)}{dt} m_i + \mathbf{R}_i \tag{2.23}$$

where a frictional force with coefficient γ_i and a stochastic force $\mathbf{R}_i$, simulating thermal noise, are applied to the system and m_i and $\mathbf{F}_i$ represent the mass and force calculated from the potential on the i^{th} particle. The stochastic force does no net work on the system (i.e. $\langle \mathbf{R}_i(t) \rangle = 0$ where $\langle \cdots \rangle$ denotes a time average). The magnitude of the stochastic force can be related to the friction coefficient using the fluctuation dissipation theorem:

$$\langle \mathbf{R}_i(t)\mathbf{R}_i(t') \rangle = 2\gamma_i m_i k_B T \delta(t - t') \tag{2.24}$$

where $\delta(t - t')$ is the Dirac delta function. The choice of γ_i determines whether the frictional or stochastic forces dominate. Appropriate choices of γ_i allow the effective maintenance of a constant temperature. If γ_i is set too high then the system moves from the inertial to the diffusive regime and Brownian dynamics are obtained. Clearly, the use of a stochastic force for temperature control means that simulations employing a Langevin thermostat are not deterministic.

Pressure is maintained in molecular dynamics simulations by scaling the coordinates (and hence volume) of the system. The Berensen barostat [59] employs a pressure bath analogous to the heat bath described above. In fact, the change in instantaneous pressure has a similar form to Equation 2.22

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}(P_{bath} - P(t)) \tag{2.25}$$

where $P(t)$ is the instantaneous pressure, P_{bath} the pressure of the bath and τ_P the pressure coupling parameter. The volume of the system is then scaled by a factor μ

$$\mu = 1 - \kappa \frac{\delta t}{\tau_P}(P - P_{bath}) \tag{2.26}$$

where κ is the isothermal compressibility. The new coordinates are given by

$$\mathbf{r}'_i = \mu^{\frac{1}{3}} \mathbf{r}_i \tag{2.27}$$

Barostats are key to molecular dynamics simulations as it is not generally practicable to build simulation systems in which the particle density is high enough to ensure that the pressure is close to atmospheric pressure. Additionally, most solvation methods leave a gap between solute and first solvation shell. The use of a barostat increases particle density and removes such gaps in the solvent [32].

2.5.2 Improving Computational Efficiency

A range of different techniques have been developed which improve the efficiency of the computation involved in simulation or increase the verisimilitude of small scale simulations to the conditions encountered in real experiments. Here we detail a few of the most commonly used.

2.5.2.1 Periodic Boundary Conditions

Periodic boundary conditions are used to increase the effective size of the simulation environment. This enables the simulation of a relatively small number of atoms in such a way that they feel forces as if they were in a bulk fluid. This allows the method to obtain results which are valid in the thermodynamic limit, in which we can use statistical mechanics to relate the microscopic behaviour observed to macroscopic thermodynamic quantities [32, 33].

The concept is to effectively create an infinite array of images which repeat the contents of the simulation box. The images are created by using integer multiples of the atomic coordinates from the simulated box. In order to conserve the number of particles in the periodic system any atom which moves past the boundary of the simulation box is replaced by an image particle entering from the opposite side of the simulation box. In order for this approximation to work it is necessary to construct the simulation box in such a way that the system does not feel the effects of the boundary. Ensuring a large enough box is used to minimise interactions between proteins and their images is usually the most important factor in avoiding such 'finite size' effects [32].

2.5.2.2 Handling Short Range Force Contributions

One of the most computationally expensive parts of a molecular dynamics simulation is the calculation of the non-bonded energies. In a pairwise model the cost of these calculations scales with the particles simulates, N as $\mathcal{O}(N^2)$. The Lennard-Jones potential is only significant over a very short range (reflecting the r^{-6} dependence of the dispersion reaction) and to calculate it's value for distant atoms will make very little effect. A common way to reduce the computational effort of calculating its effect is to impose a distance cut off beyond which the potential is set to zero [32, 50]. It is conventional in periodic systems to set the cut off such that each atom only interacts with one image of each other atom in the system.

On its own the distance cut off may not result in significant gains in efficiency. This is because it requires the additional computation of the distances between all of the atoms and their comparison with the cutoff (introducing $N(N-1)$ additional calculations). In order to avoid this problem, advantage is taken of the fact that an atoms neighbours are unlikely to change radically over 10–20 timesteps. A list of the atoms which fall within the cut off is created on this timescale, meaning that distance comparisons need to be calculated much less frequently. In many MD codes a list of those atoms just outside the cut off is also calculated, with these atoms being used in the Lennard Jones potential computation only if they move within the cut off distance [32, 50].

The inclusion of a cut off distance introduces a discontinuity in the potential energy (and hence the force) at the cut off. In order to prevent problems with energy conservation most simulation codes multiply the real potential by a switching potential which goes smoothly to zero at the cut off. This alteration to the potential is often only introduced a short distance before the cut off.

2.5.2.3 Handling Long Range Force Contributions

Unlike the Leonard-Jones contribution, the Coulombic, electrostatic, contribution to the potential is significant at long distances as it decays as the inverse of the distance between two atoms. This means that cut offs cannot be implemented without leading to spurious dynamics and consequently the full electrostatic calculation would scale as $\mathscr{O}(N^2)$. When periodic boundary conditions are used the system can be thought of as being infinitely periodic. This can be exploited to help calculate the electrostatic potential. The Ewald sum (first described in 1921 [62]) decomposes the potential into short and long range contributions. The long range contribution can be represented as a sum over the Fourier transforms of the potential and the charge density. This sum converges rapidly and so can be truncated with little error but significant gain in terms of computational workload.

The first term of the Ewald sum requires a Gaussian distribution of width β and equal magnitude but opposite charge to be centred at each atomic position. This has the effect of screening the atomic charges, diminishing the rapid changes at small separations and ensuring a real space summation converges rapidly. This shielding effect must be corrected for by a second term of identical but oppositely charged Gaussians. This second term is not efficiently summed in real space and therefore the cancelling distribution is Fourier transformed and summed in reciprocal space before conversion back into real space. The self interaction of each Gaussian is removed by a third term. Although more complicated than the simple summation of Coulombic terms the Ewald sum converges as $\mathscr{O}(N^{3/2})$ rather than $\mathscr{O}(N^2)$.

A further improvement can be implemented by using the particle-mesh Ewald (PME) method [63] which speeds up the second term summation by interpolating charges onto a three dimensional grid. This allows Fast Fourier Transform (FFT) techniques to be applied to efficiently calculate the Fourier transforms. Increasing the width of the added Gaussians, β, allows faster convergence of the real space sum but slows the computation in reciprocal space. The value of β must be tuned in order to achieve optimal performance. PME scales as $\mathscr{O}(N \ln N)$, permitting the routine calculation of electrostatics without any cut off for periodic systems. This method of dealing with electrostatics is employed throughout this thesis.

2.5.2.4 Multiple Time Step Algorithms

The majority of the time consuming force calculation step of any MD algorithm is spent calculating long range forces that vary slowly with time. It is therefore possible to compute the forces more efficiently by computing these contributions less frequently than the more rapidly varying short range forces. This is the approach used in multiple time step (MTS) algorithms [64, 65]. Most MTS implementations divide forces into three categories. The first, most frequently updated category, is the bonded forces (usually updated approximately every femtosecond). The second includes all non-bonded forces between atoms within a set distance (usually the short range force cut off) of one another updated less often. The final class is the long range electrostatic forces between distant atoms which are updated the least frequently.

2.5.2.5 Constrained Dynamics

The integration timestep that can be used in a simulation is determined by the fastest motions in the system. In biomolecular systems this is the vibrations of hydrogen atoms bound to heavy atoms. If one assumes that these vibrations do not contribute strongly to the overall dynamics of the system then the lengths of these bonds can be constrained and the integrator allowed to proceed more quickly. The SHAKE method, developed by Ryckaert *et al.* [66], assumes that the constraint forces always act along the bonds which they are constraining. In this approach the unconstrained equations of motion are solved first and then the atomic positions corrected. An analytical variant of the SHAKE algorithm, called SETTLE, is designed specifically to constrain bonds in water molecules [67]. The NAMD code used to perform the simulations presented in this thesis makes use of SETTLE to constrain hydrogen atoms within water molecules and SHAKE for those in all other atoms. This combination allows for the extension of the timestep to 2 fs, rather than 1 fs as required in unconstrained dynamics.

2.5.2.6 Accelerating Dynamics

The dynamics seen during a simulation can be accelerated by adding biasing potentials [68] or external forces into the calculation [69]. When a force is applied to a set of atoms to guide it in a particular direction (or set of directions over time) the technique is known as steered molecular dynamics (SMD). Two variants of SMD are widely implemented, ‘constant velocity’ and ‘constant force’ steering. A further variation of the technique has been developed in which a subset of atoms in the simulation is guided towards a final ‘target’ structure by means of the steering forces, this is known as targetted molecular dynamics (TMD).

2.5.3 Available Packages For Biomolecular Molecular Dynamics

A variety of molecular dynamics codes are available designed specifically for the simulation of biomolecular systems. The first packages to gain widespread usage were AMBER [70] and CHARMM [51]. Both packages consist of a suite of programs allowing the construction of the solvated system for simulation, it’s potential energy to be minimised and molecular dynamics performed. In both cases forcefields were developed that are synonymous with their simulation software. The extensive validation and optimisation of these forcefields has led them to be accepted as benchmarks for other forcefields which are developed. The desire to better exploit larger super computers in the mid-1990s led to the a second generation of simulation software. Codes which emerged at this time include GROMACS [71, 72], NAMD [50], LAMMPS [73, 74]. More recently the Desmond [75] code has been produced motivated by further developments in computational resources. These packages allow the use of existing forcefields including variants of AMBER and CHARMM.

The NAMD code has been used to perform all simulations in this thesis, primarily due to its excellent scaling and performance on the large number of processors (CPUs) now available on super computing resources. However, the tools of the AMBER suite were used both to build and analyse systems. This illustrates a common theme in molecular dynamics where a single package is often unable to fulfill all of the requirements of a particular study necessitating the use of elements of multiple codes.

2.5.3.1 Parallel Algorithms

The turn around times possible for any given system depend on two types of optimisation - that of the raw speed of the computation code and the ability of the code to make use of an increasing number of CPUs. As the scale of computational resource available have increased, the ability to partition and execute MD simulations across a larger and larger number of CPUs has been a key determinant of both the size of system and the timescales that simulations can be run on.

There is no unique way in which to parallelise an algorithm and many different techniques have been applied in molecular dynamics codes. The degree of speed up achieved as increasing numbers of processors are made available to a code is known as the scaling and is dependent on the precise algorithms used. In general as more CPUs are used the level of communication between processors increases resulting from a departure from linear scaling. For a given system size a compromise is eventually reached between the performance gained in increasing the usage of an increasing number of CPUs and the penalty imposed by communication between them.

Older MD codes, such as CHARMM and the SANDER module of AMBER, were initially designed using serial algorithms but have been developed more recently to take advantage of parallel computing resources. Initially, this involved the replication of the data of the entire system for each processor and correspondingly poor performance due to memory and communication overheads. A reduced version of SANDER (called PMEMD) has now been included in the AMBER suite which provides improved scaling [70].

Newer codes such as NAMD, GROMACS and LAMMPS decompose the calculations required for the entire system across different processors, using a variety of strategies to achieve this. LAMMPS uses ‘force decomposition’ in which the pairwise forces are evenly distributed across all CPUs [73, 74]. Until recently GROMACS used ‘particle decomposition’ where each atom is assigned to a particular CPU throughout the simulation [71]. In common with NAMD it now uses a scheme known as ‘domain decomposition’, in which the system is divided into a number of spatial regions whose size is greater than the cut off for the non-bonded terms of the potential [50, 72]. In NAMD the non-bonded interactions are additionally grouped into ‘patches’ which are also distributed across the available CPUs. At regular intervals the computational load is then balanced in order to maximise efficiency [50]. This latter strategy has proved to be highly successful for larger systems, allowing scaling to thousands of CPUs.

It is, of course, also the case that for smaller systems the bottleneck is the intrinsic speed of the algorithms used to solve the equations of motion. For example, GROMACS is very fast and for many systems will outperform codes such as NAMD or PMEMD up to 32 CPUs [71, 76]. A different approach that has emphasised sampling of small systems over timescale is that of ensembles of simulations run on single processors by Folding@Home [77] (which uses GROMACS).

More recently the ACEMD code [78] has been developed to take advantage of newly available graphics processing unit (GPU) based processors, which are more efficient than traditional CPUs for many numerical calculations. Versions of NAMD [79], GROMACS and LAMMPS have also been developed to take advantage of GPU technology.

2.6 High Performance Computing

The simulation of biological systems using molecular dynamics is a highly computationally intensive endeavor. In order to exploit its power it is necessary to couple the use of high performance computing with suitably designed codes. In recent years efficiently parallelised codes have become widely adopted within the community of molecular biologists. In order to fully utilise the power of these packages it is necessary to have access to supercomputing resources which allow the user to run programs on large numbers of processors at once.

One of the more appealing methods of being able to do this is called Grid computing. Grid computing has been defined as “distributed computing performed transparently across multiple administrative domains” [80] and aims to provide a common framework for scientists to run their simulations on local clusters or on more powerful national, or international, resources. The original vision of the Grid was to provide uniform methods of access to geographically and organisationally distributed resources (where the resources in question need not be computational but might include storage or even scientific instrumentation) and the name reflects the dream of making these resources available as seamlessly as electrical power is obtained from the electrical grid. In reality although access can be gained to a wide variety of spatially disparate supercomputing resources there is still considerable work needed on middleware that makes the process transparent. These problems continue to apply despite the existence of numerous grid projects spanning those which are national and international in scope and those general in purpose to those designed for specific purposes, examples include the UK National

Grid Service (NGS)[2], DEISA in Europe[3], the US TeraGrid [81][4] and the QCDGrid[5] (dedicated to investigating quantum chromodynamics).

The US TeraGrid in particular provides access to machines with many thousands of processors (currently the largest system on the network is Jaguar at the Oakridge National Laboratory[6] which can perform a maximum of 2.3 petaFLOPS (floating point operations per second) if all 224,256 cores are used). Use of such petascale resources offers the potential to achieve scientific results at unprecedented scales and resolution [82]. In realistic terms these resources now allow us to produce microseconds of all atom molecular dynamics trajectory for a wide variety of biomolecular systems.

A GPU based system called Lincoln[7] has recently been installed at the National Centre for Super Computing Applications (NCSA) in Illinois, as part of the Teragrid. The lower cost of GPUs mean that in the future the number of large scale GPU systems is likely to increase. Lincoln has a maximum processing power of 47.5 teraFLOPS provided by 1536 cores and 96 accelerator units. GPUGRID[8] is a distributed supercomputing infrastructure made of many commodity graphics cards joined together to deliver high-performance all-atom biomolecular simulations [83] which also offers the possibility of reaching similar levels of sampling to the CPU systems available on the TeraGrid.

An even more specialised approach than the use of GPUs comes from the development of the single purpose Anton machine by D. E. Shaw Research. It is designed solely for the purpose of performing MD simulations of proteins and other biological macromolecules [84]. Anton uses specially designed hardware to accelerate compute intensive parts of the simulation such as force calculation, which is claimed to offer much improved overall performance. A 512 node Anton machine will be made available to the research community at the National Resource for Biomedical Supercomputing (NRBSC) at the Pittsburgh Supercomputing Center (PSC)[9] during 2011.

In general, access to high performance computational resources continues to require knowledge of the target resource and can often involve considerable amounts of waiting before a job is processed. Some of the issues with using different middleware and submission processes can be hidden from the user by using the Application Hosting Environment (AHE) [85]. In this thesis we make extensive use of a set of scripts built upon

[2] NGS: www.ngs.ac.uk

[3] DEISA: www.deisa.eu

[4] TeraGrid: www.teragrid.org

[5] QCDGrid: www.gridpp.ac.uk/qcdgrid/

[6] Jaguar: http://www.nccs.gov/computing-resources/jaguar/

[7] Lincoln: http://www.ncsa.illinois.edu/UserInfo/Resources/Hardware/Intel64TeslaCluster/Doc/

[8] GPUGRID: http://www.gpugrid.net/

[9] Anton machine at NRSBC: http://www.nrbsc.org/anton_rfp/

the AHE known collectively as the Binding Affinity Calculator (BAC). A full description of BAC is provided in Appendix A.

2.7 Conclusions

Focussing on the classical molecular dynamics approach, which is to form the basis of the research in this thesis, a variety of experimental and computational approaches to investigating protein structure and conformational change have been discussed. In the case of molecular dynamics, the representation of interatomic interactions by a forcefield function and several widely used methods for integrating the equations of motions have been described. A brief review of a number of molecular dynamics packages and some of the techniques they employ in order to increase computational performance has also been presented.

Chapter 3

Binding Affinities and Molecular Simulation

3.1 Binding Constants and Equilibrium

Molecules which bind to proteins are termed ligands. Whilst in some cases ligands form covalent bonds with proteins (in a process often referred to as 'irreversible' binding), most bind via non-covalent bonds (the process of binding in this way is, unsurprisingly, known as 'reversible' binding) [4]. We concentrate here on the second case, as it is relevant in the majority of cases in which drugs are designed to inhibit a target enzyme. Consider a solution containing fixed total concentrations of a protein, A, and ligand, B, dissolved in a suitable solvent. If the protein and ligand non-covalently bind then ligands will constantly be binding to, and dissociating from, the proteins. These processes can be expressed as the chemical reaction

$$A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} AB \tag{3.1}$$

where k_1 and k_{-1} are the rate constants for binding and dissociation respectively. In equilibrium the rate of the forward and backward reactions are equal, resulting in a stable mixture of free protein, A, free ligand, B and the complex, AB. The concentrations of each of these species at equilibrium determine the equilibrium association constant, K_a (in units of M^{-1}), and the equilibrium dissociation constant, K_d (in units of M), given by

$$K_a = \frac{1}{K_d} = \frac{k_1}{k_{-1}} = \frac{[AB]_{eq}}{[A]_{eq}[B]_{eq}}, \tag{3.2}$$

where the brackets $[\cdots]$ indicate a concentration and the subscript eq indicates that these values must be considered at equilibrium. Rewriting Equation 3.2 in the form

$$\frac{[AB]_{eq}}{[A]_{eq}} = K_a[B]_{eq}, \tag{3.3}$$

equating the ratio of bound protein to free protein to the product of the binding constant and the free ligand helps to clarify the meaning of K_a. It can be seen from Equation 3.3 that the probability that a given protein atom is bound to a ligand goes up as more ligand is present, and that the odds are increased for a higher value of K_a. Thus, it is clear that K_a is a measure of the level of attraction between protein and ligand, consequently it is often referred to as the binding affinity. Despite simply being the reciprocal of K_a the dissociation constant, K_d, is often quoted due to its intuitive interpretation as the concentration of ligand for which the probability of any protein being bound within a complex is a half. To obtain this result we take the occupied protein fraction, σ, given by:

$$\sigma = \frac{[AB]_{eq}}{[A]_{eq} + [AB]_{eq}} \tag{3.4}$$

and multiply both the numerator and denominator by $\frac{[B]_{eq}}{[AB]_{eq}}$ to obtain

$$\sigma = \frac{[B]_{eq}}{[B]_{eq} + \frac{[B]_{eq}\cdot[AB]_{eq}}{[AB]_{eq}}} = \frac{[B]_{eq}}{[B]_{eq} + K_d} \tag{3.5}$$

from which expression the interpretation of K_d is self evident.

3.2 Thermodynamics and Binding Affinity

An alternative analysis of binding processes is provided by thermodynamics. In this view reactions are driven by the minimisation of a potential, the appropriate thermodynamic potential being determined by the conditions in which the reaction occurs. Here we present a brief overview of the thermodynamics (and related enzyme kinetics) relevant to protein-ligand binding, thorough descriptions and analyses of the concepts broached here can be found in the literature [4, 86, 87].

In standard experimental conditions (also known as the NPT ensemble because the number of molecules, pressure and temperature are kept constant) the appropriate potential is known as the Gibbs free energy, G, which is given by

$$\begin{aligned} G &\equiv U + pV - TS \\ &\equiv H - TS \end{aligned} \tag{3.6}$$

where U is the internal energy, p the pressure, V the volume, T the temperature, S the entropy and H the enthalpy of the system. Only if the difference in this potential between the free reactants and the complex is negative can the binding process occur spontaneously. If this is the case then the process will occur until the free energy is minimised and equilibrium is reached. The difference between the potential of the bound and free reactants, ΔG, can be calculated:

$$\Delta G = G(AB) - G(A) - G(B) \tag{3.7}$$

and provides a measure of the strength of binding (the more negative the stronger the attraction between the reactants). This change can also be related to the equilibrium association constant, K_a, via the *van't Hoff* equation:

$$\Delta G = -RT \ln K_a \tag{3.8}$$

where R is the universal gas constant and T the temperature. This equation states that for a given temperature the more negative the value of ΔG the higher the concentration of complex at equilibrium. The close relationship between the two quantities has led ΔG to also be known as the 'binding affinity', although it is interchangeably referred to as the free energy of binding. It is often instructive to break the overall change into its component enthalpic and entropic changes:

$$\Delta G = \Delta H - T\Delta S \tag{3.9}$$

3.3 Enzyme Catalysis

The catalytic function of enzymes requires them to accelerate the rate of reactions without themselves being consumed. Thermodynamics allows us to characterise the equilibrium reached by such processes but cannot describe the rate at which it is achieved. The rate at which a reaction proceeds is determined by the 'activation energy', the difference in energy between the free reactants and the highest energy state along the reaction

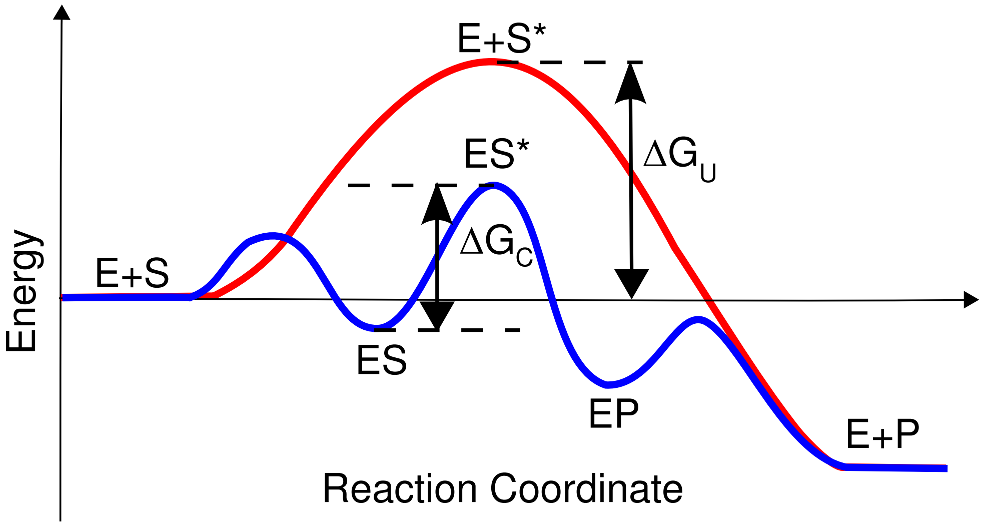


Figure 3.1: Comparison of the free energy barrier of a catalysed, ΔG_C, and uncatalysed, ΔG_U reaction, showing the reduction caused by a catalysing enzyme which causes the reaction to proceed at a more rapid rate. In the uncatalysed reaction the enzyme, E, and substrate, S, do not interact and the substrate forms the product P via the transition state S* at the natural rate of this reaction. In the catalysed reaction the enzyme binds the substrate forming a complex, ES, which then converts into a transition state ES^* which proceeds to a complex of enzyme and product, EP, which then disassociates to give free enzyme and product at a much faster rate than the uncatalysed reaction.

path to the end state of the reaction. It is this barrier that is lowered by enzymes in order to catalyse the reaction (see Figure 3.1). The energy required to do this comes from the binding free energy between the protein and bound substrate. The tighter the binding the greater the amount of energy available to reduce the activation energy. In order to demonstrate the impact this we require a model of the reaction. The most commonly used model to describe catalytic reactions is the Michaelis-Menton equation. This formulation describes a kinetic scheme in which an enzyme (E) binds a substrate (S) which undergoes a reaction and produces a product (P):

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_{cat}}{\rightarrow} E + P \tag{3.10}$$

where k_1 and k_{-1} are the rate constants of the binding and unbinding of the enzyme and substrate respectively and k_{cat} the rate constant of catalysed reaction. The Michaelis-Menton equation (Equation 3.11) is reached by applying the 'law of mass action', in which the reaction rate is proportional to the product of the concentrations of enzyme and substrate, and assuming that the overall enzyme concentration (i.e. of free *and* substrate bound enzyme) remains constant. Furthermore it is assumed that the enzyme concentration is much less than that of the substrate. The original derivation of the equation also assumed that the substrate is in instantaneous equilibrium with the complex, this is known and the ‘equilibrium’ approximation. For this to be valid it is necessary that the condition $k_{-1} \gg k_{cat}$ is met. The equation can then relate the reaction velocity, v_o, to the substrate concentration $[S]$, the theoretical maximum velocity v_{max} and the Michaelis constant, K_m:

$$v_o = \frac{v_{max}[S]}{K_m + [S]} \tag{3.11}$$

The maximum rate of reaction, v_{max} will occur at the saturation point, when all enzyme molecules are bound in a complex. At this point the concentration of the complex, $[ES]$, is equal to the total enzyme concentration $[E]_{tot}$, allowing v_{max} to be written as:

$$v_{max} = k_{cat}[E]_{tot}. \tag{3.12}$$

Using the equilibrium approximation the Michaelis constant, K_m, is defined as:

$$K_m = \frac{k_{-1}}{k_1}. \tag{3.13}$$

In this conception K_m is clearly equal to the dissociation constant of the enzyme substrate complex (see Equation 3.2) and it is apparent that the stronger the binding affinity between enzyme and substrate the faster the reaction will occur. This underlines the importance of an understanding of K_m in ascertaining the rate at which an enzyme catalysed reaction proceeds.

An alternative assumption which yields the same form of Equation 3.11, but altering the meaning of K_m, is known as the the ‘quasi-steady state’ model. Here the complex is assumed to be short lived (i.e $k_{cat} \gg k_1$) and its concentration, $[ES]$, assumed to be constant. In this approach the the Michaelis constant is given by:

$$K_m = \frac{k_{-1} + k_{cat}}{k_1}. \tag{3.14}$$

3.4 Enzyme Inhibition

The treatment of many medical conditions is facilitated by the prevention of the enzymes of an etiological agent or malfunctioning host cell performing their function. Drugs designed to perform this function are known as inhibitors. Most inhibitors bind reversibly to their targets, leaving them with no permanent alteration and unlike substrates they do not usually undergo chemical reactions upon binding. Reversible inhibitors are frequently classified into four groups according to their mode of interaction with their target; competitive, uncompetitive, mixed and non-competitive. Competitive inhibitors, as the name implies, are generally understood to compete with the substrate to bind in the same site in the enzyme. this is not always the case as some operate allosterically, but in all cases the enzyme can only bind either inhibitor or substrate at any one time. Competitive inhibitors do not alter k_{cat}, but do increase K_m. Inhibitors of this type can be out-competed by increasing concentrations of substrate, meaning that v_{max} remains unaltered by their presence. Uncompetitive inhibitors do not interact directly with the binding site but allosterically alter the function of the enzyme, reducing k_{cat} but leaving K_m unchanged. It is generally assumed that uncompetitive inhibitors bind to the enzyme-substrate complex. Mixed inhibitors are those which interact in someway with the substrate binding site (partially blocking it or becoming part of a modified binding site). This will reduce k_{cat} but may increase or decrease K_m. Non-competetive binding is a special case of mixed inhibition in which substrate binding is unaffected but k_{cat} is decreased.

Competitive binding is probably the most frequently employed mode of inhibition and the kinetic scheme described in Equation 3.10 can easily be amended to incorporate such an inhibitor, I, alongside the enzyme, E, and substrate, S:

$$\begin{aligned} E + S + I &\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_{cat}}{\rightarrow} E + P + I \\ E + S + I &\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} EI + S \end{aligned} \tag{3.15}$$

where the rate constants k_1, k_{-1} and k_{cat} are preserved unchanged from Equation 3.10 and a reversible reaction is included for the inhibitor bound to the enzyme with rate constants k_2 for the forward and k_{-2} for the reverse reactions respectively. The modified Michaelis-Menton equation can again be derived in identical form by applying either the equilibrium or quasi-steady state assumptions are applied;

$$v_o = \frac{v_{max}[S]}{K_m(1 + \frac{[I]}{K_i} + [S]}, \tag{3.16}$$

with the same differences in the definitions of K_m (see Equation 3.13 and Equation 3.14), $[I]$ and $[S]$ being the concentrations of inhibitor and substrate respectively and K_i is the disassociation for the enzyme-inhibitor complex (defined in Equation 3.17).

$$K_i = \frac{k_{-2}}{k_2}. \tag{3.17}$$

Thus K_i has an identical form to the disassociation constant, K_d, as described for more general ligands in Section 3.1. Consequently, K_i is an obvious metric for estimating the strength of inhibitor binding. The lower the K_i the stronger the binding and, consequently, the more effective the inhibitor is at preventing the target enzyme performing its catalytic function; hence inhibitor minimising this quantity is of great importance in the optimisation of inhibitor design. As previously noted, minimising the disassociation constant is equivalent to maximising the free energy of binding, ΔG.

3.5 Experimental Measurements of Binding

The free energy of reactions cannot be measured experimentally at the molecular level, meaning that experiments are generally designed to follow the binding kinetics of the enzyme in a mixture with the inhibitor (and often a natural substrate). A repository of experimentally obtained binding free energies for a wide range of proteins and small molecules can be found at the BindingDB[1] [88, 89] along with details of the methods used to obtain them. Here we give details of the three methods most frequently used to produce the binding affinity data for protein-inhibitor binding.

3.5.1 Inhibition Assays

In steady state inhibition assays the generation of products or the consumption of substrate in a reaction is monitored in order to measure the rate of the reaction being catalysed by the enzyme. A fitting procedure can then be used to convert the observed rate into the parameters of the Michaelis-Menton model (see Equation 3.11), K_m and k_{cat}. The inhibitor to be studied can then be added to the mixture and its binding affinity derived from the impact this has on the reaction rate (and hence K_m and k_{cat}) using the modified Michaelis-Menton model described by Equation 3.16.

The results of such assays are frequently not reported as K_i (or equivalently K_d) values but in terms of the concentration of inhibitor which reduces enzyme activity by 50%, which is known as the IC_{50}. This will depend on both the substrate concentration with which the experiment is performed but also how tightly it binds the enzyme. For competitive inhibitors the IC_{50} can be converted to a K_i value using the Cheng-Prussof equation [90]:

[1] BindingDB:http://bindingdb.org

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}}, \tag{3.18}$$

where $[S]$ is the substrate concentration and K_m is the Michaelis constant of the reaction involving the enzyme and substrate. As can easily be seen in general, the IC_{50} will typically be larger than K_i but when [S] is low then the two values will be essentially equal. The dependence on K_m means that, unlike K_i, it is not strictly correct to compare IC_{50} values for the same inhibitor bound to different enzymes but only those for different inhibitors binding to the same enzyme.

In many situations enzyme inhibition assays are particularly convenient as the fact that each enzyme molecule can generate many reactions means that it acts as an amplifier. The main problem with the technique is the need to devise a method of detecting the removal of substrate or generation of product. A common approach is to devise a substrate for which the fluorescence properties are altered by the reaction and use a fluorimeter to measure enzyme activity as a function of inhibitor concentration [4, 91].

3.5.2 Pre-Steady State Assays

After the addition of a ligand to an enzyme containing solution there is an initial stage of rapid complex formation before equilibrium is reached. The study of product formation during the first few milliseconds of the reaction, i.e. in the first turnover, is called pre-steady state (transient) kinetics. Observations of this phase of the reaction are much more demanding to make than those at equilibrium due to the requirement of rapid mixing and high temporal resolution measurements. Such experiments can be used to determine rate limiting steps and the transition states through which the reaction proceeds. The advantage of experiments probing this regime in the context of enzyme inhibition assays is the ability to determine both the forward and reverse rate constants of the reaction, k_1 and k_{-1} (and hence K_i, K_d, calculated as in Equation 3.2) [4, 91].

Rapid kinetic experimental techniques, such as stopped-flow methods or rapid chemical quench-flow, allow the rate of the reaction to be observed via the detection of changes in protein florescence or light scattering. The reaction rate observed in such experiments will be given by:

$$k_{obs} = k_1[I] + k_{-1} \tag{3.19}$$

where $[I]$ is the inhibitor concentration (which in these assays is in excess compared to that of the enzyme). By changing the concentration of inhibitor and fitting to this linear relationship the forward and reverse constants k_1 and k_{-1} can be obtained.

3.5.3 Isothermal Titration Calorimetry

The binding of an inhibitor and enzyme will often lead to small amounts of either heating or cooling (i.e. the reaction may be exothermic or endothermic). The detection of these changes via isothermal titration calorimetry (ITC) allows the calculation not only of the binding free energy, ΔG, but also the enthalpic and entropic components thereof (ΔH and ΔS). In ITC a cell containing a solution of one reactant is maintained at a constant temperature using a thermostat. Small injections of aliquots of the second reactant are then titrated into the cell. The chemical heat release or uptake is evaluated based upon the energy required by the thermostat to keep the solution at constant temperature. Each injection, i, of the second reactant causes the absorption or release of a quantity of heat, q_i. This heat can be related to the amount of ligand that binds to the protein and the enthalpy of the reaction ΔH via the relation:

$$q_i = v\Delta H\Delta L_i \tag{3.20}$$

where v is the cell volume and ΔL_i is the increase in concentration of bound ligand upon the i^{th} reactant injection[4, 91]. The energy requirement for each successive injection becomes less as the free reactant becomes bound. The heat change during each injection is proportional to the amount of complex formed. Consequently, the change in heat over the course of the titration can be used to calculate the binding affinity constant, K_a (and hence the Gibbs free energy, ΔG, via Equation 3.8). ΔH is calculated from the experiment directly and using Equation 3.9 the value of ΔS can also be calculated.

3.5.4 Experimental Errors

Experimental free energy differences are often quoted with errors of the order of 0.1 kcal mol^{-1} (often less than 1% of the total values) [92–96]. This does not seem credible given the number of factors that can affect measurements and the difficulty in controlling them. Sources of error include instrumentation accuracy, inadvertent inactivation of some enzymes and incorrect assessment of concentrations (for example proteins may stick to vessel walls). In some cases, results published by the same group for the same enzyme and inhibitor pairing show significant variability. For example binding affinities for the anti-cancer drugs Gefitinib and AEE788 to the human epidermal growth factor receptor (EGFR) in two studies published within a year of one another vary by 0.4 kcal mol^{-1} [97, 98].

3.6 The Theoretical Basis of Computational Free Energy Calculations

An alternative approach to calculating the binding energetics of protein-ligand interactions is via computational techniques. It is this method of investigation which will provide the main focus of this thesis. These methods have a number of advantages over experiments as they can be used to study molecules which are difficult, or impossible, to synthesise and can provide atomistic insight into the systems under study. Molecular simulations (as described in Chapter 2) generally describe single enzyme systems and consequently in order to understand how they can be used to calculate macroscopic thermodynamic variables requires a brief description of the relevant statistical mechanics that provide the theoretical link between these two scales. Here we describe how some basic statistical mechanical concepts apply to molecular simulations and how they help understand the power and limitations of using them in free energy calculations. Again more detailed treatment of the subject is available in standard texts [49, 58].

3.6.1 Free Energy and Statistical Mechanics

As noted in Section 3.2, binding reactions are driven by the minimisation of the relevant thermodynamic potential depending on conditions. The Gibbs free energy, G, used in the usual isothermal-isobaric (NPT) experimental conditions has already been described. In this section we will turn to the Helmholtz free energy, A, which is used in the canonical ensemble (constant particle number (N), volume (V) and temperature (T) conditions, also known as NVT) in order to derive results in order to simplify the mathematics. The form of the two potentials is:

$$\begin{aligned} G &\equiv H - TS, \\ A &\equiv U - TS, \end{aligned} \tag{3.21a}$$

where H is the enthalpy, S the entropy, T the temperature and U is the internal energy of the system. Consideration of the partition function in both ensembles is instructive both as to the differences between the ensembles and the simplifications allowed by using the Helmholtz free energy. The partition function is the sum of the energies of all the microstates accessibly by the system and can be used to calculate many thermodynamic quantities. In the isothermal-isobaric ensemble the partition function is given by:

$$Q_{NPT} = \frac{1}{h^{3N}N!} \int \int \int e^{-\beta(E(\mathbf{p},\mathbf{r})+pV)} \, d\mathbf{p} \, d\mathbf{r} \, dV, \tag{3.22}$$

where $\mathbf{p}$ and $\mathbf{r}$ represent the momentum and position coordinates of the system respectively, $E(\mathbf{p}, \mathbf{r})$ the energy of a particular microstate, V is the volume, p the pressure, β is the inverse temperature ($1/k_BT$, where k_B is the Boltzmann constant), h is Plank's constant and N the number of particles. The prefactor here assumes that the particles are indistinguishable. In the canonical ensemble the equivalent quantity is given by:

$$Q_{NVT} = \frac{1}{h^{3N}N!} \int \int e^{-\beta E(\mathbf{p},\mathbf{r})} \, d\mathbf{p} \, d\mathbf{r}. \tag{3.23}$$

The form of this equation is clearly simpler than that of Equation 3.22, the difference being that no volume integral is required and the Boltzmann factor does not include the pV term. Both partition functions can be interpreted as sums over phase space weighted by the Boltzmann factor. This factor is a function of the microstates of the system and it is this property that means that the partition function provides the link between the microstates and macrostates of the system it describes.

The partition function can be used to calculate the internal energy and Helmholtz free energy of the system. We will now use these connections to show that it is significantly more difficult to calculate converged values of the later than it is the former. The relationship between the partition function, Q, and internal energy, U is:

$$U = -\frac{1}{Q(\mathbf{p}, \mathbf{r})} \frac{\partial Q(\mathbf{p}, \mathbf{r})}{\partial \beta}. \tag{3.24}$$

Substituting for the canonical partition function from Equation 3.23:

$$U = \int \int \frac{E(\mathbf{p}, \mathbf{r}) e^{-\beta E(\mathbf{p},\mathbf{r})}}{Q(\mathbf{p}, \mathbf{r})} \, d\mathbf{p} \, d\mathbf{r}. \tag{3.25}$$

If we define the probability density, $\rho(\mathbf{p}, \mathbf{r})$ as:

$$\rho(\mathbf{p}, \mathbf{r}) = \frac{e^{-\beta E(\mathbf{p},\mathbf{r})}}{Q(\mathbf{p}, \mathbf{r})}, \tag{3.26}$$

then we can relate U to the ensemble average of $E(\mathbf{p}, \mathbf{r})$:

$$U = \int\int E(\mathbf{p},\mathbf{r})\rho(\mathbf{p},\mathbf{r})\,d\mathbf{p}\,d\mathbf{r} = \langle E(\mathbf{p},\mathbf{r})\rangle\,. \tag{3.27}$$

The integral involved in this expression is over all of phase space but the linear dependence on $E(\mathbf{p},\mathbf{r})$ means that those areas visited with a low probability do not make significant contributions to the overall value. A consequence of this is that calculations of U converge quickly.

Where the internal energy is related to the partition function through a derivative the relationship for the Helmholtz free energy is direct:

$$A = -\frac{1}{\beta}\ln Q(\mathbf{p},\mathbf{r}). \tag{3.28}$$

Again we substitute for the canonical partition function from Equation 3.23 to give:

$$A = -\frac{1}{\beta}\ln\left(\frac{1}{h^{3N}N!}\int\int e^{-\beta E(\mathbf{p},\mathbf{r})}\,d\mathbf{p}\,d\mathbf{r}\right). \tag{3.29}$$

Multiplication of both numerator and denominator by $\int\int e^{-\beta E(\mathbf{p},\mathbf{r})}e^{\beta E(\mathbf{p},\mathbf{r})}$ and the discarding of constant prefactors gives:

$$A = \frac{1}{\beta}\ln\left(\frac{\int\int e^{-\beta E(\mathbf{p},\mathbf{r})}e^{\beta E(\mathbf{p},\mathbf{r})}\,d\mathbf{p}\,d\mathbf{r}}{\int\int e^{-\beta E(\mathbf{p},\mathbf{r})}\,d\mathbf{p}\,d\mathbf{r}}\right). \tag{3.30}$$

Substituting for the probability density we obtain:

$$A = \frac{1}{\beta}\ln\left(\int\int e^{\beta E(\mathbf{p},\mathbf{r})}\rho(\mathbf{p},\mathbf{r})\,d\mathbf{p}\,d\mathbf{r}\right). \tag{3.31}$$

As for the internal energy this can be recast in terms of ensemble averages:

$$e^{\beta A} = e^{\langle\beta E(\mathbf{p},\mathbf{r})\rangle}. \tag{3.32}$$

The exponential nature of this relationship indicates that high energy regions of phase space, which will be infrequently visited, contribute significantly to the free energy (a similar result is obtained for the Gibbs free energy, see Equation 3.33). This means that free energy values will be considerably more difficult to converge than those of the internal energy

$$e^{\beta G} = e^{\langle \beta E(\mathbf{p},\mathbf{r}) + pV \rangle}. \tag{3.33}$$

3.6.2 Convergence and Sources of Error

Just as in experimental approaches, theoretical calculations contain various errors. This section reviews the causes of these inaccuracies.

3.6.2.1 Incomplete Sampling of Phase Space

As highlighted by Equation 3.32 the level of phase space sampling is key to obtaining converged and hence reliable free energy values. The phase space of proteins is too large for all of it to be explored, consequently we must confine our ambitions to sampling it in a representative fashion. To understand the proposition of representative sampling it is instructive to consider the traditional picture of the free energy landscape of a folded protein as a series of free energy minima separated by high energy barriers. If our simulation only visits a single local minimum then it is clear that it has not sufficiently sampled the free energy surface; if it has sampled from a large number of such minima then the chances are higher that phase space has been adequately explored. In general the topology of the phase space under investigation is unknown and consequently, in order to validate our computational results, we must use comparisons with experimentally measured values.

3.6.2.2 Accuracy of the Interatomic Potential Energy Force Field

No matter how thoroughly the free energy landscape is sampled, the accuracy of the calculations will be limited by how well our model of the protein describes the real system .In general as quantum mechanical calculations are impractical for systems the size of proteins this usually means that the accuracy of computational free energy calculations is dependent on the quality of the forcefield potential (a definition of which is given in Chapter 2) used to describe the inter atomic interactions.

3.7 Formally Exact Free Energy Calculations

A number of formally 'exact' methods, containing no empirically fitted parameters, for calculation of free energy differences from molecular simulation have been developed. In this section we shall explore two techniques illustrative of those available: (i) free

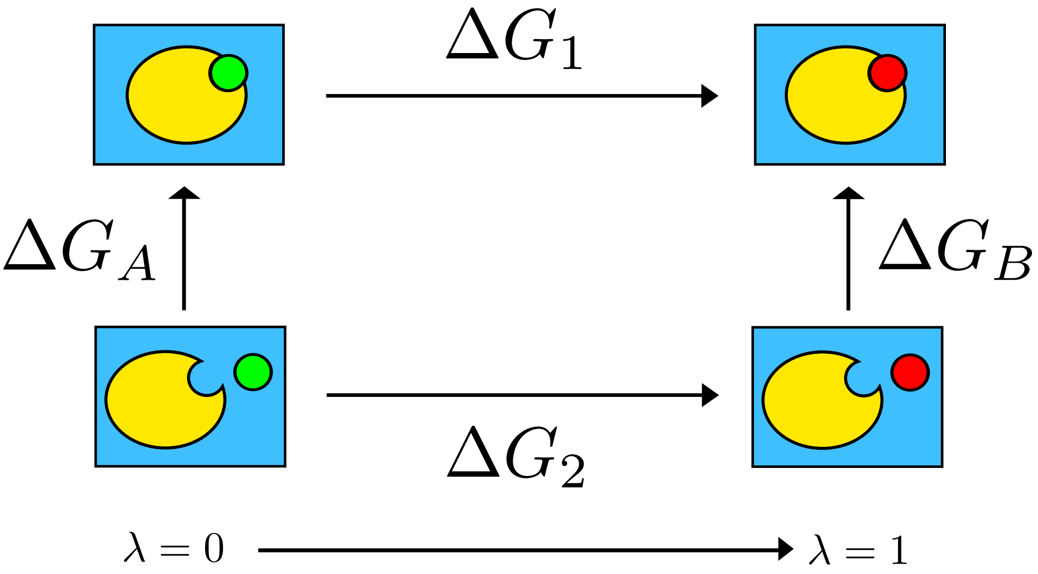


Figure 3.2: A thermodynamic cycle illustrating the indirect calculation of the relative binding free energy difference, $\Delta\Delta G_{bind}$ (using Equation 3.35). Two ligands A (green) and B(red) bind to the same protein (represented in yellow). The free energies of ligand binding (ΔG_A and ΔG_B) and the alchemical transformation of one ligand into another (ΔG_1 and ΔG_2) are labeled

energy perturbation (FEP) which is based on exponential averages of the change in the potential energy, and (ii) thermodynamic integration (TI), which is based on integrating the the change in energy as one system description is gradually changed into another. A number of reviews of a wider range of methods, many of which share similar conceptual foundations to those described here, are available in the literature [33, 99–101]. In general, considerations of computational efficiency lead to 'exact' methods being employed to compute free energy differences arising from non-physical system transformations (known as alchemical changes). Thermodynamic cycles can then be used to relate these changes to the free energy differences and relative differences which are of real interest.

3.7.1 Thermodynamic Cycles

Free energy is a function of state and as such cumulative variations around a closed thermodynamic cycle sum to zero. Figure 3.2 depicts such a cycle which could be used to calculate the relative binding affinity, $\Delta\Delta G_{bind}$, of two different ligands to the same target where

$$\Delta\Delta G_{bind} = \Delta G_B - \Delta G_A \tag{3.34}$$

The vertical paths describe the binding of the two ligands, A and B, to the receptor with binding free energy differences of ΔG_A and ΔG_B respectively. This paths involve substantial rearrangements as the ligands move towards the target before binding (which may itself involve conformational changes) and consequently the binding free energy measurements will be very slow to converge using molecular dynamics. The horizontal paths (and associated free energy differences ΔG_1 and ΔG_2) describe the alchemical transformation of one ligand into another in the bound and unbound states. The conversion of one ligand into another is clearly unphysical but is easily accomplished in a computational model. If the ligands are similar the system reorganization will be minor and the alchemical free energy calculations will converge much more quickly than the physical ones. We can now make use of the properties of state functions and calculate the relative binding affinity in terms of the alchemical free energies as

$$\Delta\Delta G_{bind} = \Delta G_1 - \Delta G_2. \tag{3.35}$$

The value obtained is independent of the path taken from system A (protein P bound to ligand X) to system B (protein P bound to ligand Y). The path taken is represented by the alchemical Hamiltonian employed, which is conventionally given as a function of a parameter λ where $0 \leq \lambda \leq 1$. The only constraint on the form of the Hamiltonian is then that it satisfies the boundary conditions that a value of 0 represents the Hamiltonian of system A and 1 that of system B. An identical argument applies in the situation in which the same ligand is bound to two slightly different targets (for example mutated forms of the same protein).

Thermodynamic cycles can also be constructed to calculate absolute free energy differences and this approach is beginning to be more widely used. The two most commonly used approaches are know as the 'double decoupling' and 'double annihilation' methods. These make use of constraints and more complicated cycles than those presented here. Excellent reviews on the topic are available by Deng & Roux [101] and Shirts *et al.* [100]. We will confine the discussion in the following sections to the calculation of relative binding free energies.

3.7.2 Free Energy Perturbation

The free energy perturbation (FEP) method calculates free energy differences of alchemical transformations from one system to another (see Figure 3.2 for a relevant example for relative binding free energy differences in protein-ligand systems). The kinetic and potential energy contributions to the system Hamiltonian are separated and it is assumed that there is no change in the kinetic component between the two systems. The potential energy contribution will obviously be altered along with the change in the molecular composition of the system. The free energy change of the system can be obtained from the ensemble average of the difference in the potential between the two systems A and B using

$$\Delta G = \frac{1}{\beta} \ln \langle e^{-\beta \Delta \mathscr{V}} \rangle_A, \tag{3.36}$$

where β is $1/k_B T$ and $\Delta \mathscr{V} = \mathscr{V}_B - \mathscr{V}_A$, $\langle \cdots \rangle$ represents an ensemble average and the subscript A denotes an average calculated from MD trajectories using the potential for system A. The values of $\mathscr{V}_B$ are calculated from the coordinates generated by the simulation using the potential for system A. For this approach to be valid the potential for system B calculated in this manner is required to represent a low energy state. In practice, this is not usually the case and a series of unphysical intermediate states between those of A and B are used, with an independent simulation run for each one. The sum of the ΔG values obtained for each step provides the value for the complete transformation. A parameter, λ, with values between 0 and 1 is used to transform the potential of one system into the other via:

$$\mathscr{V}_m(\lambda_m) = (1 - \lambda_m)\mathscr{V}_1 + \lambda_m \mathscr{V}_2 \tag{3.37}$$

with the subscript m denoting a step number which increments over the transformation process.

3.7.3 Thermodynamic Integration

The thermodynamic integration (TI) methodology requires a series of simulations to be run at different values of the parameter λ. This parameter runs from 0 to 1 and describes the gradual conversion of the system from one set of molecular constituents to another. The ensemble average of the derivative of the potential energy with respect to λ is then numerically integrated as seen in Equation 3.38. As in FEP λ is coupled to the MD potential and with values other than 0 and 1 represents an unphysical transition between the two systems.

$$\Delta G = \int_0^1 \left\langle \frac{\partial \mathscr{V}(\lambda, x)}{\partial \lambda} \right\rangle \partial \lambda. \tag{3.38}$$

Thermodynamic integration has the advantage over FEP that each average required is independent of the others (in FEP $\Delta\mathscr{V}$ values are averaged, meaning that values from two trajectories must be considered). On the other hand, the fact that sufficient simulations need to be run to calculate the numerical integral further increases the computational load.

In both FEP and TI calculations for the alchemical transformation of both bound and unbound states need to be performed in order to estimate the relative binding free energy difference of two systems. Furthermore, they both require that the increments of λ are small. In the case of FEP this is to ensure phase space overlap between the two states and in TI it is necessary in order to provide enough points for accurate numerical integration. Typically this requires of tens of simulations to be run, all long enough to gain sufficient sampling of the potential. These requirements mean that both methods suffer from the fact that they are exceptionally computationally intensive [102]. For this reason, and to allow the comparison of a wider range of ligands, more approximate methods have been developed.

3.8 Approximate Methods

The use of formally exact methods for calculating free energies has hitherto been enormously time consuming and obtaining well converged values difficult. The computational expense also means that limited numbers of systems have been studied and definitive comparison between methodologies has not yet proved possible. In response to this and the need for rapid measurements of different systems in applications such as drug discovery, a range of approximate methods have been developed. These methods employ less accurate physical models and empirically fitted parameters to allow faster turn around of calculations. In this section a variety of such methods are described in order of increasing physical rigour; excellent reviews are also available in the literature such as those by Gilson & Zhou [103] and Steinbrecher & Labahn [104].

3.8.1 Docking and Empirical Surface Area Based Methods

In applications such as the identification of drug candidates and the identification of binding poses of known drugs it is often necessary to rapidly assess the binding affinity of a variety of drugs, or drug-protein conformations. A wide variety of ‘scoring functions’ both physical and empirical, have been developed for this purpose, examples include AutoDock [105], X-Score [106], DrugScore [107], ChemScore [108], GOLD [109], FlexX [110], LigScore [111] and LUDI [112]. In general single structures are evaluated and little or, more commonly, no protein motion is incorporated into the evaluation. The simplest version of this concept is that of the empirical surface area based methods, based on the observation that ligand binding induces a reduction in the protein surface area accessible by surrounding solvent. Often the scoring function assumes that the change in free energy is lineary dependent on the changes in the areas of polar and non-polar regions which can come into contact with solvent. There is little evidence that such simplistic models have any predictive power and as such they are infrequently used. More sophisticated models accounting more accurately for the physics involved are frequently used for purposes such as deciding which preliminary drug candidates can be discarded in drug design applications where they are found to give a useful first approximation to the free energy. In these high throughput scenarios the emphasis is on the speed of calculations despite the inevitable, major trade-off in accuracy. The form of both empirical and physics based scoring functions designed for these type of docking applications is under constant development with insights from more rigorous binding free energy calculations becoming incorporated over time [113].

3.8.2 Linear Interaction Energy

The Linear Interaction Energy (LIE) is another approximate method which considers the end points of the binding process. At its core LIE simply considers the van der Waals, $\mathscr{V}_{vdW}$, and electrostatic, $\mathscr{V}_{elec}$, interactions of the ligand with its surrounding environment [114, 115]. A contribution which considers the solvent accessible area of the ligand, A, can also be added [116]. The averages of these quantities are used to compute the change in free energy according to the equation

$$\Delta G = \alpha \Delta \left\langle \mathscr{V}_{elec} \right\rangle + \beta \Delta \left\langle \mathscr{V}_{vdW} \right\rangle + \gamma \Delta \left\langle A \right\rangle \tag{3.39}$$

where the constants α, β and γ are semiempirically derived for each system. The initial variant of the method proposed a theoretical value of 0.5 for β assuming a linear response of the surroundings to electrostatic fields [114]; more recent formulations allow it to vary with the chemical composition of the ligand [117]. The requirement to fit these parameters make it ill suited to screening novel ligands in most cases.

3.8.3 Molecular Mechanics Poisson-Boltzmann Surface Area

The Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) [118, 119] methodology is the most rigorous of the approximate methods presented here and has been used successfully to perform free energy calculations on a variety of biological systems. It is a good candidate for use in drug comparisons as it can treat widely varying systems and can be calculated from a single MD run, unlike FEP or TI. The combination of a reasonably detailed physical model and rapidity of calculation (compared to exact methods) has led to its widespread usage. As it is the methodology utilised in much of the quantitative analysis presented in this thesis we will discuss its basis in some detail.

The application of the MMPBSA method involves the computation of absolute binding free energies by calculating average free energies of the enzyme/inhibitor complex, the enzyme alone and the inhibitor alone. These values are then used to calculate the change in free energy using

$$\Delta G = \langle G_{complex} \rangle - \langle G_{enzyme} \rangle - \langle G_{ligand} \rangle \, , \tag{3.40}$$

where $\langle \cdots \rangle$ indicates an ensemble average over the values calculated by post processing a series of representative frames from molecular dynamics trajectories. The MD simulations are typically performed in a periodic box solvated with explicit solvent and counter ions. Two strategies are employed to help facilitate the calculation of converged free energy differences. Firstly, the solvent and counter ions are removed from the frames and they are replaced by a continuum solvent representation and, secondly, a thermodynamic cycle is employed.

3.8.3.1 Thermodynamic Cycle

The ultimate objective of free energy calculations is the absolute free energy of binding in an appropriate (usually aqueous) solvent, ΔG_b^{aq}. However, in simulations of solvated protein ligand systems the majority of the energy contributions would come from solvent-solvent interactions, resulting in fluctuations in total energy an order of magnitude larger than the binding energy. Hence, direct calculations would require inaccessible levels of sampling to converge. The solution to this employed by MMPBSA is to use the thermodynamic cycle pictured in Figure 3.3. The *in vacuo* binding free energy, ΔG_b^{vac}, is calculated along with the solvation energies of the complex ($\Delta G_{complex}^{sol}$), enzyme (ΔG_{enzyme}^{sol}) and ligand (ΔG_{ligand}^{sol}). The binding free energy of the solvated system is then given by

$$\Delta G_b^{aq} = \Delta G_b^{vac} + \Delta G_{complex}^{sol} - (\Delta G_{enzyme}^{sol} + \Delta G_{ligand}^{sol}) \tag{3.41a}$$

$$= \Delta G_b^{vac} + \Delta G^{sol} \tag{3.41b}$$

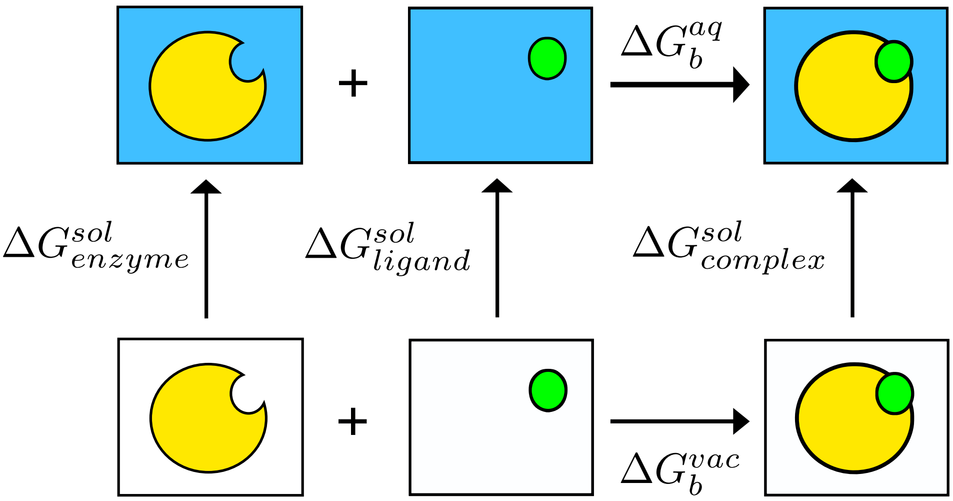

ΔG_b^{aq}
ΔG_{enzyme}^{sol}
ΔG_{ligand}^{sol}
$\Delta G_{complex}^{sol}$
ΔG_b^{vac}
+
+

Figure 3.3: Solvation thermodynamic cycle used to indirectly calculate the absolute free energy of binding, ΔG_b^{aq}, of a ligand to a target in aqueous solvent in the MMPBSA methodology. The binding free energy change *in vacuo*, ΔG_b^{vac} and the solvation free energies for the complex ($\Delta G_{complex}^{sol}$), enzyme (ΔG_{enzyme}^{sol}) and ligand (ΔG_{ligand}^{sol}) are calculated and the ΔG_b^{aq} computed via Equation 3.41.

The ΔG_b^{vac} and ΔG^{sol} components of the binding affinity are calculated separately, using different methodologies. We now discuss how they are decomposed, before discussing in detail the approaches used to compute each component of the binding affinity.

3.8.4 Decomposition of the Free Energy

The *in vacuo* binding free energy, ΔG_b^{vac}, is calculated using the molecular mechanics forcefield used to describe the system during the molecular dynamics simulation. It can be separated into a sum of electrostatic, van der Waals and internal molecular mechanics interactions:

$$\Delta G_b^{vac} = \Delta G_{ele}^{MM} + \Delta G_{vdW}^{MM} + \Delta G_{int}^{MM}. \tag{3.42}$$

The calculation of ΔG^{sol} is more complicated. This term represents the free energy change associated with taking the solute from vacuum into a solvent environment. This can be decomposed into contributions from polar and non-polar interactions between the solute and solvent:

$$\Delta G^{sol} = \Delta G^{sol}_{pol} + \Delta G^{sol}_{nonpol}. \tag{3.43}$$

The polar solvation energy is calculated using a numerical solution of the Poisson-Boltzmann equation with an implicit solvent modelled as a high dielectric constant medium and the solute as one with a lower dielectric constant (more details are given in Section 3.8.6). The non-polar contribution is estimated using a term related to the solvent accessible surface area (see Section 3.8.6).

3.8.5 Single and Component Trajectories

The necessary elements of Equation 3.40 can be calculated either from separate trajectories or from one single trajectory. In the latter case the the individual components are calculated using the snapshots for the entire system but effectively removing the parts which are not of interest. This approximation has been seen to be a good approximation in several instances [115, 118, 120]. The single trajectory approach has the advantage that contributions from parts of the system that do not affect binding exactly cancel as the same coordinates are used for the components both as part of the complex and separated (ΔG^{MM}_{int} is, trivially, zero in this approach). If separate trajectories are used then this cancellation of errors does not occur as each trajectory is free to explore different conformations. On the other hand the single trajectory approach will obviously not be able to account for alterations in the dynamic properties of the enzyme or drug induced by binding and consequently may ignore important contributions to the overall binding affinity.

3.8.6 Poisson-Boltzmann

In order to estimate the polar solvation energy, ΔG^{sol}_{pol}, we need to model the electrostatic potential surrounding the system of interest in an appropriate solvent environment. As the name suggests the approach taken within the MMPBSA methodology is to numerically solve the Poisson-Boltzmann equation. The Poisson-Boltzmann equation uses an implicit solvent model in which the solvent is treated as a high dielectric constant continuum, aqueous ions as a diffuse "charge cloud" and the solute as a collection of fixed point charges embedded in a lower dielectric continuum.

The Poisson-Boltzmann equation may be derived from statistical mechanical considerations [121] but a more straight forward approach is to begin with Poisson's equation[122, 123]

$$\nabla \cdot [\epsilon_0 \epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}), \tag{3.44}$$

used to describe the electrostatic potential $\phi(\mathbf{r})$ at a point $\mathbf{r}$ generated by a charge distribution $\rho(\mathbf{r})$ in an environment of dielectric coefficient $\epsilon(\mathbf{r})$ (relative to the permittivity of free space, ϵ_0). The dielectric coefficient describes the local polarizability within the system. In the context of biomolecular simulations the functional form of $\epsilon(\mathbf{r})$ will depend on the geometry of the system being studied with the biomolecule represented as continuum region of low polarizability embedded in a surrounding continuum region of higher polarizability representing the solvent. Typically, for biomolecular systems the dielectric constant $\epsilon(\mathbf{r})$ of the solute is chosen to be in the range 1 to 4 (although values as high as 20 are sometimes used) and a value of 80 is used to represent water [123, 124]. The boundary between the two regions is imprecisely defined with several methods used in practice [125–128]. The most common method is to use the centre of a hypothetical rolling sphere with a radius of a water molecule on the van der Waals surface of the molecule to determine the position of the boundary [129]. The charge distribution $\rho(\mathbf{r})$ can be decomposed into two parts, the fixed solute charge density, $\rho_f(\mathbf{r})$, and a contribution from the ions present in the solvent, $c(\mathbf{r})$. The former is generally described as a set of delta functions centred on each solute atom's centre and scaled by the atom's charge. The ion contribution is modelled as a continuum with charge distributed according to the Boltzmann distribution. For M ion species with charges q_j and bulk concentrations c_j^∞ the ion charge distribution is given by

$$c(\mathbf{r}) = 4\pi \sum_{i=1}^{M} q_j c_j^\infty e^{-\beta q_j \psi(\mathbf{r})}, \tag{3.45}$$

with $\beta = 1/k_B T$. Substituting this into (3.44), we obtain the following:

$$\nabla \cdot [\epsilon_0 \epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] + 4\pi \sum_{i=1}^{M} q_j c_j^\infty e^{-\beta q_j \psi(\mathbf{r})} = -4\pi\rho_f(\mathbf{r}); \tag{3.46}$$

this can be simplified for the case of two ions of equal bulk concentration, c^∞, which have opposite charges of equal magnitude, q, providing electrostatic neutrality, to

$$\nabla \cdot \epsilon_0 \epsilon(\mathbf{r}) \nabla \phi(\mathbf{r}) - 8\pi q c^\infty sinh[\beta q \psi(\mathbf{r})] = -4\pi\rho_f(\mathbf{r}). \tag{3.47}$$

This equation can be linearised by expanding the hyperbolic sine function as a Taylor series and retaining only the first term. This process results in what is known as the linearised Poisson-Boltzmann equation:

$$\nabla.\epsilon(\mathbf{r})\nabla\phi(\mathbf{r}) - 8\pi q^2 c^{\infty}\beta\psi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{3.48}$$

A wide variety of numerical methods can be employed to solve the linearized Poisson-Boltzmann equation. In applications involving biomolecules the finite difference approach is the most widely adopted [130, 131]. This is the approach which is adopted within the AMBER software suite [70]. This method involves imposing a 3D grid onto the system with the atomic charges mapped onto the grid points (charges are distributed to neighbouring grid points weighted by their displacement), the grid points are also assigned a dielectric constant depending upon whether they lie within or outside the solute and boundary conditions are invoked at the edges of the grid (often $\phi(\mathbf{r}) = 0$). This reduces the original partial differential equation into a simple linear system of the matrix equation form $\mathbf{Ax} = \mathbf{b}$ in which $\mathbf{x}$ represents the unknown electrostatic potential at the grid points, $\mathbf{b}$ the charge distribution upon the grid points that are the source of $\mathbf{x}$ and $\mathbf{A}$, the coefficient matrix, combines the dielectric constant on the grid edges and any salt related terms. This form of the equation is then solved iteratively until the potential converges to within a predefined tolerance. In order to obtain the polar contribution to the solvation free energy the system must be solved twice, once with the solvent dielectric constant chosen to represent water and then again with a choice representative of vacuum (values of 80 and 1 are normally used respectively). The polar solvation energy, ΔG^{sol}_{pol}, is then given by:

$$\Delta G^{sol}_{pol} = \frac{1}{2}\sum_i q_i(\phi_i^{wat} - \phi_i^{vac}), \tag{3.49}$$

where q_i is the charge assigned to each point i on the finite difference grid and ϕ_i^{wat} and ϕ_i^{vac} the potential at the same points in water and vacuum respectively. This calculation must be performed for the complex, protein and ligand separately in order for the total ΔG^{sol}_{pol} to be determined (in line with Equation 3.40).

3.8.7 Solvent Accessible Surface Area

The non-polar contribution to the solvation energy, ΔG^{sol}_{nonpol}, can be thought of as being composed of two contributions: one due to the van der Waals interaction, ΔG^{vdw}_{nonpol}, and the other originating from the cost of creating the cavity in the solvent occupied by the solute, ΔG^{cav}_{nonpol}. The approximation made in MMPBSA is to assume, as most of the solvent reorganisation occurs in the first solvation shell around the protein and the van der Waals interaction is short ranged, that the non-polar solvation energy is linearly related to the solvent accessible surface area (SASA) of the protein, A:

$$\Delta G^{sol}_{nonpol} = \Delta G^{vdw}_{nonpol} + \Delta G^{cav}_{nonpol} = \gamma A + b \tag{3.50}$$

The constant of proportionality γ is often referred to as the surface tension and along with the constant b its value is determined empirically [119, 132]. It is likely, given their empirical nature, that these terms implicitly incorporate some purely quantum mechanical contributions to the solvation energy. For example it is known that experimental free energies include a contribution, which is always positive, due to the polarisation of the electronic wave function in response to the change from a gas phase to condensed phase environment [133, 134]. As is the case with the other components of the MMPBSA calculation the SASA computation must be performed for the complex, protein and ligand separately in order for the total ΔG^{sol}_{nonpol} to be determined (in line with Equation 3.40).

3.8.8 Calculating The Configurational Entropy Using Normal Modes

The cavity term in the non-polar solvation term of the MMPBSA calculation provides an estimate of the entropic changes associated with the insertion of a solute into the solvent. However, no account is made for the entropic impact of changes in the configurational freedom of the enzyme and ligand upon complex formation *in vacuo*. In general, protein-ligand binding events cause restrictions to the number of conformations available to both and consequently a reduction in entropy; this contribution is known as the configurational (or conformational) entropy. This results in a free energy penalty to binding, which can be included into a calculation of the absolute binding affinity,ΔG_b, using Equation 3.51, where ΔG_b^{MMPBSA} is the MMPBSA estimate, T is the temperature and ΔS_{conf} the configurational entropy.

$$\Delta G_b = \Delta G_b^{MMPBSA} - T\Delta S_{conf}. \tag{3.51}$$

The change in entropy is calculated from the difference between the values estimated for the complex and the separate enzyme and ligand contributions (as shown in Equation 3.53). In cases in which relative binding affinities alone are under investigation then the entropic contribution can be neglected for cases in which very similar systems are under comparison. To obtain absolute binding affinities or to treat a wider range of systems it must be calculated. A variety of approaches have been used with the most commonly invoked in biological systems being normal mode and quasi-harmonic analyses.

The configurational entropy comprises of three components associated with translational, rotational and vibrational motions. These are summed to give the overall contribution:

$$S_{conf} = S_{conf}^{tra} + S_{conf}^{rot} + S_{conf}^{vib}. \tag{3.52}$$

The enzyme and ligand have 3 rotational, 3 translational and $3N$-6 vibrational degrees of freedom (where N is the number of particle in the system) which can be impacted by binding into a complex. Like the MMPBSA contribution to ΔG the change in the configurational entropy is calculated from the difference between the values calculated for the complex and free enzyme and ligand:

$$\Delta S_{conf} = S_{conf}^{complex} - (S_{conf}^{enzyme} + S_{conf}^{ligand}). \tag{3.53}$$

The entropic changes associated with the contributions shown in Equation 3.52 are well described from statistical mechanics [135–137] considerations by:

$$S_{conf}^{tra} = \frac{3}{2}RT - RT\left[\frac{5}{2} + \frac{3}{2}\ln\left(\frac{2\pi m k_B T}{h^2}\right) - \ln(\rho)\right] \tag{3.54a}$$

$$S_{conf}^{rot} = \frac{3}{2}RT - RT\left[\frac{3}{2} + \frac{1}{2}\ln\left(\pi I_A I_B I_C\right) + \frac{3}{2}\ln\left(\frac{8\pi^2 k_B T}{h^2}\right) - \ln(\sigma)\right] \tag{3.54b}$$

$$S_{conf}^{vib} = \sum_{i=1}^{3N-6}\left[\frac{1}{2}h\nu_i + \frac{h\nu_i}{e^{h\nu_i/k_BT}}\right] - \sum_{i=1}^{3N-6}\left[\frac{h\nu_i}{e^{h\nu_i/k_BT}} - RT\ln\left(1 - e^{h\nu_i/k_BT}\right)\right] \tag{3.54c}$$

where ρ is the number density at 1 mol L^{-1}, m is the mass, I_A, I_A and I_A the principal moments of inertia and σ the symmetry factor. The S_{conf}^{vib} contribution depends on the normal modes ν_i of the molecule. Consequently normal mode analysis has become a popular way of estimating the change in entropy upon binding. Normal modes describe the concerted motions of the constituent atoms of the system under study, which are assumed to behave harmonically close to an energy minimum. The higher the frequency of a normal mode the smaller the amplitude. This makes normal modes useful for separating the slow, global, motions of the protein from local vibrations (for example those of hydrogen atoms).

The assumption made in normal mode analysis that the oscillations of the protein occur within a single minimum is a major limitation of the method due to the rugged nature of free energy landscape of real proteins. This means that normal mode analysis can be ineffective for systems which undergo large structural changes. This has led to the use in some circumstances of alternative approaches such as quasi-harmonic analysis [135, 138] which allow sampling over several minima, although there is evidence that it overestimates vibrational entropy in such systems [139].

3.9 Conclusion

In this chapter, the thermodynamics which govern the binding of proteins and ligands have been discussed and a brief outline of the statistical mechanical considerations, which allow us to link these considerations to the interactions of the microscopic constituents of the systems under study, presented. The property which allows us to quantify the strength of protein and ligand association was identified as the free energy of binding. A discussion of a variety of methods by which the binding free energy can be estimated, both experimentally and from molecular simulations, was presented. Particular attention has been given to the description of the MMPBSA and normal mode methodology, as this is the approach used to approximate the values for HIV enzymes binding inhibitory drugs in the rest of this thesis.

Chapter 4

HIV & AIDS

4.1 Introduction

Acquired Immunodeficiency Syndrome (AIDS) was the only major epidemic of the 20th century to be caused by a previously unknown infectious agent. The disease is characterised by the susceptibility of sufferers to opportunistic pathogens and increased risk of Kaposi's sarcoma and other rare forms of cancer. Since its identification in 1981 AIDS has been responsible for the deaths of more than 25 million people with 2.8 million loosing their lives in 2005 alone [140], while 38.6 million people are currently infected with Human Immunodeficiency Virus (HIV), the causative agent of AIDS [141]. As these vast numbers suggest HIV/AIDS has become a global pandemic and is found through out the world.

As its name suggests, HIV primarily infects vital components of the human immune system including CD4+ T cells (a sub-group of lymphocytes, a type of white blood cell, which express the surface protein CD4) and macrophages (a variety of white blood cell which absorb and then digest both pathogens and cellular debris). One of the primary symptoms of HIV infection is the loss of CD4+ T cells. When enough T cells have been destroyed by HIV the immune system can no longer fully perform its function. This loss of immune system response is the cause of AIDS.

HIV has itself been further classified into two groups HIV-1 and HIV-2. HIV-2 is endemic in West Africa (and has begun to spread into India) but most AIDS worldwide is caused by the more virulent HIV-1 [140]. For this reason most studies of HIV (including this thesis) concentrate on the HIV-1 subtype.

As the worldwide AIDS epidemic continues, a cure of HIV-1 remains elusive. In the absence of a cure, research has focused on suppressing viral replication. Modern treatment

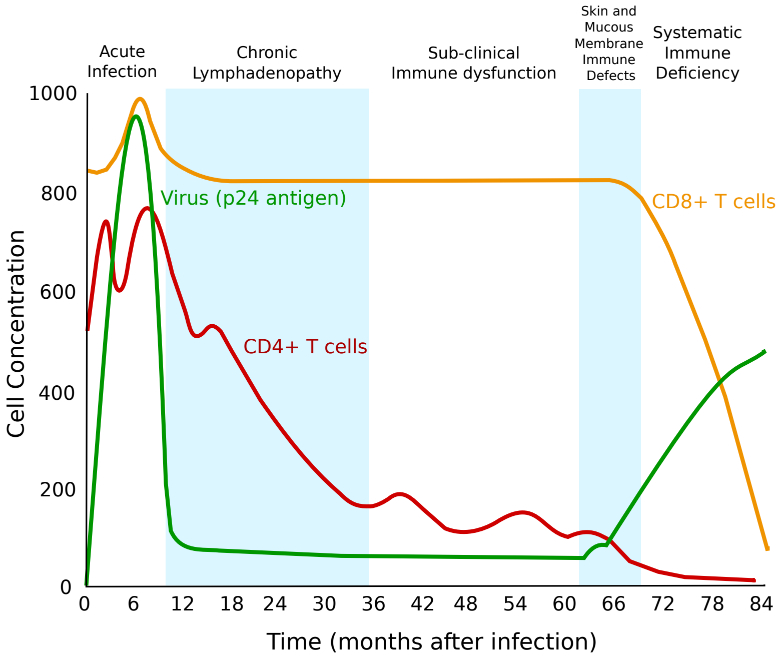


Figure 4.1: The time course of the populations of CD4+ and CD8+ T-cells and the viral load over the course of a clinical HIV infection. An initial peak in both CD4+ cells and viral load is observed, the viral load then stabilises while the CD4+ cell are gradually depleted. Eventually the CD4+ cell population declines terminally and the viral population explodes causing full blown AIDS. Adapted from [143]

strategies (discussed in Section 4.5) have achieved considerable successes, indeed for those with access to modern combination therapy HIV infection has been transformed from a 'death sentence' to a controllable disease which requires early diagnosis and life long treatment. However, the efficacy of treatment is compromised by the emergence of drug resistant mutant strains of the virus [142].

The development of effective treatments has been built upon an understanding not only of the clinical impact of infection but also of the replicative process of HIV. What follows is a description of the clinical progression of the HIV infection, a description of the life cycle of the virus and the methods of treatment currently in clinical use.

4.2 The Clinical Course of HIV Infection

Infection by HIV is characterised by complex interactions with the host and a chronic course of disease. From the point of initial infection there is usually a period of between 8 and 12 years before the patient experiences the chronic stage of infection known as AIDS [144, 145]. Disease progression happens in several phases, as indicated in Figure 4.1 and described below.

Initial infection with HIV-1 is often associated with an illness referred to as ‘acute retroviral syndrome’ or ‘primary HIV infection’. The syndrome exhibits many similar symptoms to the flu: infected individuals commonly experience fevers, sore throats, swollen lymph nodes and rashes [146, 147]. These symptoms usually subside within 1 to 2 weeks as this phase of the disease is self limiting (many HIV positive people remain asymptomatic during these early stages of infection or at least report no significant symptoms) [146, 148].

This early stage of the infection leads to the loss of mucosal CD4+ T helper lymphocytes [149]. These are the main targets of HIV, although the virus can infect several other cell types including macrophages [150, 151]. It is the loss of CD4+ lymphocytes which brings about the suppression of the immune system, resulting in AIDS. The main function of the T helper cell is to regulate immune responses by the secretion of specialised factors that activate other white blood cells to fight infections [152]. In particular they control CD8+ lymphocytes which are responsible for directly killing certain tumour cells, cells infected by viruses and some parasites.

Significant declines in the levels of CD4+ cells occur within the first 2 to 8 weeks of infection [149]. The initial fall in the number of CD4+ cells is accompanied by a rise in HIV-specific CD8+ killer cells which target infected CD4+ cells [150, 153]. Levels of both, however, quickly return to close to their pre-infection values Figure 4.1. During this period, although most CD4+ cells remain uninfected, there is a high level of viral replication in the peripheral blood system [154].

The body launches a strong immune defense to the initial high levels of virus, with infected CD4+ cells rapidly being eliminated. The initial results of this response are a lowering in the amount of viral particles in circulation and the temporary recovery of CD4+ cell levels [143]. The destruction of infected cells is balanced by the body’s production of new CD4+ cells and a steady state, in which most CD4+ cells are uninfected, is attained [148].

CD4+ cells have two states, one activated and the other a resting memory state [155, 156]. While most of the infected cells are destroyed some survive long enough to revert to the quiescent state. In this state CD4+ cells no longer express viral antigens and hence cannot be recognised as infected by the immune system [143]. Such quiescent cells can persist for many years, still carrying the viral genome, only to be reactivated at a later time, thus acting as a latent viral reservoir.

The result of this is that, although the virus continues to replicate and persists throughout the body, the patient enters a period of clinical latency [144, 145]. This period regularly lasts for 10 years or more during which time the number of CD4+ cells gradually reduces. This decrease can partly be accounted for by the ingestion of infected cells by CD8+ killer T cells and the rupturing of infected cells as millions of viral particles bud out of the host cell. However, fewer than 1 in 100,000 CD4+ T cells in the blood of AIDS patients are actually infected with the virus [157]. The loss of other cells can be explained by the fact that infected CD4+ cells can induce apoptosis (death) in uninfected cells [158, 159].

Eventually, CD4+ levels reach a point where they can no longer perform their vital signalling role in the immune system and the patient's immune response drops precipitously with an accompanying increase in viral load (again shown in Figure 4.1) [143]. Once this stage in the infection has been reached normally benign opportunistic pathogens are able to infect the patient. It is this catastrophic loss of immune function which is termed AIDS.

4.3 The Structure And Genome Of HIV

HIV belongs to the class of viruses known as retroviruses. These are enveloped viruses, possessing an RNA genome and rely on the enzyme reverse transcriptase (RT) to convert it into DNA [160]. This DNA copy of the viral genome can then be integrated into the chromosomal DNA of a target cell [161]. All retroviral genomes contain three coding regions - *gag* (group specific antigen), *pol* (polymerase) and *env* (envelope) - which always occur in this order in the genome and contain the information for the structural proteins and enzymes necessary for replication in all retroviruses [162]. All three of these genes encode multiple proteins, which perform vital roles in virion production and the viral lifecycle, the names and functions of which are given in Table 4.1. The *gag* and *env* genes are transcribed into separate Gag and Env polyproteins but *pol* is only read in combination with *gag* (utilising a shift in reading frames) to form Gag-Pol. The polyprotein chains are later cleaved to form functional proteins. HIV is further subclassified as a lentivirus, a class of retrovirus which share the characteristics of having high pathogenicity, complex genomes and long incubation periods (the name is derived from *lenti-* the Latin for "slow") [162, 163]. The genomes of lentiviruses (such as HIV, Simian Immunodeficiency Virus (SIV) and Feline Immunodeficiency Virus (FIV)) contain regulatory genes in addition to the *gag*, *pol* and *env* found in all retroviruses which are thought to be responsible for their increased pathogenicity [162]. The regulatory genes included in the HIV genome include; *tat*, *rev*, *nef*, *vif*, *vpr* and *vpu* [162, 164]. The regulatory genes can be divided into two categories. Transactivator genes are responsible for changing host gene expression and are essential for viral replication *in* vitro.

Table 4.1: Descriptions of all protein products of the *gag*, *pol* and *env* genes. All information in this table taken from [164, 165]

Chain	Protein	Description
Gag	MA (p17)	Mysrylated matrix protein. Stabilizes the viral particle by remaining attached to the inner surface of the virion lipid bilayer after viral maturation.
	CA (p24)	Core antigen capsid protein. Forms the conical core of the viral particle.
	p6	Mediates interactions between Gag and Vpr, leading to the incorporation of Vpr into assembling virions.
	NC (p7)	Nucleocapsid protein. Binds to the HIV packaging signal on viral RNA and is sufficient for incorporation of RNA into virions.
Pol	PR (p11)	Protease enzyme. Catalyses the proteolytic cleavage of the polypeptide chains into functional proteins.
	RT (p51/p66)	Reverse transcriptase enzyme. Forms a heterodimer with two active sites which reverse transcribes viral RNA into DNA.
	IN (p34)	Integrase enzyme. Mediates the insertion of viral DNA into the host genome.
Env	SU (gp120)	Surface glycoprotein. Acts as a receptor on the surface of the viral particle, and binds to CD4 and secondary receptors on macrophages and T lymphocytes.
	TM (gp41)	Transmembrane glycoprotein. Transverses the virions lipid membrane and along with SU is involved in viral entry to the host cell.

Accessory genes are those which have not been found necessary for the production of viable virions *in vitro* [164].

Retroviral genomes, such as that of HIV, are encoded in RNA. The HIV genome is carried in two identical, 9.2 Kb (kilobase), single strands of RNA (ssRNA) [161]. In order to be integrated into a host cell's genome these strands must be translated into DNA in a process known as reverse transcription (see Section 4.4.3). The DNA produced by the reverse transcription process is known as the proviral genome. As well as transcoding the HIV genome into DNA, reverse transcription replicates some regions at the 3′-end of the HIV genome at the 5′-end of the viral DNA (see Figure 4.2) creating long terminal repeats (LTRs) at either end of the genome [162].

Figure 4.3 shows the organisation of these genes in the proviral genome. As this picture shows, the information carried within the HIV genome is increased by having some regions code for more than one product. This is achieved by a process known as frameshifting [162, 163].

The structural arrangement of the gene products in an assembled virion is shown schematically in Figure 4.4. Each virion contains two copies of the ssRNA genome contained within a nucleus along with PR, RT and IN. The nucleus is a bullet shaped

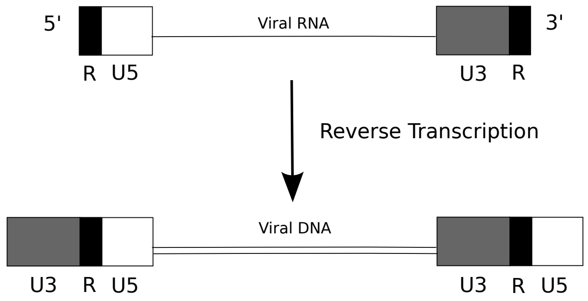


Figure 4.2: During reverse transcription of the viral genome short sections at the 3′-end and 5′-end of the viral genome, known as unique 3′ (U3) and unique 5′ (U5), along with a short repeated sequence (R) are duplicated. The resultant DNA is longer than the transcribed RNA. This is the origin of the long terminal repeats (LTRs). Adapted from [162]

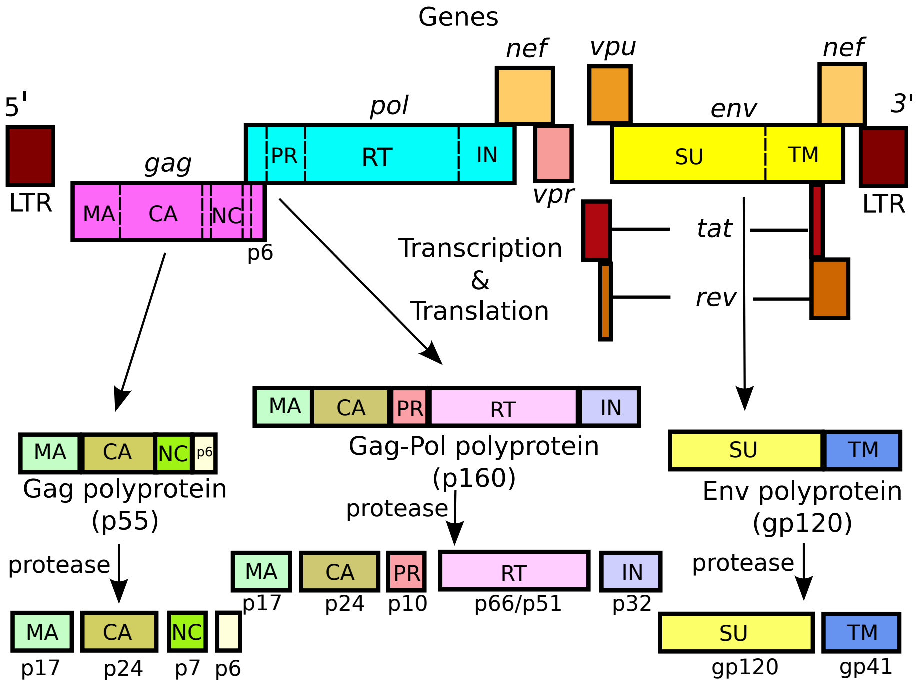


Figure 4.3: The HIV genome. The positions of the LTRs and genes which code for proteins (such as those listed in Table 4.1 and the regulatory genes *tat*, *rev*, *nef*, *vif*, *vpr* and *vpu*) are shown. Where two sections of the genome forming one product are separated (such *tat* and *rev*) this indicates that splicing is required to form the full RNA. Adapted from [166]

structured formed from CA protein units. The nucleus itself is contained first within a casing made of MA matrix protein and then a host cell derived lipid bilayer. The membrane bilayer is transversed by trimeric units of TM non covalently bonded to trimers of SU on the surface [161, 162]. The overall virion is a near spherical structure.

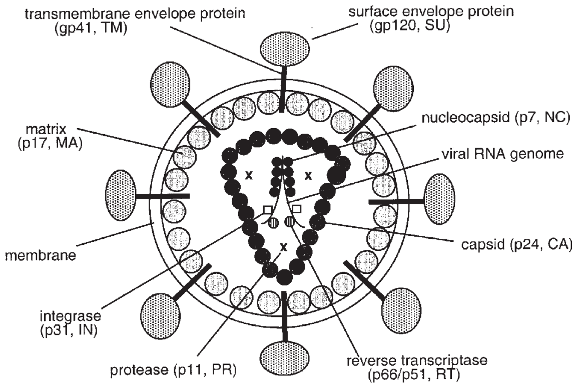


Figure 4.4: A schematic representation of the locations of the various proteins contained within an HIV-1 virion. This figure has been removed due to copyright restrictions but is available in Freed [161]

4.4 Life Cycle

A major factor in facilitating the development of anti-HIV treatment strategies has been the detailed understanding of the viral life cycle. Here we give an overview of how the virus reproduces and the roles played by several of the proteins described in the previous section within this process.

4.4.1 Virus Entry

The infection of a target cell is initiated by the binding of the envelope glycoprotein SU (also known as gp120), expressed on the virion surface, to CD4+ receptors on the surface of the target cell [167]. SU forms the surface exposed element of the viral envelope assemblage. This assemblage consists of a trimer of SU non-covalently linked to a trimer of the transmembrane protein TM (also known as gp41) [168–170]. CD4+ binding causes an alteration in the conformation of SU which exposes a chemokine receptor binding surface, facilitating the interaction of the envelope proteins with coreceptors (usually the chemokine receptors CXCR4 or CCR5) also found on the target cells surface [171–173]. The formation of the ternary complex between SU and the coreceptor is believed to lead to changes in the conformation of TM. The organization of many such complexes at the fusion site allows formation of a fusion pore [174]. The creation of the fusion pore initiates the membrane fusion reaction between the lipid bilayers of the viral envelope and the target cell plasma membrane. The contents of the virion, including the RNA genome and viral enzymes, are then released into the host cell's cytoplasm [161].

4.4.2 Post-Entry Events

The events which follow membrane fusion remain some of the least well characterised in the HIV life cycle. The first step that needs to occur is a process known as uncoating. In this process the core of the virion (defined as the structural remains of the virion after the membrane has been lost) is reconfigured to create a complex known as the reverse transcription complex (RTC), which later in the cycle is converted into the preintegration complex (PIC) [161]. During these steps most, if not all, of the viral capsid (CA) protein is lost, while at least some of the matrix (MA) and nucleocapsid (NC) proteins are used to form the RTC and PIC along with the viral enzymes reverse transcriptase (RT) and integrase (IN), and the regulatory protein Vpr [175].

4.4.3 Reverse Transcription

Once the viral core is within the host cell, the next key stage in the life cycle is conversion of the single stranded viral RNA genome into double stranded DNA which can be incorporated into the host cell chromosomes [161]. This task is performed by the RT enzyme. The enzyme uses the single stranded RNA viral genome as a template to create a single strand of DNA which is in turn used as a template to create a double stranded DNA (*ds*DNA) copy of the genome [161]. The *ds*DNA copy is then suitable for integration into the chromosomes of human host cells. HIV-1 RT is a multifunctional enzyme with distinct polymerase and RNaseH active sites [161, 162]. At the polymerase active site incoming nucleotides matching the template RNA or DNA are incorporated into the growing complementary DNA chain (see Figure 4.5). The RNaseH active site catalyses the breakdown of the RNA genome, freeing the DNA copy to act as a template for the creation of the final double stranded DNA genome. The copying process is known as reverse transcription (a more detailed description of the steps within the process is given in Appendix B).

The reverse transcription process contains a number of events (known as strand transfers) in which the enzyme must change template. One example of this is the switch from the original viral RNA template to the ssDNA copy. As a consequence of the need to change template, the interaction between the reverse transcriptase and the template is of relatively low affinity [176] and template switching is frequent. If the two RNA genomes in the original virion are not identical (or if the host cell is infected by multiple virions of varying genomic make up) this can lead to the creation of novel recombinant genomes [177]. The reverse transcription process in HIV replication also has a very low fidelity with an *in vivo* mutation rate of $3{\times}10^{-5}$ per base pair per replicative cycle [178]. In comparison to other RTs that of HIV is 10 to 100 times more likely to incorporate

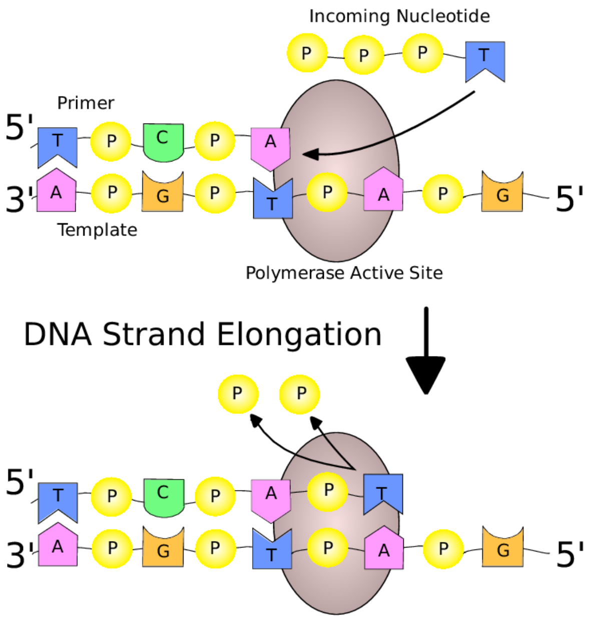


Figure 4.5: The reverse transcriptase catalyses the incorporation of nucleotides into the $3'$ end of a DNA chain, causing it to be elongated in the $5'$-$3'$ direction. The reaction pairs the incoming nucleoside triphosphate with the complementary base in a template strand, releasing two phosphate molecules in the process. Initially the template is the viral RNA genome. As it is copied the RNA strand is degraded at the RNaseH active site. Once transcription of the RNA genome is complete then the resultant single strand of DNA is used as the template to create a double strand of DNA capable of being incorporated into the host genome.

errors into DNA [179, 180]. It is the high frequency of recombination allied to the low replicative fidelity and high level of viral production (approximately 10^8 to 10^9 are produced per day [181]) which accounts for the high level of genetic diversity of the viral populations in a single patient. The generation of such diversity allows for the rapid selection of resistant strains when the virus encounters an environment in which anti-HIV drugs are present if replication suppression is incomplete [182].

4.4.4 Nuclear Import and Integration

The next stage of the viral life cycle requires the DNA genome to be transported to the host cell nucleus for integration into the chromosomal DNA. In order to achieve this the RTC (with which the genome is associated throughout reverse transcription) is converted into the PIC [161]. As mentioned above the PIC is known to contain MA, NC and accessory proteins including Vpr. Early models suggested that MA was the main signal responsible for nuclear import [183]; however, it is now believed that this is not the case, with Vpr being the most important viral factor. Vpr does not contain a viral import signal itself but is believed to attach the PIC to the cellular import machinery [184].

Once the PIC has entered the nucleus, the viral enzyme IN catalyses the insertion of the viral DNA into the host cell's chromosomal DNA. First IN processes the $3'$ termini of both strands of the viral dsDNA, resulting in a DNA duplex with staggered ends. It then creates a staggered cleavage in the host DNA, inserting the viral DNA into the gap created in the host genome [161]. Integration is accompanied by the duplication of a short sequence from the target site, typically 5 base pairs long [162]. This process results in an intermediate with gaps flanking the inserted element. Repair enzymes, from the host cell, then fill these gaps, joining the host and viral DNA [185]. Once the viral DNA is integrated it is known as the "provirus" and effectively acts like a cellular gene.

4.4.5 Gene Expression

The integrated provirus provides a template which the host cell translates into RNA. In fact it actually encodes more than 30 RNAs including the HIV genome [162, 186]. These different RNA transcripts are created by splicing of the complete genome product. In order to form the full range of of products some RNAs must be doubly spliced. This presents a challenge for HIV as most cellular protein coding RNAs (known as messanger or mRNAs) are only exported to the cytoplasm for translation when fully spliced. To overcome this potential problem HIV uses the viral Rev protein. This binds to a *cis*-acting RNA element called the Rev responsive element (RRE) [187], which is located in the *env* gene and is present in all unspliced and partially spliced genes. Over time Rev forms a multimer around the RRE which results in a complex capable of binding to the cellular export machinery. Once in the cytoplasm the RNAs can be translated in the same way as cellular mRNAs to produce proteins.

4.4.6 Virus Particle Production

The next stage of the lifecycle is the assembly of the Gag and Gag-Pol chains. It is thought that the Gag chain is responsible for the formation of multimers with Gag-Pol which then recruit a copy of the viral RNA genome. This complex is then transported to the cell membrane assisted by host factors and cellular machinery [188]. The assembled protein complex attaches to the inner membrane of the host cell via covalent bonds involving the myristic acid moiety of the Gag chains [189]. The complex now induces curvature of the host cell membrane, which leads to the formation of a bud. The new virion is formed as a section of the host membrane breaks away from the host cell.

Table 4.2: List of all ARVs approved by the FDA for the treatment of HIV broken down by drug class. The classes are: protease inhibitors (PIs), nucleoside analogue reverse transcriptase inhibitors (NRTIs), non-nucleoside analogue reverse transcriptase inhibitors (NNRTIs), integrase inhibitors (INIs) and fusion inhibitors (FIs).

PI	**NRTI**	**NNRTI**	**INI**	**FI**
Amprenavir	Abacavir	Delavirdine	Raltegravir	Enfuvirtide
Atazanavir	Didanosine	Efavirenz		Maraviroc
Darunavir	Emtricitabine	Etravirine		
Indinavir	Lamivudine	Nevirapine		
Lopinavir	Stavudine			
Nelnavir	Tenofovir			
Ritonavir	Zalcitabine			
Saquinavir	Zidovudine			
Tipranavir				

4.4.7 Maturation

At this point although a new virion is formed it is not infectious. This is because the essential viral proteins remain inactive within the Gag and Gag-Pol chains. The last step in the HIV lifecycle is the maturation of the virus. It is at this stage that the viral protease plays its vital role by cleaving the polypeptide chains into active enzymes (this process includes extracting itself from the Gag-Pol chain) [182, 190]. If the protease does not perform it's role correctly then the virion will not be able to infect another cell, and consequently the virus will not be able to replicate further [191].

4.5 HIV-1 Drug Treatment

Understanding of the life cycle of HIV has led to the development of a variety of drugs targeted at specific steps in the viral reproductive process. The first drug developed was Zidovudine, targetted at the reverse transcriptase, which was approved for clinical use in 1987 [192]. As more drugs were developed it became apparent that combining several drugs, which are active against different target enzymes, was the most effective way to treat HIV infection [193]. This treatment approach is often referred to as highly active antiretroviral therapy (HAART). HAART based combination therapy is usually capable of reducing viral loads to undetectable levels [193, 194] and where available has extended life expectancy to 21.5 years [195].

Currently, 24 antiretroviral drugs (ARVs) have been approved by the US Food and Drug Administration (FDA)[1] for the treatment of HIV. These drugs are usually separated

[1] FDA:www.fda.gov

into five classes: protease inhibitors (PIs), nucleoside analogue reverse transcriptase inhibitors (NRTIs), non-nucleoside analogue reverse transcriptase inhibitors (NNRTIs), integrase inhibitors (INIs) and fusion inhibitors (FIs) [196]. A list of all currently approved inhibitors broken down by class is presented in Table 4.2. The different classes of ARV affect separate steps in the HIV life cycle and most HAART cocktails feature drugs intended to inhibit at least two different targets [194]. Most initial (known as 'first line') drug cocktails comprise two NRTIs and one NNRTI or PI. Other types of inhibitor are usually used as part of so-called 'salvage regimes' after failure of previous treatment selections [194].

PIs and NRTIs are competitive inhibitors and will be described further in Section 4.6.4 and Section 4.7.4.1 respectively. NNRTIs bind to the RT altering its conformation and preventing the enzyme from correctly performing its DNA polymerase function [197]. Further details of this class of drug are provided in Section 4.7.4.2

INIs are designed to inhibit the IN enzyme and prevent the incorporation of the HIV provirus into the host DNA [198]. INIs are recent additions to the library of clinically available ARVs, with Raltegravir, the first INI to gain FDA approval, in use only from 2007 [199]. Despite the relatively short period of clinical use, mutations at three locations within the IN enzyme (at positions 143, 148 and 155) [200] have been strongly linked to Raltegravir resistance and other mutations suggested to have some impact [201]. A further drug, known as Elvitegravir is currently undergoing phase 3 clinical trials [202].

FIs operate by preventing viral entry via the CD4 receptor and CCRC5 coreceptor. The two FDA approved FIs operate in different ways: Enfuvirtide binds to TM and prevents conformational changes required in order to create an entry pore for the viral capsid into the target cell [203], whereas Maraviroc blocks binding of the viral envelope protein, SU, to CCR5 [204]. In the case of Maraviroc a test is required to determine the tropism of the virus infecting the patient being treated (the virus may make use of either the CCR5 or CXCR4 coreceptor) as it is only effective in blocking entry using the CCR5 coreceptor [205].

4.5.1 Drug Resistance

As mentioned in Section 4.4.3, the low fidelity and high throughput of HIV RT results in acquisition of mutations in the viral genome. Most mutations will have an adverse impact on enzymatic function. However, in the presence of ARVs a trade off between reduced efficacy and the ability to evade inhibition allows some mutations, which induce reduced affinity to drugs, to gain an evolutionary advantage [206]. Hence, if viral replication is not complete then these strains will come to predominate within a patient and the

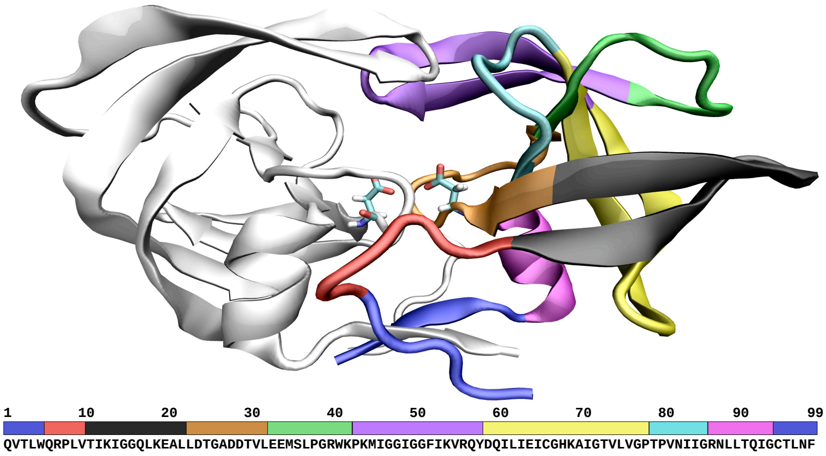

1
10
20
30
40
50
60
70
80
90
99
QVTLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF

Figure 4.6: The structure of the HIV-1 protease dimer is shown in cartoon representation. The structural elements identified in the 'bulldog's face' description of the structure are shown highlighted in the right hand monomer with the 'whiskers' in blue, 'nose' in red, 'cheek turn' in black, 'eyes' in brown, 'ears' in green, 'flaps' in purple, 'cheek sheet' in yellow, 'wall turn' in cyan and the final helix in pink. The catalytic dyad is depicted in chemical structure. The HXB2 wildtype protease sequence is also shown with the positions within it of each of the structural elements indicated using a colour bar.

treatment employed will no longer be effective. This is a particular problem in situations in which monotherapy is used [182].

In order to ensure that a drug regimen will work for the HIV strain present in a particular patient it is now routine for viral sequences to be taken upon diagnosis and treatment failure [207, 208]. The relationship between resistance causing mutations are complicated and clinical decision support systems, based on statistical analysis of data collected from patient databases and the published literature, are used to assess the likely susceptibility to particular drugs [209–213]. The focus of this thesis is on investigating the possibility of extending these existing systems through the use of predictive modelling to assess the impact of mutations upon drug binding. This topic will be discussed in more detail in Chapter 6.

4.6 Protease Structure, Function and Inhibition

The HIV-1 protease is a homodimer, meaning that it is formed from two identical monomeric subunits. Each monomer is a protein chain 99 amino acid residues in length [214]. The complete protease structure exhibits approximate rotational symmetry through the dimer interface. In cases when it is necessary to distinguish residues of the two monomers that form the homodimer, the two chains are conventionally numbered from 1 to 99 and 101 to 199 respectively. A great deal of effort has been expended

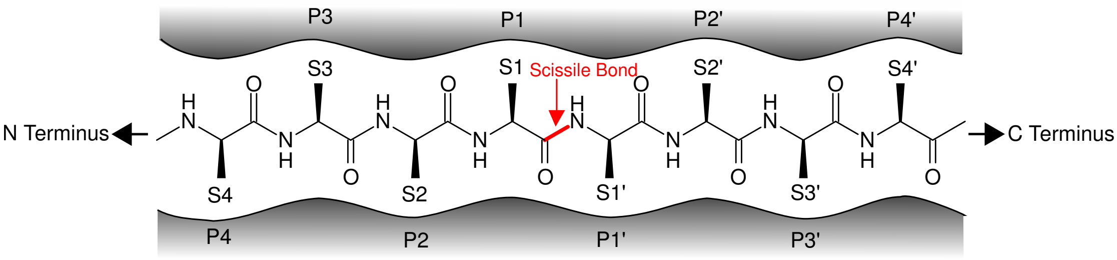


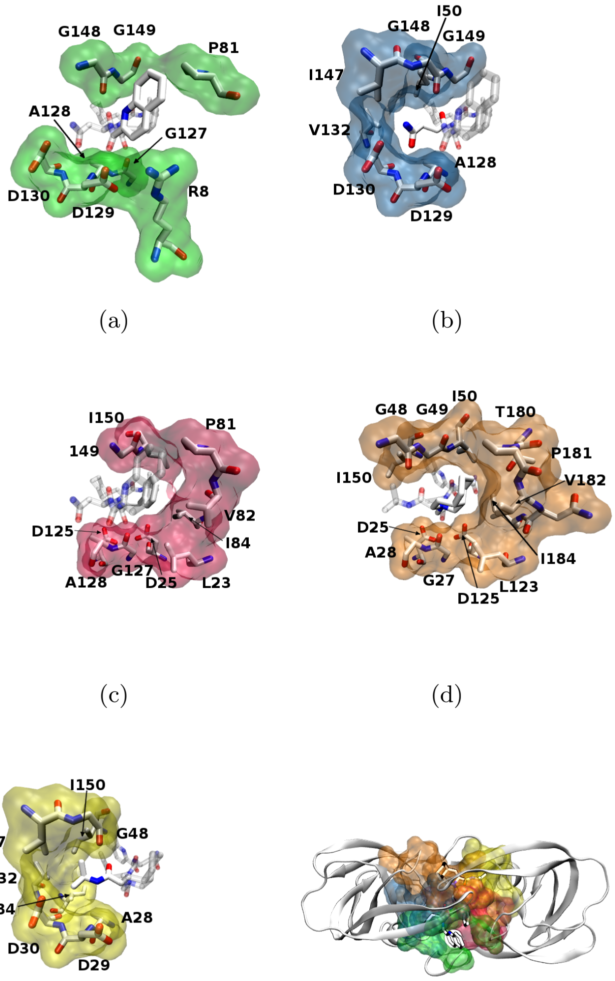
Figure 4.7: Schematic representation of the location of substrate residues in the protease active site. Towards the C-terminus of the substrate peptide chain they are labelled S1′, S2′, S3′ ... and S1, S2, S3 ... towards the N-terminus. The complementary protein pockets retain the numbering of the substrate position but are denoted with a P.

in investigating the structure of HIV-1 PR and 342 crystal structures are available in the PDB at the time of writing.

In order to describe the structure of the protease, we will be adopting the terminology set out by Perryman *et al.* [215]. According to this naming scheme the various sections of the protease structure are labelled according to the resemblance of the protease backbone to a bulldogs face (see Figure 4.6). The active site of the protease is situated in a cleft, which is covered by the region known as the 'flaps' (residues 43-58). The flaps are connected to the 'ears' (residues 33-42), the catalytic ASP containing 'eyes' (residues 23-32, which form the base of the active site), the 'cheek sheet' (residues 59-78), the 'wall turn' (residues 79-85), the 'cheek turn' (residues 11-22), the 'nose' (residues 6-10) and the 'whiskers' (residues 1-5 and 95-99).

The two dimers interface at three points; the whiskers, the eyes and the flaps. The whiskers are the terminal domain, in which the N and C termini of each subunit interlock to form a compact four-stranded β sheet. The interlocked structure is crucial for the formation and stability of the active protease. The whiskers connect to the rest of the protease via the wall turn, which terminates at a helix formed by residues 86 to 94 and a turn encompassing residues 4 to 9.

The core domain is made up of the eye and cheek sections of the protease chains. The eye section of each monomer is not only involved in the dimer interface but also includes the triad D25-T26-G27. This motif is responsible for the cleavage of the Gag and Gag-Pol polypeptides and is consequently referred to as the active site. Despite the infidelity of the HIV reverse transcriptase, the active site structure is highly conserved over generations as it is essential to the enzymes function. The D-T-G motif is common to a wide range of retroviral proteases, with the T and G residues implicated in both correct alignment of the catalytic Asp and dimer stability [216–218]. The pair of D25 residues (one from each monomer) is often referred to as the 'catalytic dyad'.



(a) (b)

(c) (d)

(e) (f)

Figure 4.8: The amino acid composure of the subsites of the protease subsites in the saquinavir bound 1HXB crystal structure. The amino acids that form the binding pocket and the substrate sidechains which fit into them are shown explicitly for the (a) S3/P3, (b) S2/P2, (c) S1/P1, (d) S1′/P1′ and (e) S2′/P2′ subsites. The position of all of the subsites within the overall PR structure is shown in (f).

The interface between the core domain and the wall turn is mostly made up of small, hydrophobic residues, while the ear section is a mostly solvent exposed loop, which precedes the flaps. The flaps not only form part of the dimer interface but also play a crucial role in substrate capture. In order for the catalytic processing of a substrate to occur it must pass into the active site containing the catalytic Asp dyad. The entry of the substrate into this area is controlled by the flaps, which form a flexible gate for an approaching ligand [219, 220].

When a substrate is bound the protease forms a closed conformation. In this state either side of the active site predominantly hydrophobic pockets are formed. These pockets are made up of residues 23, 50, 81, 82 and 84 from one monomer and residues 28, 32 and 48 from the other and interact with the sidechains of any bound substrate. Despite the hydrophobic nature of the residues, substrate side chains with varying chemical character can be found within these pockets [162]. The natural substrates of the protease are cleavage sites within an extended polypeptide chain and a labelling system is required in order to specify the location of specific interactions. Conventionally, the positions of the substrate amino acids are numbered from the cleaved peptide bond (often known as the scissile bond), towards the C-terminus of the substrate peptide chain they are given a labelled S1′, S2′, S3′ ... and S1, S2, S3 ... towards the N-terminus. The pockets in the protease structure into which they fit are correspondingly names P1′ - P4′ and P1 - P4 as shown in Figure 4.7. The precise protease residues involved in each pocket varies depending on the ligand bound. The same naming convention is adhered to for inhibitors as is used for natural substrates. The composition of the pockets in the 1HXB crystal structure [221] (which is bound to the inhibitor saquinavir) is shown in Figure 4.8.

4.6.1 Flexibility and Conformations

NMR and crystal structure evidence suggest that the PR flaps are highly flexible and it is intuitive that they would play a role in the recognition and binding of the natural substrate and inhibitory drugs [222–224]. All ligand bound proteases have shown the flaps to be overlapping and closed over the active site (this conformation can be seen in Figure 4.9) in a conformation that is relatively insensitive to the particular ligand [224]. This contrasts with NMR, molecular dynamics and crystal structure evidence that suggests that the apo enzyme exists in an ensemble of more open states where the structure freely interconverts between the closed, open and an intermediate semi-open state [220, 222–225]. Recent NMR data indicates that the predominant conformation in this ensemble involves weak interactions between the flaps [226] suggesting that it is dominated by semi-open forms of the enzyme similar to those seen in some crystal structures (such as that shown in Figure 4.9). While any such definition is inevitably some what arbitrary, the 1HHP crystal structure has been used in previous studies to define the semi-open conformation [225, 227]. Figure 4.9 shows that accompanying the opening of the flaps is a change in the curling of the flap tips. In the closed conformation the flap tips are seen to curl into one another and in the semi-open conformation the flaps cross over with the tips curling away from one another. This change in relative orientation is known as handedness, with the closed structure said to display *cis* handedness and conversely the semi-open structure is said to have *trans* handedness. Unfortunately, only one crystal structure of the fully open protease exists (the 1TW7 PDB structure) and it has been shown that this is stabilised by crystal packing effects [228].

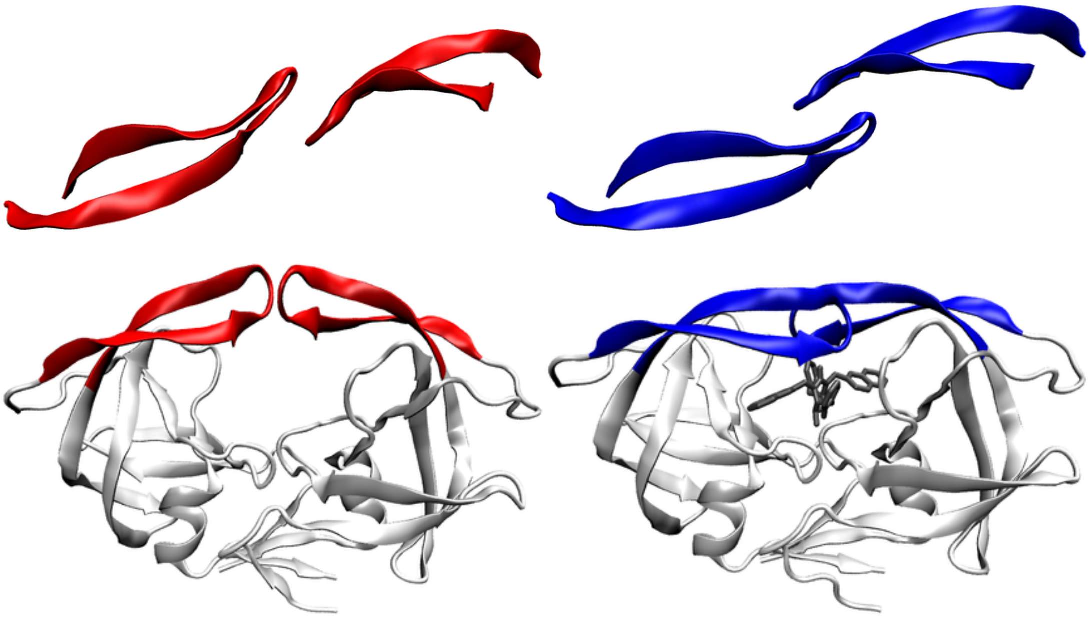


Figure 4.9: The two different protease conformations identified by crystallography. On the left the apo PR is seen in the semi-open form (based on PDB 1HHP), whilst on the right is the the inhibitor bound PR (based on PDB 1HVR) in the closed conformation. Above the main figures the flaps region is pictured from above, showing the change in handedness that accompanies flap opening. The semi-open structure exhibits *trans* handedness where the flaps cross over, whereas the closed conformation (seen in all substrate bound structures) does not and is said to have *cis* handedness.

Flap dynamics is an attractive area for study via molecular simulation. The free energy penalty of changing the conformation of the protease flaps from semi-open to closed upon ligand binding has been investigated, with recent studies suggesting a change of 2.4 ± 0.4 kcal mol^{-1} [229]. The model of ligand access to the active site being mediated by the flaps has also been given credence by molecular dynamics simulations, which indicate that when a ligand is introduced to the semi-open apo enzyme it closes [225] and when an inhibitor is removed from a closed structure it moves to a semi-open one [230]. In addition, coarse-grained, Brownian dynamics models have shown that the flaps act act as a gate controlling substrate entry to the active site [231, 232]. This type of simulation distinguished between the binding pathway of long polypeptide chains, where the substrate must enter from the top of the PR, inducing full flap opening, and shorter ligands, such as typical inhibitors, which can enter directly from the ends of the substrate binding cleft.

4.6.2 Catalytic Mechanism

The exact nature of the mechanism involved in the cleavage of peptide bonds by the HIV-1 protease remains a matter of debate. Mutational studies have, however, shown that the aspartic acid residues in position 25 play a vital role in enzyme function [191, 233] and the HIV PR is only active in dimeric form [234, 235]. Experimental evidence has long suggested an acid-base mechanism and indicated that a contribution is made by a lytic water molecule in the catalytic process [236, 237]. Many different reaction pathways have now been suggested, most of which posit that the active site Asp dyad activates a water molecule which then acts as a nucleophile and attacks the carbonyl carbon of the scissile bond [236–240]. Crystallographic studies in conditions which greatly reduce catalytic activity suggest that the reaction proceeds via a tetrahedral intermediate as shown in Figure 4.10 [241, 242]. In this mechanism one of the catalytic aspartic acids must be negatively charged (i.e. unprotonated) in order to activate the water molecule bound between the catalytic dyad [243].

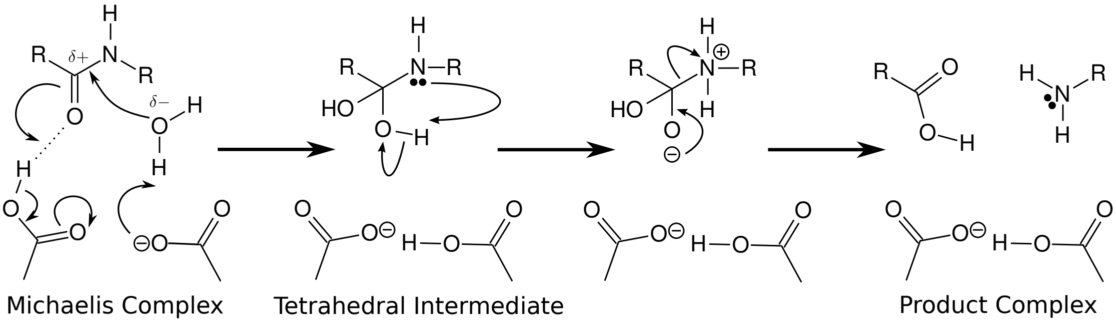


Figure 4.10: Schematic diagram of the HIV-1 PR reaction mechanism based on recently acquired crystal structure snapshots (PDBs 3MIM and 2NPH). The initial Michaelis complex is converted via a tetrahedral intermediate into the product complex. Adapted from from [242].

The potential impact on the catalytic function of the aspartyl dyad protonation state has prompted a range of studies of this property [243]. Four possible states are available dianionic (D-), diprotonated (D25/D125), position 25 protonated (D25) and position 125 protonated (D125). Aspartic proteases function over a wide range of pH values (2 to 7.4) [244]. An experimental study of how HIV protease activity varies with changing pH by Hyland *et al.* [236] produced a bell shaped profile, centred on pH 5. These results also suggested that substrates bind only to a form of HIV-1 protease in which one of the two catalytic aspartyl residues is protonated. In contrast NMR studies of the ^{1}3C enriched apo enzyme at pH 6 suggest that the dianionic, D- state, is prevalent. NMR evidence, from studies with inhibitors bound, suggests that the protonation state depends on the character of the ligand. Results for a symmetric inhibitor [245] indicate a diprotonated state, whereas those for an asymmetric one indicate monoprotonation [246]. A high level of dependence on local chemical environment or ligand induced structural distortion is also suggested by a wide range of computational studies [243, 247–250].

4.6.3 Substrate Processing

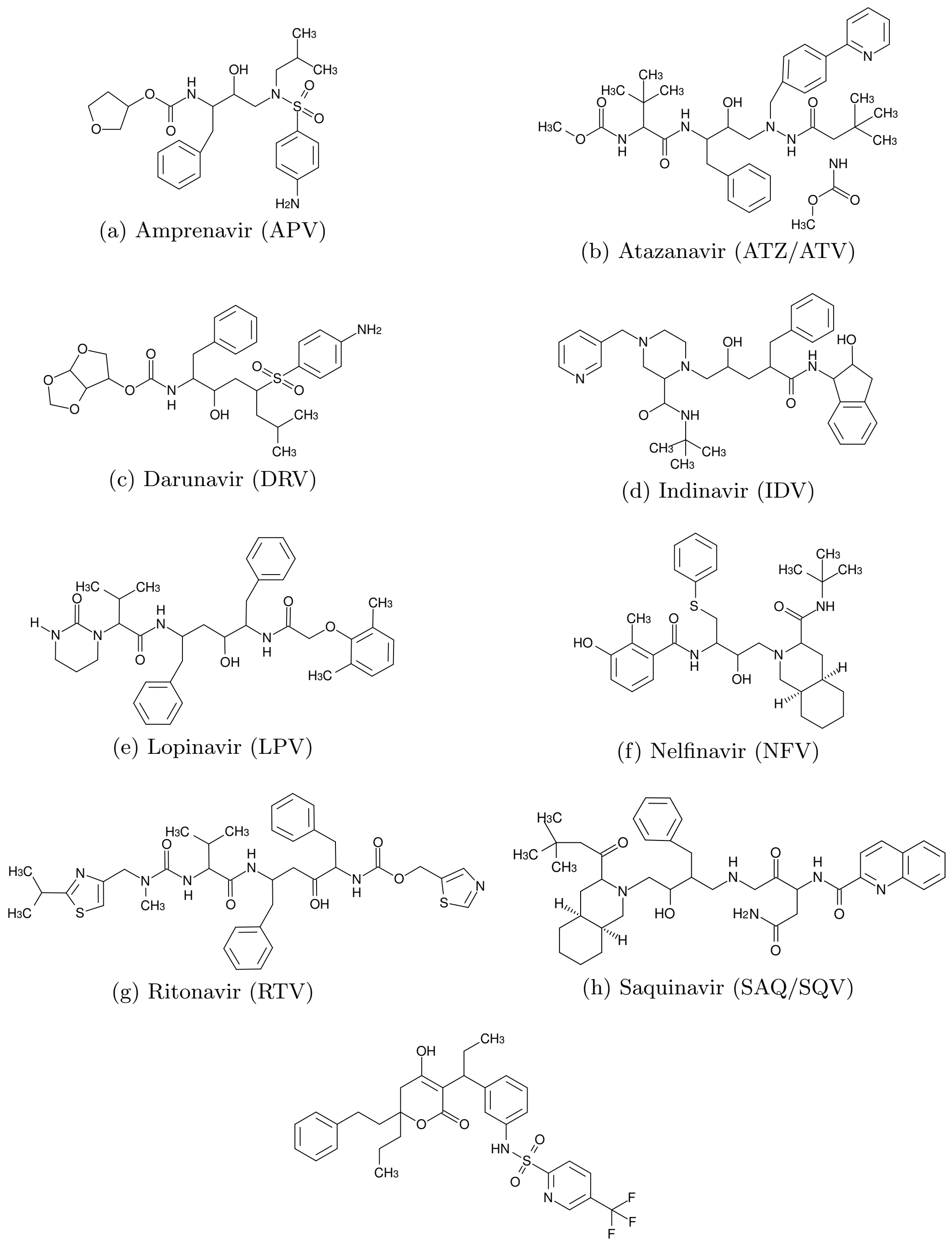
A fully detailed description of the process by which PR cleaves Gag and Gag-Pol into functional proteins has yet to emerge, however, it is clear that in order to perform it's function HIV-1 PR must recognise specific sites in the polyprotein chains where cleavage is necessary. Cleavage sites differ in their amino acid composition, but HIV protease most efficiently cleaves peptide substrates seven amino acids long (running for S4 to S3′) [162, 251, 252]. Whilst the cleavage sites show significant variation some properties are well conserved with P1 and P1′ largely hydrophobic and asparagine found in P2 for four of the sequences. These two positions are thought to play significant roles in substrate specificity although prediction of cleavage sites remains an active area of research [253–255].

Both *in vitro* and tissue culture studies have been used to elucidate the maturation process. These studies strongly suggest that the processing of the various cleavage sites in the Gag and Gag-Pol precursors occurs sequentially and is tightly regulated, with mature proteins formed as products of primary, secondary and tertiary lytic events [234, 256, 257].

4.6.4 Inhibitory Drugs

As a consequence of the quantity of available crystal structures, protease has become a canonical example of structure assisted drug design [216, 258]. Currently, nine protease inhibitors (PIs) are approved for clinical use by the FDA and they play a key role in many recommended HAART drug cocktails [194]. The structures of all of these drugs are shown in Figure 4.11.

All licensed PIs act competitively and a general principle of their design has been to mimic the natural peptide substrate of PR but with the cleavable bond replaced by an uncleavable hydroxyethylene moiety (with the exception of tipranavir, which is based on a coumarin scaffold and was discovered using high throughput screening) [192, 196, 216, 258]. This moiety is known to bind to the catalytic dyad and the side chain groups of peptidomimetic inhibitors are conventionally labelled using the same convention as the natural peptide substrate (as illustrated in Figure 4.7). As might be expected, given the commonalities of their design, the clinically used PIs all bind to PR in a broadly similar fashion. In all cases a water molecule (referred to as WAT301) is bound above the inhibitor, tetrahedrally hydrogen-bonded to oxygen molecules either side of the hydroxyethylene moiety of the drug and the backbone nitrogens of residues 50 and 150 of the PR. This observation has inspired an alternative design strategy, based around



(a) Amprenavir (APV)

(b) Atazanavir (ATZ/ATV)

(c) Darunavir (DRV)

(d) Indinavir (IDV)

(e) Lopinavir (LPV)

(f) Nelfinavir (NFV)

(g) Ritonavir (RTV)

(h) Saquinavir (SAQ/SQV)

(i) Tipranavir (TPV)

Figure 4.11: Chemical structures of each of the nine FDA approved HIV protease inhibitors.

central cyclic urea groups, in which sections of the inhibitor aim to displace this water molecule [259]. As yet no inhibitor based on this design principle has been approved for clinical use.

Saquinavir was the first approved inhibitor and, like other early inhibitors, achieved a binding affinity in the sub-nM range to wild type protease. However, the emergence of

Table 4.3: Binding affinity values for all FDA approved HIV-1 protease inhibitors with the wild type protease. The first six values are taken from Ohtaka *et al.* [92] and were all produced using the same experimental conditions. The remaining values were obtained from the BindingDB database [88] and were produced under a variety of conditions.

Inhibitor	ΔG **(kcal mol^{-1})**	K_i **(nM)**
APV	-13.2	0.20
IDV	-12.4	0.76
LPV	-15.1	0.008
NFV	-12.8	0.44
RTV	-13.7	0.098
SQV	-13.2	0.28
ATZ	-13.2	0.48
DRV	-14.8	0.014
TPV	-15.1	0.014

resistant mutations (discussed in Section 4.6.5) has forced the continuing development of inhibitors with recent inhibitors such as lopinavir and tipranivir achieving sub-pM potencies [196]. A comparison of the binding affinities of all nine FDA approved inhibitors is given in Table 4.3.

Increasing inhibitor potency is obviously desirable, but is not the only factor determining the usefulness of a drug *in vivo*. Factors such as solubility can have a significant impact on the ability of an inhibitor to reach it's intended target. Administered alone, lopinavir suffers from low bioavailability (a measure of the fraction of the administered drug that reaches the circulatory system) and so is combined with sub-therapeutic doses of ritonavir, which increase the quantities of lopinavir reaching the bloodstream [260]. The addition of ritonavir to other PIs to gain the same benefits is now commonplace, and combined PI tablets are known as 'boosted' PIs [261]. Further to the issue of bioavailability, it is vital that inhibitors are highly selective in order to reduce side effects induced by binding to non-target human proteins. Despite efforts to minimise them in the design and trial stages of drug development all of the available PIs are associated with some level of toxicity.

4.6.5 Drug Resistance in Protease

The structure of PR is able to tolerate mutations in up to 50% of the amino acids in its 99 residue sequence and remain functional [224]. This observation is in line with the finding that monomers of HIV-1 and HIV-2 protease are functionally interchangeable in the dimeric enzymes, despite differing by between 45 and 50 mutations [262]. Such high levels of mutational robustness suggest that many mutations have little, or no, effect on catalytic function. Competition between multiple viral strains ensures that only highly efficient phenotypes survive and some sections of the protein, such as the catalytic Asp containing 'firemans grip' motif, are highly conserved [263]. Studies of HIV-1 sequences

Figure 4.12: The mutations associated with PI resistance in clinical studies. The locations within the protein sequence are given within the horizontal bar with the wild type amino acid above and the mutant form(s) beneath. This figure has been removed due to copyright restrictions bit is available in [200].

taken from patients have identified 17 polymorphisms[2] and 37 locations which are rarely seen to mutate [264].

It is unsurprising that many mutations that leave the protease functional alter its specificity. The use of PIs introduces a selective pressure, which may give an evolutionary benefit to enzymes which are more likely to bind natural substrate than an inhibitor even if this comes at the cost of reduced catalytic efficacy. This phenomenon is the basis for the emergence of resistance protease mutants. Many clinical studies have been undertaken to characterise mutations that are associated with resistance [200, 264–266]. Figure 4.12 shows a summary of the key mutations associated with PI treatment failure for all of the FDA approved inhibitors. Mutations have arisen for all of the currently licensed PIs. However, the most worrying development has been the emergence of viral strains which exhibit cross class resistance, known as multi-drug resistant (MDR) viruses [267, 268]. MDR viruses have limited the effectiveness of salvage therapy in which one PI is replaced by another once resistance is acquired.

The primary cause of resistance is the reduction in binding affinity to inhibitors. For resistance associated mutations which occur in the PR active site (such as V82A and I84V) this can intuitively be understood as resulting from direct alteration of the contacts between drug and protein. Other mutations, such as L90M, are located far from the active site yet influence the binding of a range of inhibitors and their impact is much harder to explain. At least a partial explanation for this observation is that most resistant PR sequences contain multiple mutations [264, 269] with several studies indicating that protease accumulates mutations in a ordered fashion under selective pressure from PIs [270–273]. Resistance associated PR mutations are frequently divided into two classes; primary mutations which are highly correlated with treatment failure, and accessory mutations which either only exhibit resistance when many mutations are combined or enhance the resistance of existing primary mutations. Sometimes as many as 7 or 8 accessory mutations are observed to accrue in resistant sequences.

Studies of the positioning of many resistance associated mutations have suggested that PIs that fit within the overlapping consensus van der Waals volume of the natural substrates are less likely to be impacted by their presence. The hypothesis is that mutations impacting such inhibitors would simultaneously reduce the processing of the

[2] Polymorphisms are locations at which more than one type of amino acid coexists in a population without dependence on specific selective pressure.

substrates [274–276]. The inhibitor darunavir is designed to fit snuggly within this envelope and has been observed to present a higher genetic barrier to the development of resistance and higher efficacy against multi-drug resistant HIV relative to other protease inhibitors [277].

Commonly, resistant protease sequences involve point mutations and double mutations, such as G48V/L90M which arises after saquinavir treatment [278] and V82F/I84V in response to ritonavir [279], which may evoke more than thousand fold reductions in binding affinity. In other cases similar reductions in affinity occur through the combination of several mutations which do not alone induce significant energetic changes. It is becoming increasingly recognised that super additive combinatorial effects may lead to significant levels of resistance [92, 269, 280]. It is however more common that accessory mutations act to enhance the effects caused by existing primary mutations [281, 282].

In many cases the mechanism by which mutations away from the active site cause reduced inhibitor binding is unclear. The difficulty of kinetic experiments makes this a fertile ground for molecular simulation. One area that has been extensively studied is the impact of mutations upon flap dynamics [220, 225, 229, 232]. For example, studies have suggested that M46I stabalises the closed PR conformation [283] and that V82F/I84V induces greater sampling of the semi-open flap positions [284]. MD studies by Foulkes-Murzycki *et al.* [285] have also suggested that 19 residues, which form a hydrophobic core in PR[3], facilitate conformational changes in the flaps. These residues slide by one another with little energy penalty as they simply exchange one van der Waals contact for another as they move. They hypothesise that mutations to residues in this area (such as L90M) change the packing in the hydrophobic core, altering flexibility and hence the selectivity of PR. Other simulations have suggested that the G48V/L90M mutations may increase the accessibility of a lateral unbinding pathway for inhibitors [286]. MD simulations have also been widely used to directly calculate binding free energies of PR mutants [279, 287–289].

In some cases resistance-causing mutations can only be supported by PR in the presence of polymorphisms elsewhere. One example of this phenomena is the D30N mutation which is associated with resistance to nelfinavir. Alone D30N renders PR non-functional but the presence of N88D rescues catalytic activity [290]. This is an extreme case of what is known as a ‘compensatory’ mutation. This describes an accessory mutation which becomes fixed in the viral population as it reduces the fitness cost associated with associated resistance causing mutations [291, 292]. It is worth noting that *in vivo* viral fitness does not necessarily correlate directly with enzymatic efficacy, for example the

[3] The following positions are identified as forming the PR hydrophobic core: 5, 11, 13, 15, 22, 24, 33, 36, 38, 62, 64, 66, 75, 77, 85, 89, 90, 93 and 97

L90M mutation has been seen to improve protease activity [293, 294], but its absence in the wildtype suggests that overall it carries a fitness cost to the virus.

Some mutations that arise during treatment with one inhibitor may, in fact, cause the protease to be more susceptible than wild type to other drugs. Two examples of this are I47A, a rare LPV resistant mutation which enhances saquinavir binding [295], and I50L, which emerges in response to ATZ treatment and is hypersusceptible to other PIs[296].

Another form of resistance associated with PI treatment is the mutation of Gag precursor cleavage sites [297]. Mutations in the substrate are generally believed to act as compensatory mutations, restoring the fitness of the virus when primary resistance mutations in the PR alter substrate specificity [293, 294] (although recently this idea has been challenged [298]). An example is the A431V mutation in the P2 subsite of the NC/p1 cleavage site which is seen to emerge in virus populations containing the V82A mutation in response to treatment with ritonavir [299].

4.7 Reverse Transcriptase Structure and Inhibition

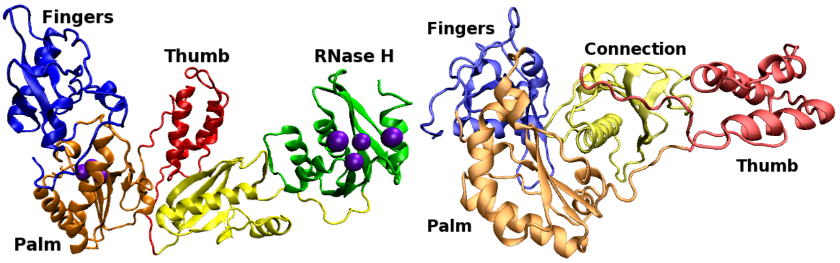
Active reverse transcriptase consists of an asymmetric heterodimer. This structure is created from a homodimer of two 66 kDa subunits, both subunits containing 560 residues. One of the subunits is subsequently proteolytically cleaved by the viral protease. This results in one 51 kDa subunit which is missing 120 C-terminal residues compared to the larger unit. The removal of these residues induces a considerable difference in the conformation of the two units, which are now referred to as p66 and p51 respectively. The p66/p51 heterodimer is resistant to further hydrolysis by the protease [300].

The p66 subunit contains both the polymerase and RNaseH active sites. The residues which were cleaved from the p51 subunit included the RNaseH domain and the residues which form the polymerase active site in p66 are buried in the RT structure and perform no catalytic function.

A large number of crystal structures of RT have been produced and the structure which has emerged has been likened to that of a right hand. The analogy has led to the naming of the subdomains as see in Figure 4.13. The template/primer duplexes bind in a large cleft between the fingers and thumb domains [301].

4.7.1 Active Sites

The polymerase and Ribonulease H (RnaseH) active sites are separated by a distance of 17 to 18 nucleotides of the template (approximately 60Å) [301]. The polymerase active



(a) (b)

Secondary Structure	p66	p51
Fingers (1–84)		
β0	7–12	7–12
β1	18–24	19–22
αA	28–44	28–44
β2	49–51	49–51
β3	56–63	56–63
β4	73–77	72–76
αB	78–83	78–84
Palm (85–119)		
β5a	86–90	87–90
β5b	94–96	94–96
β6	105–112	105–112
αC	114–117	112–115
Fingers (120–150)		
αD	122–127	122–127
β7	128–134	128–134
β8	141–147	141–147
Palm (151–243)		
αE	155–174	155–174
β9	178–183	179–183
β10	186–191	186–191
αF	195–212	198–212
β11a	214–217	214–219
β11b	219–222	–
β12	227–229	–
β13	232–235	–
β14	238–242	239–242
Thumb (244–322)		
αH	255–268	254–270
αI	278–286	277–283
αJ	298–311	289–310
β15	316–321	316–321
Connection (323–437)		
β16	326–333	325–333
β17	336–341	336–343
β18	350–358	350–358
αK	364–382	364–381
β19	388–391	386–392
αL	395–404	395–404
β20	406–412	410–416
β21	421–424	–
β22	427–430	–
RNaseH (438–560)		
$\beta1'$	438–447	
$\beta2'$	452–459	
$\beta3'$	462–470	
$\alpha A'$	474–488	
$\beta4'$	492–497	
$\alpha B'$	500–508	
$\alpha D'$	516–527	
$\beta5'$	530–536	
$\alpha E'$	544-555	

(c)

Figure 4.13: The RT subunits (a) p66 and (b) p51. The domains are named after the structures supposed likeness to a right hand. It is in fact, the folding of the p66 seen which results in the likeness. However the subdomains retain their name in p51 as seen in (b). The secondary structure elements and their position in the amino acid sequence of both subdomains are shown in (c).

site is located in the palm domain between the fingers and thumb subdomains with the RNaseH domain at the far end of the enzyme (see Figure 4.13 and Figure 4.14).

The polymerase active site consists of residues D110, D185, D186 all of which mutational studies have found to be essential for the enzyme to exhibit polymerase activity [302]. Further to this, studies which mutated these residues in p51 alone showed this had no effect on catalytic activity [303], placing the catalytic site in the β6-β9-β10 area of the palm subdomain of p66.

The RNaseH active site is responsible for the degradation of RNA/DNA and RNA/RNA duplexes. It consists of D443, E478, D498 and D549. Mutagenetic studies show that mutations of D443 and E478 [304, 305] result in the loss of catalytic function whereas changes to D498 destabilise the dimer [304].

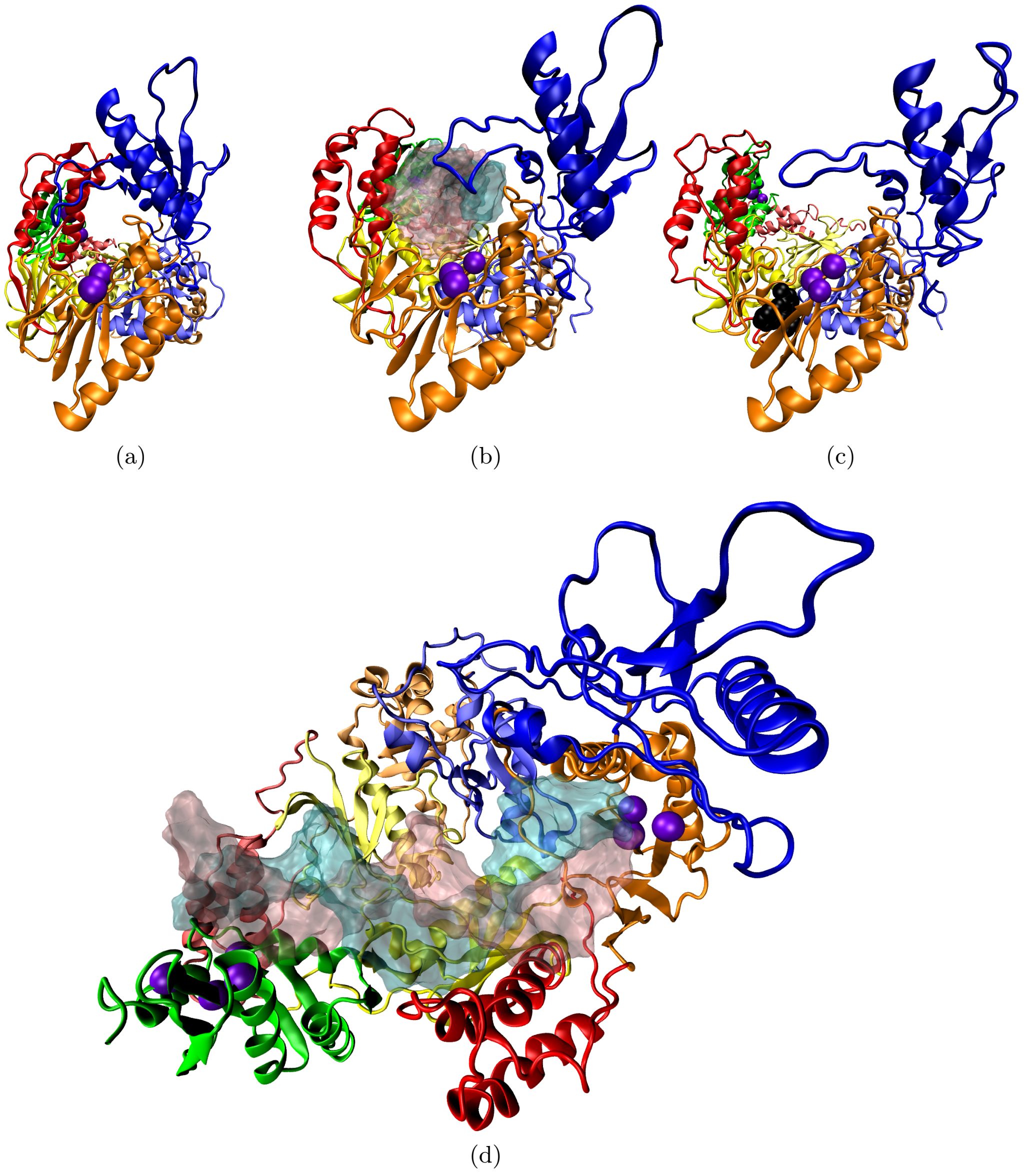
4.7.2 dNTP Binding Pocket

The crystal structure of RT in complex with an incoming dNTP shows that, upon binding of the dNTP, the loop between residues 60 and 75 of the fingers subdomain bends inwards towards the active site. In particular, residues K65 and R72 make contact with the incoming dNTP, forming salt bridges with the γ and α phosphates [306]. R72 also interacts with Q151 resulting in a flexible binding pocket which accommodates the $3'$ OH of the incoming dNTP. The rest of the binding pocket consists of the sidechains of A114, Y115 and the backbones of D113 and Y115. The interaction of R72 and Q151 is found to be template dependent. With a DNA template the amide nitrogen of Q151 interacts with the nucleotide base, where in the RNA template structures the amide oxygen is seen to interact with the first primer base and stabilises the side chain of R72 [306]. Additionally M184, which takes the X position in the YXDD motif conserved in all reverse transcriptases, is positioned close to both the $3'$ OH primer terminus and the bound dNTP [307].

Mutations at these locations have been seen to result in changes in the specificity of RT. M184V and Q151M do not significantly alter the error rate but change the type of errors made [308, 309], whereas Y115A results in a four fold decrease in transcription fidelity [310]. These results indicate that Tyr115 plays an important role in selecting the dNTP to be bound.

4.7.3 Global Conformational Variation

It is well established that crystallised RT structures show large scale conformational changes when bound to either a template/primer or NNRTI ligand [311]. Comparison



(a) (b) (c)

(d)

Figure 4.14: The structure of the reverse transcriptase enzyme is shown bound in apo form (based on the 1DLO PDB structure) and bound to both DNA and the NNRTI NVP (based on the 2HMI and 3HVT PDB structures respectively). The fingers domain is shown in blue, the palm in orange, the thumb red, the connection in grey and the RNaseH in green (the darker shades indicate the regions of the p66 subunit, lighter ones p51). The locations of the residues involved in the polymerase and RNaseH active sites are shown in purple. (a), (b) and (c) show the enzyme viewed along the DNA binding cleft for the apo enzyme and DNA and NVP bound systems respectively. The DNA strand is shown in cyan and pink, while the NNRTI NVP is shown in black, highlighting the location of the NNRTI binding pocket. (d) shows the entire enzyme bound to DNA from above the binding cleft.

of the apo and ligand bound structures in Figure 4.14 shows the movement of the thumb and fingers away from the binding cleft along the middle of the RT structure, a change which is often described using the analogy of an opening hand [312]. The apo structures 1JLE and 1RTJ both show RT in an open conformation, similar to that of the NNRTI or template/primer ligated structures; however, this can be explained by the fact that they were created by soaking out an NNRTI [313, 314]. For this reason, here we the analyse a subset of 92 RT crystal structures (details of the selected PDBs are provided in Appendix C) separated into three classes ignoring the open apo structures; the closed form unliganded, the NNRTI bound and the template/primer bound. In order to investigate the structural differences between the classes, average structures of the three conformations were created[4].

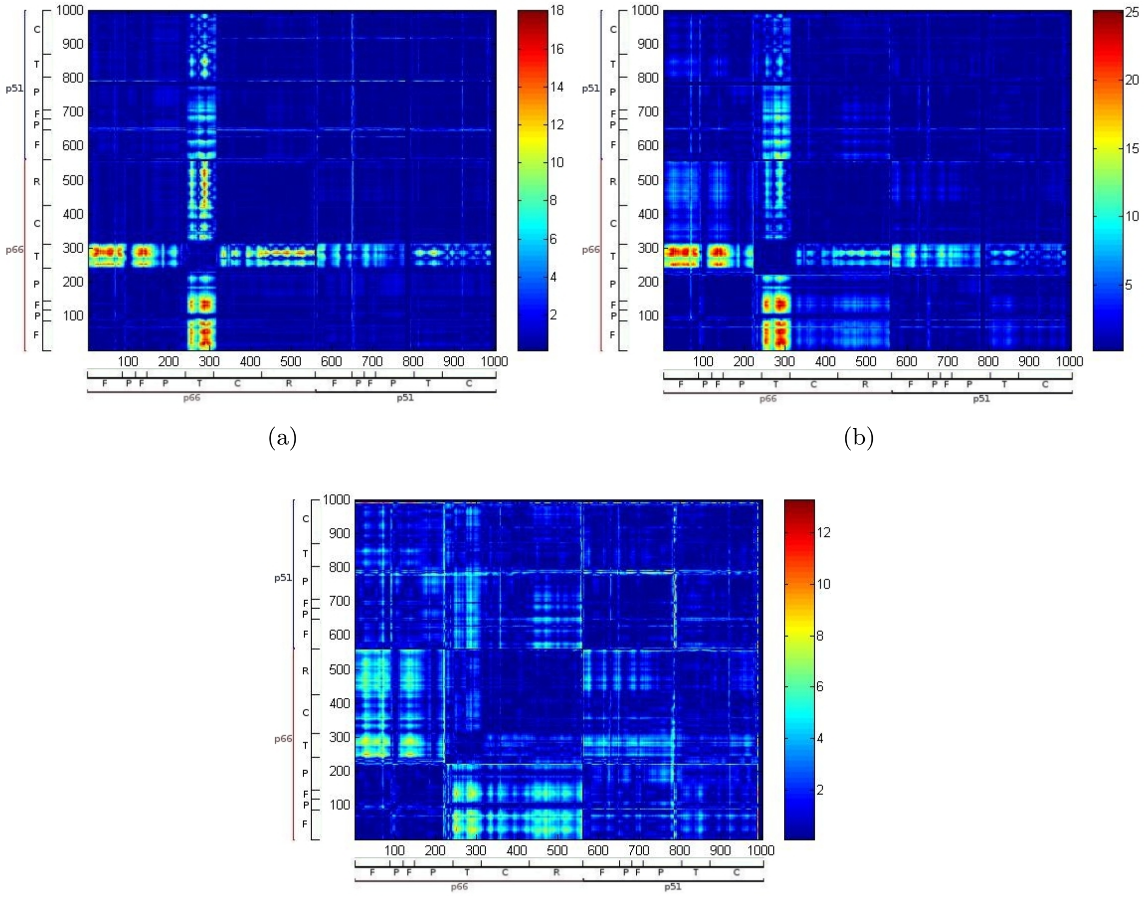
Information about differences between two structures can be gained by using difference distance matrices (DDMs). The first stage of the process of creating a DDM, is to create a difference matrix (DM) describing the conformation of each structure. A DM consists of a $N \times N$ matrix (where N is the number of residues in the structure) holding the distances between each pair of elements, in all of the matrices calculated here these will be the Cαs of each residue in RT. Using internal coordinates avoids the need to align the structures being compared. In order to calculate the DDM from a pair of DMs you simply need to subtract one from the other and take the magnitude of this change. Thus each element of the DDM is given by

$$D(i,j) = \left|\Delta r_{ij}^{A} - \Delta r_{ij}^{B}\right|, \tag{4.1}$$

where Δr_{ij}^{A} is the distance between the C_α of residues i and j in one structure and Δr_{ij}^{B} is the same distance in a different structure. Each element of the DDM represents the change in distance between a pair of residues [316] and hence higher values show where the greatest degree of conformational change has occurred.

The average structures were used to create the DDMs shown in Figure 4.15. These show that the dominant change in both liganded forms is the expected movement of the p66 thumb subdomain relative to all of the other subdomains. In the DNA bound form the thumb is the only section to show significant rearrangement with respect to the apo structure, whereas the NNRTI structure shows alteration in the p66 palm, connection and RNaseH subdomains as well. The differences between the two structures seen in Figure 4.15c are almost all in the p66 subunit beginning with the section of palm immediately preceding the thumb (which contains the NNRTIBP), with the biggest

[4] This was achieved by aligning all of the structures using VMD [315] and then taking the mean value of the coordinates for each residue.



(a)

(b)

(c)

Figure 4.15: DDMs showing the differences between average structures of differently ligated RT: (a) compares DNA bound and closed apo structures, (b) NNRTI bound to closed apo and (c) the NNRTI to the DNA bound. The colouring is normalised in each panel with the bar to the side indicating the scale used in each picture.

being in the distance between the thumb and the palm and fingers. This is consistent with the general observation that the thumb to fingers separation is greater in the NNRTI bound structures than the template/primer liganded.

In order to examine the variation of the separation between thumb and fingers subdomains exhibited in the different classes of structure, the distance between the C_α of W24 (located in the fingers, close to the thumb in the apo structure 1DLO) and K287 (the top of the thumb subdomain) was measured. Table4.4 confirms that the 'opening of the hand' is found throughout the available liganded structures. The average separation between the thumb and fingers subdomains of the NNRTI containing complexes is considerably larger (around 7 Å on average) than that of the template/primer liganded systems. In fact, the smallest separation seen in the NNRTI bound structures is still greater than the distance seen in any of the template bound systems.

Table 4.4: Distance in Å between the C_αs of residues 24 (in the fingers) and 287 (in the thumb)

Ligand Type	**Min.**	**Max.**	**Average**	**Std Dev.**
Unliganded	12.01	12.85	12.48	0.36
Template/Primer	30.09	36.15	34.54	2.33
NNRTI	37.43	46.44	41.58	2.61
Unliganded (open)	39.48	40.23	39.86	0.53

Table 4.5: The average angles between the thumb subdomain and αF which runs across the front of the palm and αK in the connection (in degrees) in different classes of crystal structures.

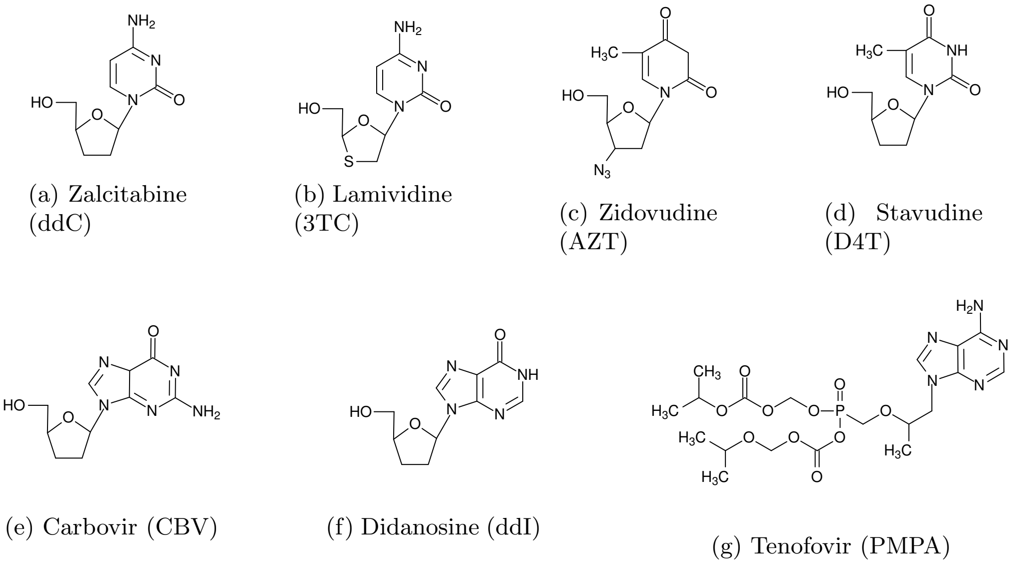
Ligand	**Thumb to αF**	**Thumb to αK**
Unliganded	32.75	115.34
Template/Primer	57.6	146.5
NNRTI	72.24	144.24
Unliganded (open)	71.64	145.11

The changes in the separation of the thumb and fingers are a result of a rotation of the thumb induced by ligand binding. The angles between αJ of the thumb and αF (which runs across the front of the active site between the base of the thumb and the base of the fingers) and αJ and αK (which is part of the connection and runs between the thumb and the RNaseH subdomains) both change (see Table 4.5). On average NNRTI binding induces a 40° increase in the αJ-αF angle from the unliganded form compared to 25° in the template/primer case. Both, however, show similar rotations of αJ relative to αK (of around 30°) when compared to the unliganded structures.

As described in Section 4.4.3, RT must bind both RNA/DNA and DNA/DNA template/primer duplexes in order to elongate the primer strand by adding additional nucleosides. In order to accommodate the template and primer the RT undergoes considerable structural rearrangement. The most obvious change is the positioning of the thumb subdomain, which rotates approximately 20° and consequently moves away from the fingers opening a binding cleft into which the template (and primer) can fit (see Figure 4.14). Only crystal structures of the unliganded RT are all in the closed conformation (1JLE [314] and 1RTJ [313] are discounted here as they were created by soaking out a weakly bound NNRTI) but a spin labelling study has shown that it exists in a temperature dependent equilibrium between open and closed states. At 273 K 65% of the population was found to be closed, rising to 95% at 313 K [317].

4.7.4 Inhibitory Drugs

Two classes of drugs have been developed which target the HIV-1 RT. Here we briefly describe their structure and method of operation. We focus upon the non-nucleoside



(a) Zalcitabine (ddC)

(b) Lamividine (3TC)

(c) Zidovudine (AZT)

(d) Stavudine (D4T)

(e) Carbovir (CBV)

(f) Didanosine (ddI)

(g) Tenofovir (PMPA)

Figure 4.16: Chemical structures of each of the 11 FDA approved NRTIs.

analogue inhibitor (NNRTI) class of drug as these are the subject of the simulations presented in Chapter 7.

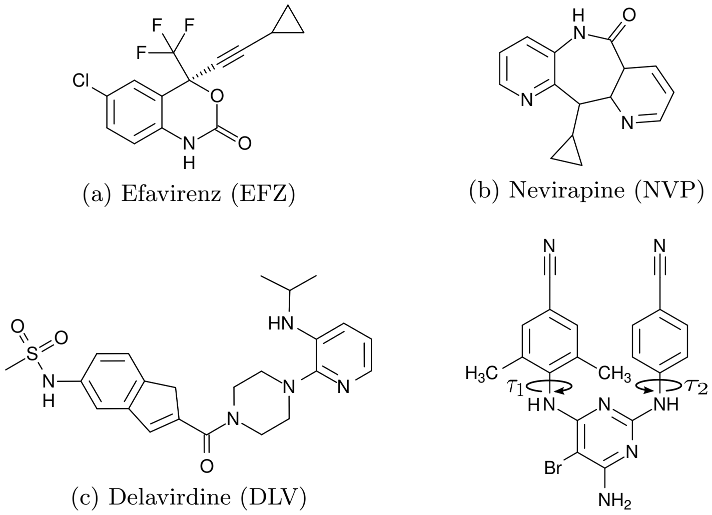
4.7.4.1 Nucleoside Analogue Inhibitors

As the name implies NRTIs compete with the natural dNTP substrate of RT, however, once incorporated into the nascent DNA chain they prevent further elongation (they are designed to have no 3′ OH) [318]. As expected crystal structures show that NRTIs bind to and interact with the residues of the dNTP binding pocket, with R72 in particular playing a role in stabilising the sugar moiety [318].

NRTIs are administered as pro-drugs, which need to be recognised and processed by cellular kinases to become active (this facilitates their penetration of the target call membrane). Nucleoside analogues must be tri-phosphorylated [319, 320], while nucleotide analogues need only be di-phosphorylated (as they already contain one phosphate group) [321]. Currently there are seven nucleoside and one nucleotide analogue approved by the US FDA, the structures of which are shown in Figure 4.16. Nucleotide analogues are an advance over the nucleoside analogues as they require cellular processing before they become active [318]

4.7.4.2 Non-Nucleoside Analogue Inhibitors

NNRTIs are, in general, small (less than 600 Da) hydrophobic compounds but have a diverse range of structures (Figure 4.17) [197]. Three have been licensed by the US FDA;



(a) Efavirenz (EFZ)

(b) Nevirapine (NVP)

(c) Delavirdine (DLV)

(d) Etravirine (ETV)

Figure 4.17: Chemical structures of each of the four FDA approved NNRTIs.

efavirenz (EFV), delavirdine (DLV) and Nevirapine (NVP). Delavirdine is not licensed in the UK and Etravirine is approved for treatment-experienced patients only in both the US and the UK [322]. NNRTIs bind in a pocket which does not exist in RT when there is no drug bound [323–325]. The so-called NNRTI binding pocket (NNRTIBP) is situated approximately 10 Å from the polymerase active site, between the β6-β10-β9 and β12-β13-β14 sheets[324, 326]. This is in the area where the thumb and palm subdomains are hinged. The pocket is hydrophobic in character and made up of L100, K101, K102, K103, V106, T107, V108, V179, Y181, Y188, V189, G190, F227, W229, L234 of p66 and E138 of p51 [323]. In the unliganded enzyme the sidechains of Y181 and Y188 fill the pocket but upon NNRTI binding they rotate away from the hydrophobic core creating space for the ligand [323, 324]. Twisting of the β12-β13-β14 sheet also expands the binding pocket [323]. Several entrances to the pocket have been proposed. The most commonly described is the area surrounded by K101, K103 and V179 near the p66/p51 interface. It has also been proposed that some NNRTIs may enter via an opening near P236 or from the active site region [327, 328].

Method of Inhibition

A number of theories have been proposed to explain the inhibition of reverse transcriptase by NNRTIs. Below is a short description of the main candidates.

NNRTI liganded crystal structures show short range deformation of the palm domain. In particular the YMDD motif containing two of the catalytic aspartates (residues 185 and 186) is distorted, altering the geometry of the polymerase active site [313]. It is proposed that the process of polymerisation is highly dependent on the alignment of the catalytic aspartates and this distortion of the active site prevents catalytic function. Similarly, the structures also show that the primer grip is also distorted and it is proposed that this may prevent the primer from being correctly aligned for catalysis [323].

Another hypothesis is that the NNRTIBP may disrupt the hinge between the thumb and palm subdomains resulting in a reduction in the mobility of the thumb (called the "arthritic thumb" model). It is contended that the motion of the thumb is essential to allow the translocation of the template/primer duplex to facilitate continuing DNA strand elongation [329].

As has already been noted the NNRTIBP is at the dimer interface, with p66 residues L100 K101, K103, V179 and Y181 along with p51 residue E138 being involved in both the interface and the NNRTIBP. Several experiments have shown that NNRTI binding affects the stability of the dimer, either increasing or decreasing it depending on the specific NNRTI [197]. It has been suggested that as dimerisation is essential for enzyme function [330], changes in dimer stability may prevent correct enzyme function.

There is no reason to assume that any of these explanations are exclusive and it may well be that a combination of factors contribute to inhibition. One study examined the steps of reverse transcription in the presence of NNRTIs and found that they blocked the polymerisation reaction but did not interfere with the binding of dNTPs [331]. Nevirapine is also known to alter the rate at which RT can flip orientation along a nucleic acid substrate[332], although it is not clear whether this is related to inhibition or a side effect of the structural changes induced by drug binding.

Molecular simulations have been employed to investigate the arthritic thumb model but the results have been inconclusive. A steered molecular dynamics study has demonstrated a reduction in the motion of the thumb [333], however another study, using a network model, disagreed. This second study compared the motions available to the unliganded RT with those available to RT with NNRTI bound, finding that the binding of an NNRTI did not reduce the flexibility of the thumb but did change the way it was correlated with the motion of the rest of the enzyme [47].

Binding Modes

The first three clinically approved NNRTIs were found by random screening but ETR [334, 335], along with more recent drug candidates, was rationally designed, with molecular level studies playing an important role. The first set of drugs are referred to as 'first generation' NNRTIs with the new drugs, which are usually found to be more potent than their predecessors, referred to as 'second generation' [197].

All of the clinically relevant, and several other, first generation inhibitors have been crystallised bound to RT. Examination of these structures reveals a common binding

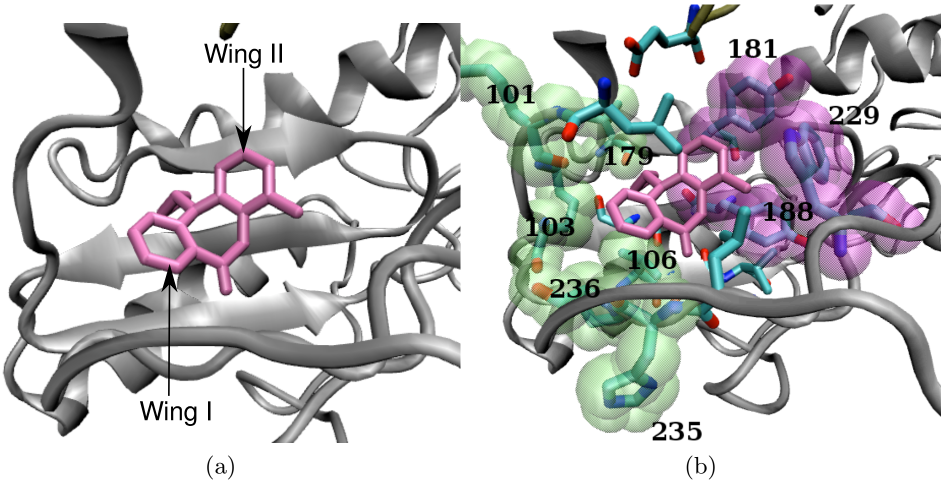


Figure 4.18: Nevirapine shown in the NNRTIBP in PDB 3HVT. a) shows the position of the drug resting on the $\beta6$-$\beta10$-$\beta9$ sheet. In b) the residues shown in green interact with Wing I of the drug those in purple with Wing II.

mode [197]. The appearance of the bound NNRTIs has frequently been likened to a butterfly. The drugs rest on the $\beta6$-$\beta10$-$\beta9$ sheet, as seen in Figure 4.18a. The head of the butterfly points down towards the hypothetical binding pocket entrance near residues 101 and 103 of p66 (from now on when describing residues within the NNRTIBP they should be assumed to be from p66 unless otherwise stated). Figure 4.18b shows an example of this type of binding by Nevirapine. Wing I of the butterfly interacts with the K101, K103, V106, V179 and Y318 sidechains (it may additionally interact with the backbone of H235 and P236). The body of the butterfly interacts with the backbone of residues Y188, Y189 and G190 and the sidechains of L100 and L134 (which also interact with both wings). Wing II has considerable hydrophobic contacts, interacting with Y181, Y188 and W229. There is one exception to this mode of binding in the first generation of NNRTIs and that is exhibited by DLV. DLV is considerably larger than the other NNRTIs and its elongated shape means that it extends out from the NNRTIBP and into the solvent surrounding it [336].

Etravirine has been designed to have a large degree of flexibility. It is one member of a class of drugs called DAPY derivatives, which have been rationally designed and show significant ability to rotate around the torsion angles, $\tau1$ and $\tau2$, shown in Figure 4.17d [334]. The very flexibility which is designed to allow it to retain activity against drug resistant RT variants has the side effect of preventing its crystallization in complex with wild type RT. A complex is available bound to the K103N resistant mutant and further insight into its binding conformation can be made by analogy with other DAPY derivatives. These studies along with molecular dynamics simulations indicate that the drug can adopt two different binding modes Figure 4.19, one more extended than the other. The more compact of these, called the “horseshoe” conformation, is almost U

Figure 4.19: (a) shows TMC120 in the "horseshoe" binding mode which modelling shows that Etravirine can also adopt. (b) shows Etravirine exhibiting a second binding mode closer to residues 100 and 103 with substantially different wing orientations relative to the body. This figure has been removed due to copyright restrictions but is available in Das *et al.* [334]

Figure 4.20: The mutations associated with NRTI resistance in clinical studies. The locations within the protein sequence are given within the horizontal bar with the wild type amino acid above and the mutant form(s) beneath. This figure has been removed due to copyright restrictions but is available in Johnson *et al.* [200].

Figure 4.21: The mutations associated with NNRTI resistance in clinical studies. The locations within the protein sequence are given within the horizontal bar with the wild type amino acid above and the mutant form(s) beneath. This figure has been removed due to copyright restrictions but is available in Johnson *et al.* [200].

shaped and shows one wing of the drug surrounded by P95, K100, Y181, Y188, W229 and K234. The $\tau 1$ and $\tau 2$ angles appear to orient the wings relative to one another in such a way as to give the drug favourable self interactions. The second conformation seen in the crystal structures shows the pyramidine ring undergoing enhanced interactions with K100 and N103 and wing 2 with Y318. Computational studies have shown that the energetic barrier between these conformations is low (around 1.2 kcal mol^{-1}) [334].

4.7.5 Drug Resistance in Reverse Transcriptase

4.7.5.1 Resistance to NRTIs

Figure 4.20 shows the most important NRTI resistance mutations found in clinical sequences. There are two routes via which HIV-1 RT can gain resistance to NRTIs. The first is for the enzyme to evolve greater specificity for the natural substrates [337–339], the second is for it to increase the efficiency of an excision reaction [340, 341]. The mutations can be divided into two distinct categories, those close to the dNTP binding site which result in increased specificity and those which are distal and generally result in an increase in the efficacy of the template removal reaction. The distal mutations function by altering the position of the template/primer complex at the polymerase active site. The consequence of this is a change in the primer position which favours the excision reaction [318].

4.7.5.2 Resistance to NNRTIs

Figure 4.21 shows the locations of the most common clinically relevant NNRTI resistance implicated mutations in the RT sequence. All of these mutations are seen within the NNRTIBP and are generally believed to reduce binding affinity by sterically interfering with the specific interactions of the NNRTIs with residues of the binding pocket [197]. In general, unlike the case of PIs, single NNRTI mutations often produce significant reductions in binding affinity.

However, this is not true of the K103N mutation which is not seen to alter the drug enzyme contacts in crystal structures in any significant way [342]. Both crystal structures and modelling studies [342–344] have suggested that the observed resistance is a consequence of N103 to Y188 hydrogen bonding which increases the energy barrier that must be overcome to create the binding pocket. What is clear from Figure 4.21 is that mutations which cause resistance to one NNRTI will in general cause resistance to all of the clinically administered drugs. This is a problem exacerbated by the fact that single mutations are generally enough to produce resistance against NVP and DLV [197, 345]. TMC125 was specifically designed to exploit conformational flexibility to be active against resistant strains of RT [334]. *In vitro* trials have shown that significant resistance to TMC125 is associated with double or more usually triple mutants [197, 334].

The explanation of NNRTI resistance as solely caused by mutations in the binding pocket is an incomplete one. Statistical analysis of sequence data has shown that high levels of resistance in addition to the primary mutations in the NNRTIBP are associated with mutations at a range of other codons which do not directly interact with the drugs. Twenty five positions have been implicated (6, 20, 35, 39, 43, 53, 68, 90, 98, 101, 122, 179, 200, 203, 208, 218, 221, 223, 228, 284, 318, 320, 348, 359 and 371) as the locations of these so-called accessory mutations [346]. It has become apparent that mutations in the connection subdomain may also play a part in NNRTI resistance. N384I has been linked with NVP resistance [347] and the D549N, Q475A, and Y501A mutants (known to reduce RNase H cleavage) have been observed to produce resistance to NVP and DLV, but not to EFV or ETR [348]. Furthermore, combining the D549N mutant with known resistance causing mutations in the NNRTI binding pocket results in increased loss of NNRTI binding affinity.

Intriguingly, some NRTI resistance associated mutations (at positions 41, 128 and 210) have been found to be associated with an increased susceptibility of RT to inhibition by NNRTIs, this phenomena is known as hypersusceptibility [349]. At least 23 different codons in RT have been shown to be associated with an increase in NNRTI susceptibility

[350, 351]. The method by which this occurs is unknown but clinical data indicates that they have a significant effect on drug efficacy [351].

4.7.6 Computational Modelling of Reverse Transcriptase

Despite its importance as a drug target, RT has been less extensively computationally modelled than PR. This is due in large part to the greater size of RT (it contains approximately five times the number of residues) and number of the simulations which have been conducted use implicit solvents [312, 342, 344], restrained atoms [352] or only investigate subsections of the enzyme [333] in order to reduce the computational workload. Although it has been claimed that up to one third of the residues of RT are immobile [352] it is not clear that this practice, particularly when applied to residues close to the NNRTIBP or hinge areas of the enzyme, does not affect the dynamics of the system. As available computational power increases more studies are being undertaken and the design of etravirine involved considerable use of computer simulation [334]. Further discussion of MD simulations of RT is provided in Chapter 7.

4.8 Conclusions

In the last 30 years significant progress has been made in the treatment of HIV. Nonetheless drug resistant viral strains provide a continuing challenge with mutations often interacting with one another in non linear ways in order to produce resistance. The assessment of such mutants is key to successful treatment of patients. Currently, this task is performed using clinical decision support tools relying on statistical analysis of existing data. One possible approach that could enhance such systems would be to apply the molecular modelling techniques discussed in Chapter 2 and Chapter 3 to predict the resistance level of patient derived viral sequences. In Chapter 5 and Chapter 6 we develop and apply a MD protocol designed to perform exactly this function for PI resistance. In Chapter 7 we focus on extending the system to NNRTIs.

Chapter 5

Analysing the Effects of Multi-drug Resistance Causing Mutations on the Binding of Lopinavir to HIV-1 Protease

Molecular association, such as that between drugs and their protein targets, is governed by changes in free energy. This has led binding free energy differences to become one of the most studied physical quantities in biochemistry. It is through the lowering of the free energy difference between the free drug and protein and the bound complex of both that mutations in pharmaceutical target enzymes induce drug resistance.

As described in Chapter 3, a wide variety of experimental and theoretical methods are available to determine free energy differences. It is neither feasible nor economic to perform high throughput experimental studies and those studies that are possible can offer little insight into the atomic origin of changes in the binding affinity. To differing degrees molecular dynamics (MD) simulations can be used to address both of these issues. The atomic detail inherent in the MD approach is ideally suited to gaining structural and dynamic information alongside thermodynamic quantities. The methods available for calculating free energy differences from MD trajectories vary from the accurate but extremely slow to highly approximate methods designed to produce rapid results. In this chapter the MMPBSA and normal mode methodology, which seeks to reach a compromise between these two extremes, is used to assess the binding affinity of a series of HIV-1 protease sequences to the inhibitory drug lopinavir. The aim of the study is investigate the level of phase space sampling required to reliably reproduce existing experimental free energy values and to define an efficient simulation protocol

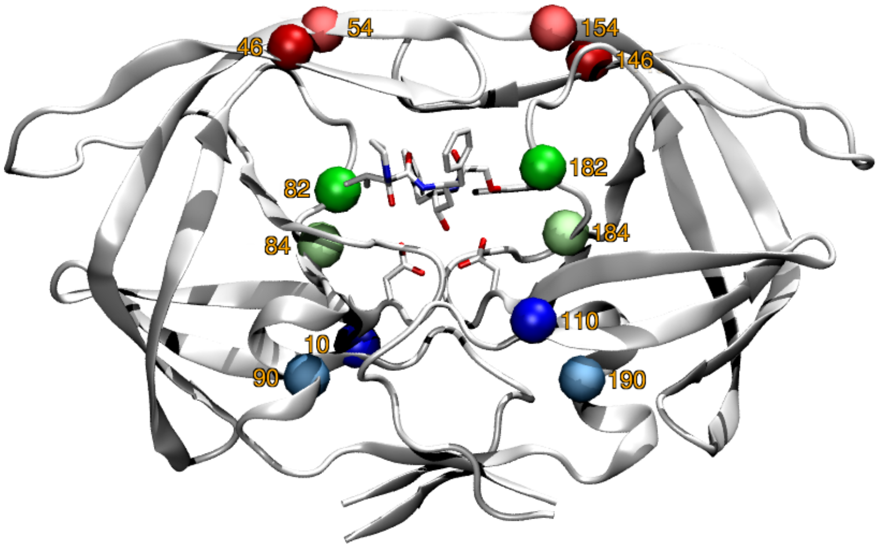


Figure 5.1: Wildtype protease structure (based on entry 1MUI from the PDB) with the locations of the residues whose effects are investigated highlighted in the following colours: L10 dark blue, L90 light blue, M46 red, I54 pink, V82 dark green and I84 light green.

that achieves this. Structural factors which may lead to resistance are also examined. The work presented in this chapter was performed in collaboration with colleagues within the CCS and published as Sadiq *et al.* [353].

5.1 Multiple Drug Resistance

As described in detail in Chapter 4, the HIV-1 protease allows the virus to form new infectious virions by cleaving the Gag-Pol polypeptide chain into functional enzymes. This vital role in the viral life cycle has made it a major target for drug design, with ten protease inhibitors currently approved by the FDA (`http://www.fda.gov`). These drugs all act competitively with the natural gag-pol substrate and, as a consequence of their common peptomimetic design, mutations which confer resistance to one often also have lowered susceptibility to several others. This phenomena is known as multi-drug resistance.

Recently, strains of HIV-1 which are resistant to all available FDA approved protease inhibitors have emerged. The main mutations implicated in conferring this resistance occur at the following residues 10, 46, 54, 82, 84 and 90 [354]. The most common mutations seen in patients are L10I, M46I, I54V, V82A, I84V and L90M and it is these that we will focus upon in this study. These mutations are not clustered in particular parts of the protease structure but spread throughout the enzyme (Figure 5.1). The set of mutations can be split into pairs according to where in the tertiary structure of the protease they occur; residues 10 and 90 appear in the dimer interface, 46 and 54 in the flaps and 82 and 84 in the wall turn which flanks the active site. Four of these residues lie in locations which are close to the active site (46, 54, 82 and 84) and are hence likely to directly affect ligand binding: the remaining pair (10 and 90) can only affect binding indirectly.

Table 5.1: Two letter codes and sequence composition for the protease sequences investigated.

Code	**Description**	**Mutations**
WT	Wildtype	HXB2
HM	MDR hexa-mutant	L10I, M46I, I54V, V82A, I84V, L90M
QM	MDR quatro-mutant	M46I, I54V, V82A, I84V
AS	Active site mutant	V82A, I84V
FL	Flap mutant	M46I, I54V
DM	Dimer interface mutant	L10I, L90M

Experiments carried out by Ohtaka *et al.* [92], on all of the FDA approved inhibitors, investigated these pairs of similarly located mutations and their combined affect upon binding affinities. In this study, the pairs and combinations thereof have been labeled with two letter codes which are shown in Table 5.1 and this nomenclature will be used for the remainder of this thesis. The group of six sequences will be referred to as the MDR Test Set.

The results for lopinavir are summarised in Table 5.2. The greater the affinity between protein and ligand the more negative the binding free energy difference. A useful term to be introduced here is the relative binding free energy, $\Delta\Delta G$, which is defined as

$$\Delta\Delta G = \Delta G_{system} - \Delta G_{ref}, \tag{5.1}$$

where ΔG_{system} and ΔG_{ref} are the binding affinities of the system of interest and a reference system respectively. In this case, the reference system is the values for wildtype; elsewhere in this chapter differences between experimental and theoretical values are expressed in this way.

Experimentally the DM and FL pairs exhibit only minor reductions in binding affinity, 0.2 kcal mol^{-1} in magnitude, whereas the AS mutant shows a considerably greater change of 1.2 kcal mol^{-1}. When combined in the QM and HM mutants the reduction in affinity is superadditive, resulting in binding affinity changes of 3.3 and 3.8 kcal mol^{-1} respectively, indicating cooperative interactions between the mutational pairs. The aim of this study is to reproduce the relative ranking of these resistant mutant genotypes from molecular dynamics simulations, analysed using MMPBSA and normal mode analysis to calculate binding free energies.

Table 5.2: Experimental free energy values (and differences compared to wildtype) for all protease sequences studied here, taken from Ohtaka *et al.* [92] with errors shown in brackets. All values are in kcal mol^{-1}.

Sequence	$\mathbf{\Delta G_{expt}}$	$\mathbf{\Delta\Delta G}$
WT	-15.1(0.09)	-
HM	-11.3(0.08)	3.8(0.17)
QM	-12.8(0.04)	3.3(0.13)
AS	-13.9(0.10)	1.2(0.19)
FL	-14.9(0.09)	0.2(0.18)
DM	-14.9(0.05)	0.2(0.14)

5.2 Methods

The first step towards the aim of reproducing the experimental binding affinity ranking obtained by Ohtaka *et al.* [92], using fully atomistic molecular dynamics simulations, is to establish a methodology which achieves sufficient phase space sampling to produce converged the binding affinities. In order to investigate how the relevant regions of phase space might most efficiently be sampled two simulations strategies were employed: (i) single simulations generating 50 ns of production trajectory and (ii) ensembles of 50 simulations of shorter duration. Individual simulations in both cases were conducted using the protocol established in a previous study of HIV-1 protease binding to the inhibitor saquinavir [288], with the sole source of initial variation between replicas within an ensemble being the randomly seeded velocity distribution assigned to the atoms of the system. The protocol involves the *in silico* incorporation of the mutations into a wildtype crystal structure before using them as the basis for molecular dynamics simulations. The simulation runs are divided into an equilibration phase during which the model is restrained and heated to physiological temperatures followed by an unrestrained production run. The production run is then analysed using the MMPBSA and normal mode methodology to provide binding free energies. Much of this process has now been automated in a tool called Binding Affinity Calculator (BAC) [355] which is described in detail in Appendix A.

5.2.1 Model Preparation

The 1MUI crystal structure [356] was used as the basis of all the models of protease bound to lopinavir created in this study. In order to distinguish the residues of the two identical monomers that form the HIV-1 protease homodimer, the two chains are numbered from 1 to 99 and 101 to 199 respectively. The monomer with the lower indices, labelled chain A in the PDB, contains the P1 and P2′ subsites whilst the monomer containing the higher indexed residues contains P2′ and P1, and is labelled chain B

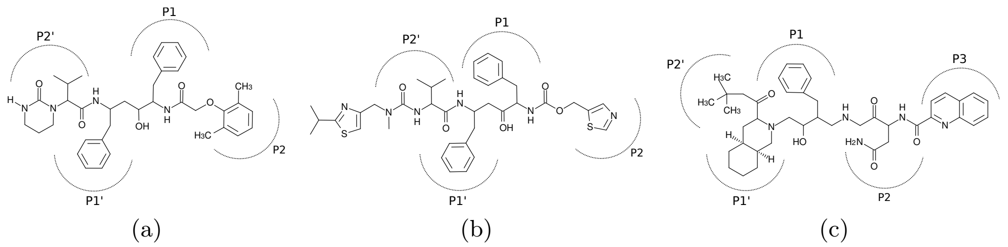


Figure 5.2: Chemical structures of the HIV-1 protease inhibitors (a) lopinavir (LPV), (b) ritonavir (RTV) and (c) saquinavir (SAQ). The structures of LPV and RTV bind to the protease with similar moieties occupying each of the protease subsites (dotted lines surrounding sections of each drug indicate the sections which interact with each subsite). This is unsurprising as LPV was originally designed as a refinement of RTV [260].

in the PDB (protease subsites were discussed in Chapter 4). The mutant protease models were derived from the 1MUI structure using the mutational protocol of the VMD [315] visualisation package, which was also used to insert the hydrogen atoms not contained in the crystal structure. Each dimeric mutation corresponds to two amino acid substitutions, one on each monomer. In addition to the substitutions required to recreate the mutant sequences under study, the mutation S37N was incorporated into all of the models in order that the model sequence matches the canonical HXB2 subtype B sequence (Genbank accession number K03455) used by Ohtaka *et al.* [92]. Inserting mutations into the structure is not expected to disrupt the protease structure as comparisons of crystal structures indicate that the tertiary structure of the enzyme is stable despite considerable variation in the amino acid sequence [224]. The wildtype system bound to saquinavir was created from the 1FB7 crystal structure [278] using the same method but requires the additional mutations V3I, V48G and M90L to recreate the HXB2 sequence.

Uniquely among crystal structures of the HIV-1 protease bound to peptomimetic inhibitors, 1MUI does not contain a water molecule bound between the drug and flap residues 50 and 150 (this molecule is conventionally labelled WAT301). In order to investigate whether this was an artifact of the crystal structure or an important distinguishing feature of lopinavir, two sets of simulations were conducted: the first used the unprocessed 1MUI structure lacking an active site water molecule; in the second a water molecule was inserted into the WAT301 position. The precise location of water insertion was determined by superimposing the 1MUI structure onto that of 1HXW, which contains the inhibitor ritonavir (RTV), and copying the location of WAT301. The 1HXW structure was chosen as ritonavir has a similar structure and chemical composition to lopinavir and exhibits a similar binding mode, with analogous moieties in each of the protease subsites (see Figure 5.2). The two structures have very similar conformations, with all atom and backbone RMSDs of 1.47 Å and 0.69 Å respectively (Figure 5.3 shows

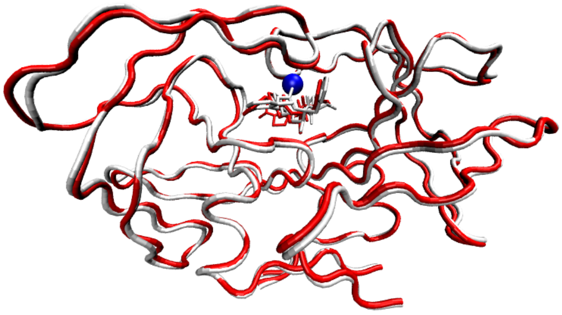


Figure 5.3: Superimposition of the backbones of the 1MUI (displayed in white) and 1HXW (shown in red) HIV-1 protease crystal structures. Minimal deviation is observed between the structures (the backbones of which have an RMSD of 0.69 Å). The active site water molecule (WAT301) of the ritonavir bound 1HXW is highlighted in blue.

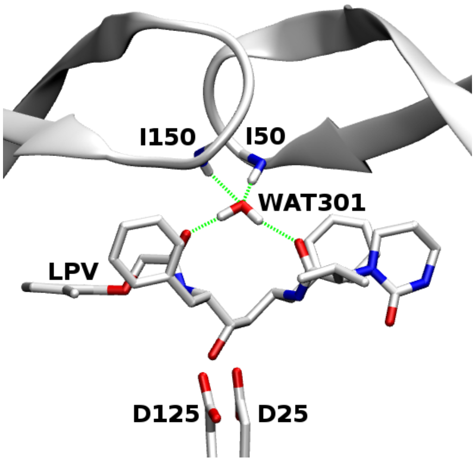


Figure 5.4: Conformation of the active site of the 1MUI crystal structure after minimisation with a water molecule inserted between the bound inhibitor, lopinavir (LPV), and the protease flaps. A water molecule, tetrahedrally hydrogen bonded to the inhibitor and residues 50 and 150, is present in this position (referred to as WAT301) in all other crystal structures of peptomimetic inhibitors bound to the HIV-1 protease. The green lines indicate the positions of putative hydrogen bonds between the water molecule and lopinavir, and the backbone nitrogens of I50 and I150 (these are analogous to those found in other inhibitor bound HIV-1 protease crystal structures).

a superimposition of the two structures where the only significant deviation is in the ear region of the first monomer). Figure 5.4 shows the final positioning of WAT301 in the 1MUI structure indicating that the putative hydrogen bonds between the water molecule and the inhibitor, and the backbone nitrogens of I50 and I150 observed in crystal structures of other drugs, are reproduced.

The drug coordinates were extracted from the crystal structure and missing hydrogens incorporated using the PRODRG tool[1] [357]. The resultant structure was then optimised employing Gaussian 98 [358] using the Hartree-Fock method and 6-31G** basis functions. The partial atomic charges were then assigned using the Restrained Electrostatic Potential (RESP) procedure, which is part of the AMBER 9 package [70]. The forcefield parameters were described using the General Amber Force Field (GAFF) [53].

The processed inhibitor and mutated protease structures were recombined using the Leap module from the AMBER 9 suite of programs. The system was then solvated by placing it in a cubic box of TIP3P [359] water molecules with at least 14 Å surrounding the complex at all points. Cl^- counter ions were added to neutralise the system. The protein potential parameters were taken from the standard AMBER forcefield for bioorganic systems (ff03) [360].

5.2.2 Minimisation and Equilibration Protocol

The molecular dynamics package NAMD2 [50] was used throughout the minimisation, equilibration and production phases of the simulations. The minimisation phase was conduced using the conjugate gradient algorithm in NAMD2 for 2000 iterations with all heavy atoms (of both protease and lopinavir) restrained using a force constant of 4 kcal mol^{-1}Å^{-2}.

The equilibration protocol used was the same as that employed in Stoica *et al.* [288] (which was itself adapted from that used by Perryman *et al.* [215]). Non-bonded interactions were cut off at 12 Å and long range Coulomb interactions were handled using the Particle Mesh Ewald (PME) method. In order to obtain an integration timestep of 2 fs the SHAKE algorithm was applied to all atoms covalently bonded to hydrogen atoms in both the equilibration and production simulations.

Each system was heated from 50 to 300 K over 50 ps and then maintained at a temperature of 300 K using a Langevin thermostat, with a 5 ps^{-1} coupling constant, for the rest of the equilibration and production phases. Once the system had been heated to the correct temperature in all subsequent simulation steps a Berensen barostat [59], with a target pressure of 1 bar and a pressure coupling constant of 0.1 ps, was applied to the system. This resulted in the system sampling an isothermal isobaric (NPT) ensemble. The restraining forces were applied for a further 200 ps in order to avoid premature flap collapse, as has been reported elsewhere [361].

The next stage of the equilibration process is a mutational relaxation protocol in which each mutated residue and residues within 5 Å are released in turn from the restraints for

[1] PRODRG: `davapc1.bioch.dundee.ac.uk/prodrg`

50 ps. This should allow the residues to reorientate into more favourable conformations if necessary. After the 50 ps relaxation period the restraints are reapplied to each region.

The final equilibration stage is the gradual reduction of the restraining force on the ligand from 4 to 0 kcal mol^{-1}Å^{-2} during a 200 ps period. The restraints on the protease were then reduced from 4 to 1 kcal mol^{-1}Å^{-2} over 150 ps. In both cases the force was reduced in equal steps of 1 kcal mol^{-1}Å^{-2}. Following this all restraints were removed and the system allowed to evolve freely. The entire equilibration stage was designed to take 2 ns for all systems meaning that this final stage varied in length according to the number of mutations which required relaxation in the previous stages.

5.2.3 Production Simulations

In both sampling strategies, the equilibrated systems were maintained in the same isothermal isobaric ensemble defined for the final equilibration stage. Ensemble simulations were initially extended for 1 ns whilst those for the single trajectory strategy had a total duration of 50 ns. In all simulations snapshots of the system coordinates were output every 10 ps for analysis. Henceforth simulation durations will always refer to the length of the production phase alone, as the equilibration phase is equivalent for all runs. The simulations described in this chapter were performed on the Ranger and Kraken machines on the US Teragrid achieving a simulation rate of approximately 4 h/ns on 64 Opteron cores/replica.

5.2.4 Data Analysis

The use of the MMPBSA and normal mode methodologies to calculate binding free energies was described in detail in Chapter 3. Equation 5.2 shows how the enthalpically dominated MMPBSA estimation of the Gibbs free energy (ΔG_{MMPBSA}) is combined with a normal mode assessment of the configurational entropy ($-T\Delta S_{NM}$) to produce an overall value for the free energy of binding (ΔG_{theor}).

$$\Delta G_{theor} = \Delta G_{MMPBSA} - T\Delta S_{NM}. \tag{5.2}$$

In the following sections we describe the specific implementation and parameters used during this study.

5.2.4.1 MMPBSA Calculations

The MMPBSA estimate of the free energy is given by

$$\Delta G_{MMPBSA} = \Delta G^{MM}_{vdW} + \Delta G^{MM}_{ele} + \Delta G^{sol}_{pol} + \Delta G^{sol}_{nonpol} \tag{5.3}$$

where ΔG^{MM}_{vdW} and ΔG^{MM}_{ele} represent the decomposition of the molecular mechanics free energy difference into van der Waals and electrostatic components, and ΔG^{sol}_{pol} and ΔG^{sol}_{nonpol} the polar and non polar contributions to the solvation free energy difference, respectively. Modules of the AMBER 9 package [70] were used in the evaluation of all components of the MMPBSA calculation. The SANDER module was employed to calculate both molecular mechanics terms (ΔG^{MM}_{vdW} and ΔG^{MM}_{ele}), with no cut off being applied to the non-bonded energies. The electrostatic free energy of solvation, ΔG^{sol}_{pol}, was calculated by the PBSA module. Internal and external dielectric constants were of 1 and 80 were used respectively. A thousand iterations of the linear Poisson-Boltzmann equation were performed on a cubic lattice grid with a spacing of 0.5 Å. The non-polar solvation energy, ΔG^{sol}_{nonpol}, was calculated from the solvent accessible solvent area (SASA) using the MSMS program [132] with a 1.4 Å radius probe. The surface tension (γ) and offset (b) were set to the standard values of 0.0052 kcal mol^{-1} and 0.92 kcal mol^{-1}, respectively.

Every output snapshot was post-processed using MMPBSA, meaning that a hundred sets of coordinates were analysed for each nanosecond of simulation. Thus both the 1 × 50 ns single trajectory and 50 × 1 ns ensemble strategies generated a total of 5000 measurements of ΔG_{MMPBSA}.

5.2.4.2 Normal Mode Calculations

The changes in configurational entropy, ΔS, were evaluated by normal mode analysis performed using the AMBER NMODE module. In order to ensure that the protease structure used in the calculation is within an energetic minimum, each snapshot was subjected to a minimisation with a distance dependent dielectric constant $\epsilon = 4r$ and a RMS gradient convergence tolerance of 10^{-4} kcal mol^{-1}Å^{-1}. Every twentieth output snapshot was post processed using normal mode analysis, meaning that 5 sets of coordinates were analysed for each nanosecond of simulation. Thus both the 1 × 50 ns single trajectory and 50 × 1 ns ensemble strategies generated a total of 250 measurements of $-T\Delta S_{NM}$.

5.2.4.3 Convergence Analysis

The convergence of the free energy calculations is a vital property which requires consideration when comparing the two simulation strategies under investigation. Assessment of convergence was performed using two primary methods, the assessment of the extent to which the data sets could be described as having a Gaussian distribution and the root mean squared difference between forward ($< \Delta X >_{\tau}^{for}$) and reverse ($< \Delta X >_{\tau}^{rev}$) cumulative means. This metric as a function of the snapshot number, ε, was designated $\sigma(\varepsilon)$ and was calculated such that:

$$\Delta\Delta X_{\tau} =< \Delta X >_{\tau}^{for} - < \Delta X >_{\tau}^{rev} = \frac{1}{\tau}\left[\sum_{i=1}^{\tau}\Delta X_i - \sum_{i=N+\tau+1}^{N}\Delta X_i\right] \tag{5.4}$$

where X denotes either ΔG_{MMPBSA}or $-T\Delta S_{NM}$, i is the i, τ the number of snapshots over which the mean was evaluated, N the total number of snapshots available in the trajectory for analysis and $\Delta\Delta X_{\tau}$ the instantaneous difference, giving:

$$\sigma(\varepsilon) = \sqrt{\frac{1}{\omega}\sum_{j=\varepsilon}^{\varepsilon-1+\omega}\Delta\Delta X_j^2} \tag{5.5}$$

where ω is the number of snapshots over which the RMS difference was calculated. The value of ω was set to represent the number of snapshots analysed per nanosecond; 100 for the MMPBSA calculation and 5 for the normal mode analysis. The value of $\sigma(N/2)$, representing a comparison of the first and second halves of the trajectory, utilises the maximum possible sample size for non-overlapping datasets. Beyond this value $\sigma(\varepsilon)$ will decay to zero at the point of equivalence of the forward and reverse trajectories by definition. With this consideration in mind, $\sigma(\varepsilon)$ was only calculated up to the mid point of the trajectory.

The Gaussian nature of the measurements was assessed by comparing the data produced from the simulation analysis with an expected normal distribution, centred on the same mean and of the same standard deviation.

5.2.5 Model Finalisation

The choice of initial model required two further decisions to be made. Firstly, the necessity and desirability of inserting a water molecule in position WAT301, between the flaps and the inhibitor lopinavir, in the active site of the 1MUI crystal structure

Table 5.3: The number of replicas in 10 member ensembles for which water molecules have entered the initially unoccupied WAT301 position by the end of the equilibration and production phases of simulation. Results are shown for each of the four possible protonation states of the HIV-1 catalytic dyad. The production phase in each case lasted for 4 ns and in all systems, once a water molecule entered the WAT301 position, it persisted there until the end of the simulation.

	No. Replicas Containing WAT301	
Protonation State	**Equilibration**	**Simulation**
D25	4	4
D125	1	2
D25/D125	1	2
D-	3	6

was assessed and secondly, the protonation state of the catalytic aspartic acid dyad was determined.

In the following discussions the protonation states of the catalytic aspartic acid dyad are denoted as the dianionic (D-), diprotonated (D25/D125), position 25 protonated (D25) and position 125 protonated (D125) states.

5.2.5.1 Active Site Water Insertion

In order to ascertain whether the initial system studied should contain a water molecule placed at the WAT301 position, ensemble simulations containing ten replicas were performed for all four possible protonation states of the HIV-1 protease catalytic aspartic acid dyad both with and without this additional water molecule. Each replica was run for 4 ns of production simulation in addition to the 2 ns of equilibration detailed above. Table 5.3 shows the frequency of water entering position WAT301 in the ensembles containing no water when intialised. For all protonation states in at least two of the ten replicas water was observed moving to occupy the WAT301 locality and once this had occurred, in all instances, it persisted until the end of the production phase.[2] For those systems where a water molecule was inserted into the WAT301 locus prior to the execution of molecular dynamics it was observed to remain for the full simulation duration. Due to the unpredictable length of time taken for water to enter the active site it was decided that for the rest of the study the systems used would be based upon the structures with an inserted water molecule in the WAT301 position.

[2] Once the protonation state was determined a D25 protonated system was extended to 27 ns of production simulation. A water molecule entered the WAT301 position around 7 ns into the production phase and persisted for the remainder of the run.

5.2.5.2 Protonation State Assignment

As discussed in Chapter 4, the question of which protonation state of the catalytic aspartic acid dyad is favoured in physiological conditions remains an open one, but the answer is believed to be an important factor in accurately determining binding affinities [362] and to vary depending upon the particular ligand to which the protease is bound [247–250].

Table 5.4: Assessment of protonation state performed on the lopinavir bound wildtype HIV-1 protease.

Protonation State	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{-T\Delta S_{NM}}$	$\mathbf{\Delta G_{theor}}$
D25	-47.70 (0.05)	37.29 (0.79)	-10.41 (0.84)
D125	-46.13 (0.05)	38.92 (0.80)	-7.21 (0.85)
D25/D125	-47.67 (0.05)	39.85 (0.83)	-7.82 (0.88)
D-	-32.15 (0.11)	38.23 (1.23)	6.08 (1.34)

Mean energies are in kcal/mol. Standard errors are shown in parentheses.

In order to determine the protonation state to be used for the rest of the study, the MMPBSA and normal mode methodology was applied to wildtype systems in all four possible protonation states of the catalytic dyad. Ensembles of twenty replicas were run for each system (except for D- for which only ten were performed due to the consistently positive ΔG_{theor} values obtained). The computed binding affinities along with the decomposed enthalpic and entropic contributions are presented in Table 5.4. The D25 protonated system is observed to have the most attractive binding affinity, ΔG_{theor}, by over 2.5 kcal mol^{-1}. The D25 state is substantially favoured over the other monoprotonated state both by the total binding affinity but also by the enthalpically dominated ΔG_{MMPBSA}. The D25 and D25/D125 systems exhibit much closer ΔG_{MMPBSA} values but the level of separation between them in ΔG_{theor} allows the confident selection of D25 as the protonation state to be used in the rest of the study (we will assume this to be invariant for all the mutant systems under investigation). This determination is in line with a number of other experimental and theoretical studies of HIV-1 protease-inhibitor complexes [246, 247, 363, 364] which find the catalytic dyad to be monoprotonated and in particular with the findings of Wittayanarakul *et al.* [362] and Ode *et al.* [365] for lopinavir which both suggest a D25 protonation state.

It should be noted that this analysis of the protonation state takes no account of the possibility of the protonation state of the protease altering during binding. Wittayanarakul *et al.* [362] noted that this omission affects all the comparisons between the systems except that between the two monoprotonated systems.

5.3 Comparison of Single Trajectory and Ensemble Simulation Strategies

Key to the calculation of meaningful free energies is the need to obtain sufficient sampling and the assessment of the convergence of the values computed. The aim of this study is to investigate the sampling and convergence properties of the single trajectory and ensemble simulation strategies for computing binding free energies with the combined MMPBSA and normal mode methodology. In order to compare the two approaches protocols were defined for each strategy which produced identical numbers of system configurations to be analysed. The protocols selected were a single 50 ns simulation (henceforth referred to as the 1 × 50) and 50 replica simulations each of 1 ns duration (labelled 50 × 1 ns). Both strategies produced 50 ns of production trajectory for each system run, containing 5000 snapshots. All of these snapshots were analysed using MMPBSA with 250 additionally processed using normal mode analysis.

5.3.1 Sampling and Convergence

The normalised frequency distributions of both the enthalpically dominated MMPBSA values, ΔG_{MMPBSA}, and configurational entropies, $-T\Delta S_{NM}$, are shown in Figure 5.5 for all systems in the MDR Test Set using both sampling strategies. Most of the systems for both the 1 × 50 ns and 50 × 1 ns strategies exhibit approximately normal distributions of the ΔG_{MMPBSA} values. However, the AS and QM in the 1 × 50 ns data set are significant exceptions. The distributions for both of these systems show multiple peaks, for AS these are located at approximately -45 and -37 kcal mol^{-1} and for QM close to -26 and -28 kcal mol^{-1}. These divergences from normality indicate that the simulations are visiting at least two specific energy minima but sampling them inadequately. The lack of such deviations in the 50 × 1 ns data set indicates that the minima visited by these simulations are well sampled. The sampling of the configurational entropy shows greater similarity between the two strategies (see Figure 5.5b). The distributions are shallower than those seen for the MMPBSA calculations and range between 0 and 80 kcal mol^{-1} in all systems. The approximation to a normal distribution is less convincing than in the ΔG_{MMPBSA} case for both the 1 × 50 ns and 50 × 1 ns strategies. A number of points fall substantially above or below the expected normal distribution curve, with several systems showing significant variation between the mean and modal averages. A plausible explanation of this is that it is indicative of the entropy values being derived from a reduced subset of specific minima which have varying levels of accessibility for each system.

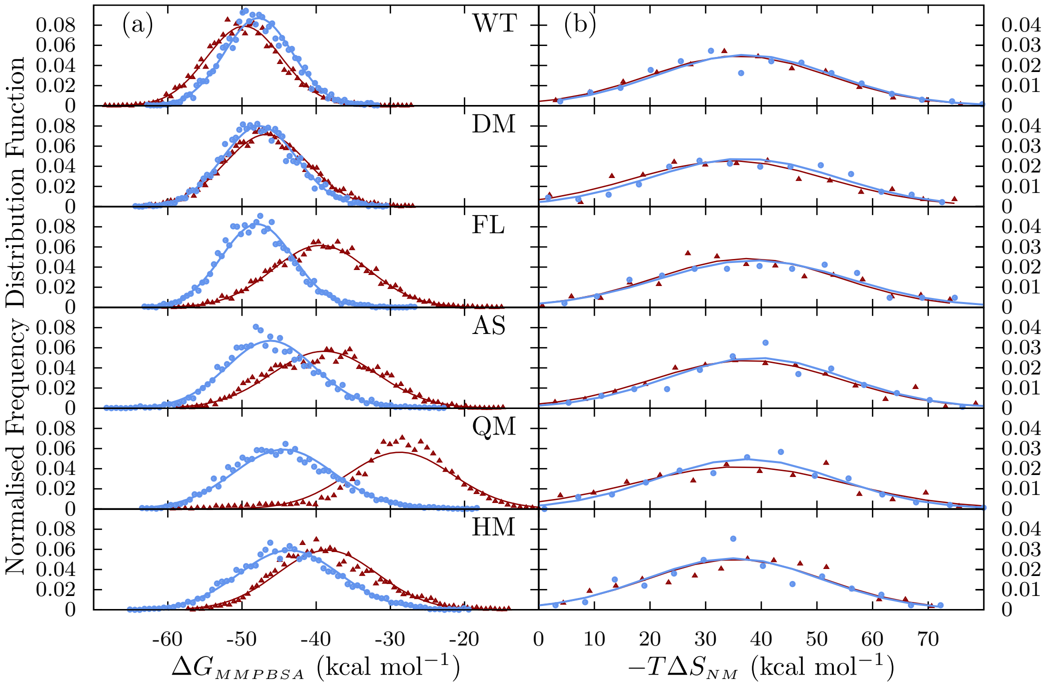


Figure 5.5: Normalised frequency distribution analysis of (a) the mainly enthalpic ,ΔG_{MMPBSA}, and (b) the entropic, $-T\Delta S_{NM}$, components of the binding free energy for the 1 × 50 ns (red triangles) and 50 × 1 ns trajectories (blue circles) for each system in the MDR Test Set. The expected normal distribution given the same mean and standard deviation for each data set is shown by the red and blue lines, respectively.

The sampling achieved by the two strategies is also differentiated by the convergence analysis performed here. The RMS difference in cumulative means for the MMPBSA component, σ_{MMPBSA}, in particular shows the 50 × 1 ns ensembles to be achieving much higher levels of internal consistency that 1 × 50 ns simulations. Figure 5.6a shows the σ_{MMPBSA} value to be greater than 2 kcal mol^{-1} at 25 ns for all the 50 × 1 ns systems (with the DM and AS sequences particularly poorly converged with values of 5 and 9 kcal mol^{-1}, respectively). In contrast all of the systems in the 50 × 1 ns data set converge to within 1.5 kcal mol^{-1}, with four (WT, FL, AS and QM) below 0.5 kcal mol^{-1}. The consistently low values of $\sigma_{MMPBSA}(\varepsilon)$ seen for the 50 × 1 ns systems indicates that statistically relevant sampling of different minima is being achieved independent of which set of replicas are used. The differences seen in the values of $\sigma_{NM}(\varepsilon)$ derived from the measurements of the entropic component of the binding free energy show much less marked differences between the two strategies (see Figure 5.6b). With the exception of the DM and HM sequences in the 1 × 50 ns data set (which fall to approximately 6 and 4 kcal mol^{-1} respectively) all systems are seen to converge to within 2 kcal mol^{-1} at 25 ns.

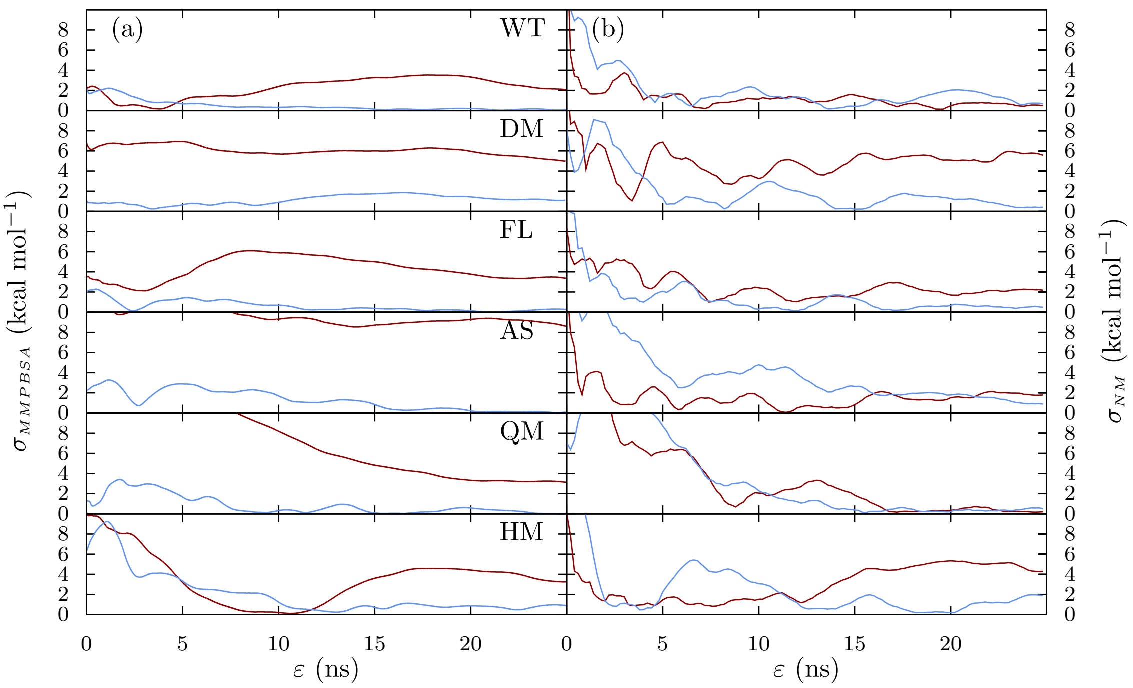


Figure 5.6: RMS difference in cumulative means, $\sigma(\varepsilon)$, of (a) the mainly enthalpic ,ΔG_{MMPBSA}, and (b) the entropic, $-T\Delta S_{NM}$, components of the binding free energy for the 1 × 50 ns trajectory (red lines) and 50 × 1 ns concatenated trajectories (blue lines).

5.3.2 Absolute and Relative Binding Free Energy Rankings

Table 5.5 shows the means computed for all six mutants within the MDR Test Set using both simulation strategies. The absolute binding affinity, ΔG_{theor}, values calculated from the 1 × 50 ns simulations show poor agreement with the experimental results, with a RMS $\Delta\Delta G_{theor-expt}$ of 11.42 kcal mol^{-1} and individual sequence deviations varying by up to 18 kcal mol^{-1}. The positive binding affinity observed for the QM system in particular stands out as particularly difficult to explain. The binding affinities for the 50 × 1 ns also show a significant RMS $\Delta\Delta G_{theor-expt}$ of 5.11 kcal mol^{-1} but the range of deviation is much smaller at 4.30 kcal mol^{-1}. In only one case (the WT sequence) does the 1 × 50 ns approach produce a result closer to the experimentally derived binding free energy than that from the 50 × 1 ns strategy. Neither strategy reproduces even the correct relative ranking order of the sequences The 1 × 50 ns results having the HM system less resistant than the FL (which was experimentally barely distinguishable from WT) and the QM system as by far the most resistant. The 50 × 1 ns data set has the DM as slightly more attractive that WT (although the separation is much less than the standard error) and reverses the ordering of the AS and QM sequences. This failure is reflected in the low correlation coefficients, κ, of the ΔG_{theor} values relative to experiment (0.56 and 0.55 for the 1 × 50 ns and 50 × 1 ns datasets, respectively). Despite these observations the 50 × 1 ns dataset exhibits separation between the susceptible WT, DM and FL systems and the highly resistant AS, QM and HM sequences.

Further to this, limited, ability to distinguish resistant mutants using ΔG_{theor} the 50 × 1 ns ensemble exhibits a very strong correlation (having a κ of 0.98) between enthalpically dominated ΔG_{MMPBSA} values and the experimental results. This indicates that using the ΔG_{MMPBSA} values alone, at least in the ensemble strategy, is a viable approach if all that is required is reproduction of the relative ranking of sequences.

Table 5.5: Computed free energy differences of binding (ΔG_{theor}) compared with experimental results (ΔG_{exp}) for all six HIV-1 protease sequences in the MDR Test Set with LPV using both 1 × 50 ns single-trajectory, 50 × 1 ns ensemble strategies. The enthalpically dominated MMPBSA and the normal mode entropic components are also shown. Correlation coefficients, κ, are provided for each theoretically computed dataset compared to the experimental data.

Sequence	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{-T\Delta S_{NM}}$	$\mathbf{\Delta G_{theor}}$	$\mathbf{\Delta G^{*}_{exp}}$	$\mathbf{\Delta\Delta G^{*}_{theor-exp}}$
			Single Trajectory (1 × 50 ns)		
WT	-49.72 (0.07)	35.85 (1.01)	-13.87 (1.08)	-15.1 (0.09)	1.23 (1.17)
HM	-38.52 (0.10)	35.20 (1.10)	-3.32 (1.10)	-11.3 (0.08)	7.98 (1.18)
QM	-28.71 (0.10)	35.86 (1.20)	-7.15 (1.30)	-12.8 (0.04)	19.95 (1.34)
AS	-38.91 (0.10)	37.17 (1.06)	-1.74 (1.16)	-13.9 (0.10)	12.16 (1.26)
FL	-39.41 (0.09)	37.35 (1.04)	-2.06 (1.13)	-14.9 (0.09)	12.84 (1.22)
DM	-46.70 (0.08)	34.28 (1.11)	-12.47 (1.19)	-14.9 (0.05)	2.43 (1.24)
κ	0.62		0.56		
			Ensemble (50 × 1 ns)		
WT	-47.79 (0.06)	37.12 (1.00)	-10.67 (1.06)	-15.1 (0.09)	4.43 (1.15)
HM	-43.63 (0.10)	34.95 (0.98)	-8.68 (1.08)	-11.3 (0.08)	2.62 (1.16)
QM	-44.40 (0.10)	37.33 (1.01)	-7.07 (1.11)	-12.8 (0.04)	5.73 (1.15)
AS	-46.15 (0.08)	39.17 (1.01)	-6.98 (1.09)	-13.9 (0.10)	6.92 (1.19)
FL	-47.95 (0.07)	38.74 (1.08)	-9.21 (1.15)	-14.9 (0.09)	5.69 (1.24)
DM	-47.62 (0.07)	36.87 (1.06)	-10.75 (1.13)	-14.9 (0.05)	4.15 (1.18)
κ	0.98		0.55		

*Experimental results are taken from Ohtaka *et al.* [92].
Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

5.3.3 Structural and Energetic Sampling

The root mean square deviation (RMSD) of the HIV-1 backbone provides a measure of the gross structural changes which each simulation undergoes. Figure 5.7 shows the backbone RMSD compared to the average structure over the trajectories for each sequence. The values for all systems under study in both the 1 × 50 ns and the 50 × 1 ns data sets exhibit stable behaviour fluctuating around 1.5 Å. This measure indicates none of the systems, simulated using either strategy, undergoes significant structural rearrangement during the production phase of the simulation. In the ensemble case this is also indicative that all of the replicas remain close to the initial crystal structure. The RMSD can say nothing, however, about motions on the scale of individual residues (such

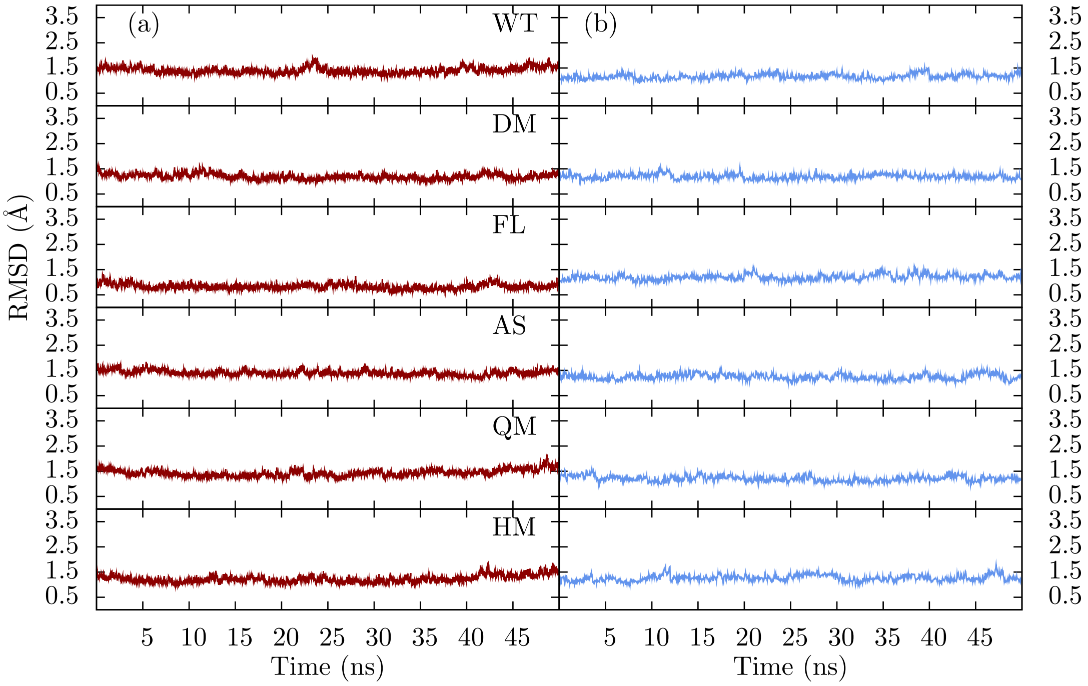


Figure 5.7: The RMS fluctuations of the backbone of all six sequences in the MDR Test Set compared to the average structure for the trajectories produced by both (a) 1 × 50 ns and (b) 50 × 1 ns strategies.

as the exploration of different rotamers) which are likely to be important in determining the binding free energies. A comparison with the binding affinities shown in Figure 5.8 indicates that changes in the binding affinity occur without the need for gross structural changes.

Clear differences in the energetic sampling between the two strategies can be seen. Figure 5.8 shows running averages of the ΔG_{MMPBSA} values for both strategies, with both exhibiting a wide range of free energies. The 1 × 50 ns track however shows long periods of sampling around a single energy level potentially indicating that the simulations become trapped in local minima. This phenomena is most easily observed in the AS and QM systems. The more extreme values seen in the ensemble strategy are only seen as brief peaks and troughs. Each individual member of the ensembles in the 50 × 1 ns strategy may similarly become trapped and sample only a small region of phase space but the randomised initial velocities makes the sampling of a larger number minima more probable. Whilst it is likely that the same arguments can be applied to measurements of the configurational entropy the large variability of individual normal mode measurements make the effect less apparent. The 1 × 50 ns $-T\Delta S_{NM}$ values in Figure 5.9 do not show the clustering apparent in the single trajectory ΔG_{MMPBSA} values.

The differences in sampling seen between the two strategies is consistent with the widely

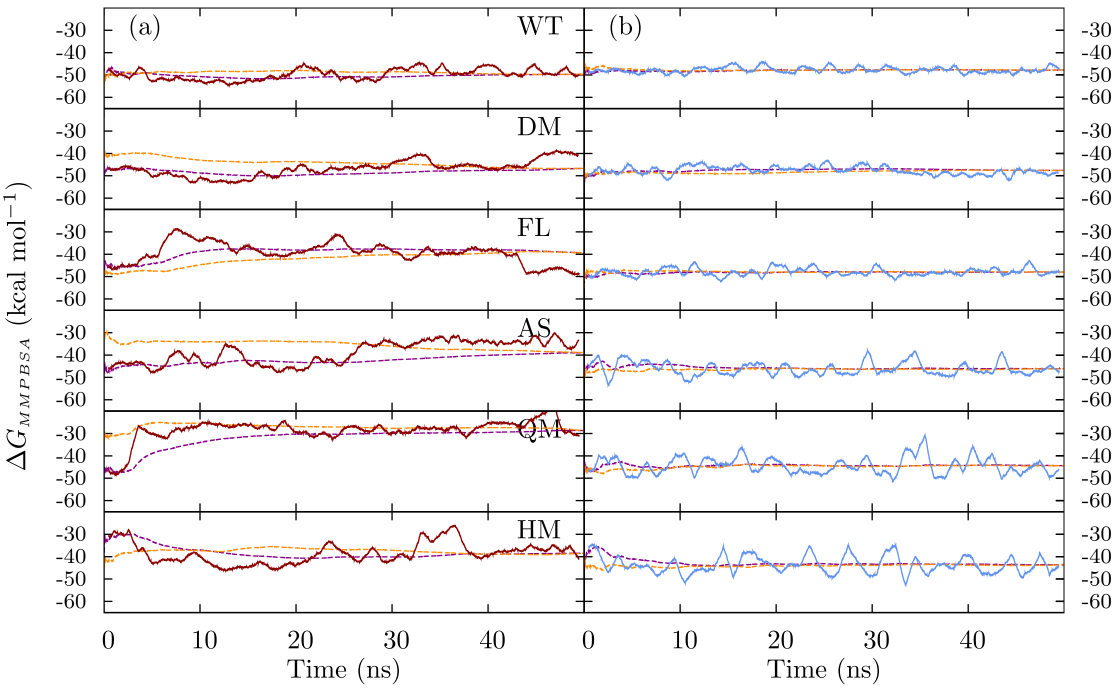


Figure 5.8: Convergence of the enthalpically dominated ΔG_{MMPBSA}component of the binding free energy difference for (a) 1 × 50 ns and (b) the 50 × 1 ns strategies. Forward and reverse cumulative means are shown in orange and magenta respectively, 1 ns running means are shown in red and blue.

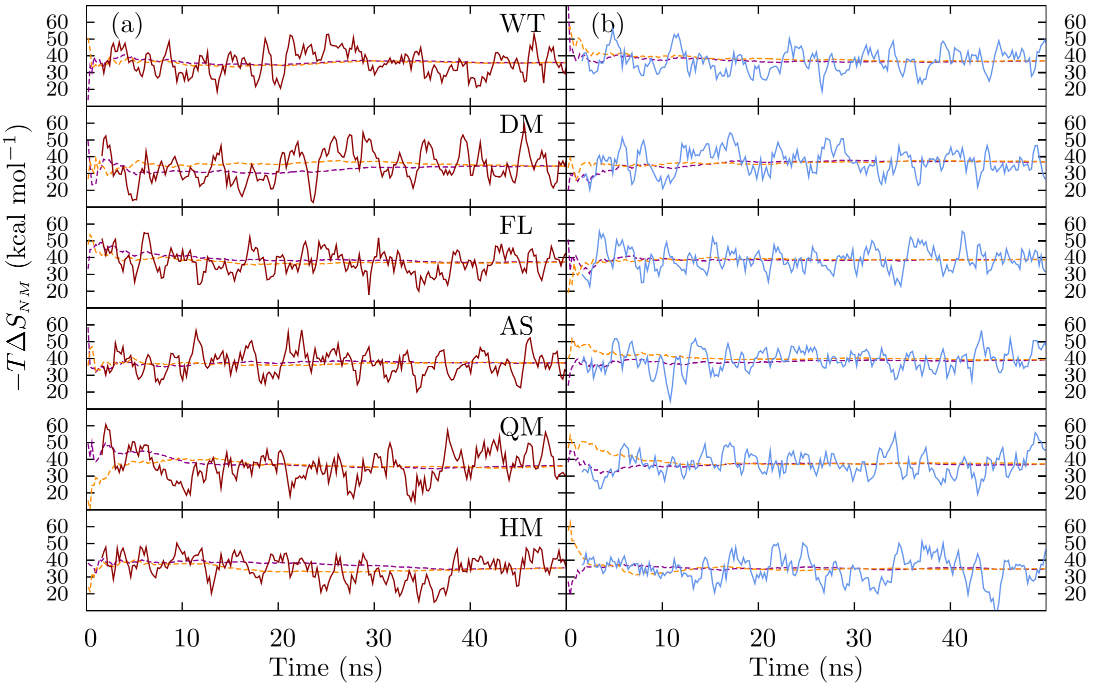


Figure 5.9: Convergence of the entropic $-T\Delta S_{NM}$component of the binding free energy difference for (a) 1 × 50 ns and (b) the 50 × 1 ns strategies. Forward and reverse cumulative means are shown in orange and magenta respectively, 1ns running means are shown in red and blue.

Table 5.6: The population density of catalytic water occupancy (ρ) in the 50 × 1 ns simulations of all sequences in the MDR Test Set.

Sequence	ρ
WT	0.011
HM	0.382
QM	0.319
AS	0.131
FL	0.019
DM	0.010

believed picture of the protein-ligand free energy landscape being rugged with many low energy minima separated from one another by high energy barriers. It is the potential to visit a large number of these minima which motivates the use of ensembles of simulations. The MMPBSA method is an end point approach and hence the only area of phase space which it is relevant to sample is that where the complex is well defined. In a previous study the inhibitor saquinavir bound to a mutant HIV-1 protease was seen to exhibit structural drift on a multi-nanosecond timescale. It is thus reasonable to conclude that for simulations on this timescale conformations that form part of the exit pathway of the drug may be sampled. MMPBSA analysis of such conformations in the intermediate region of phase space between the free and bound states should be avoided. The use of shorter simulation timescales in ensembles serves to limit the structural drift whilst also enhancing the energetic sampling.

A recent accelerated dynamics study of HIV-1 protease bound to a short section of its natural polypeptide substrate by Pietrucci *et al.* [366] indicates that a number of interconverting states exist around that seen in the ligand bound crystal structure. One of the defining features differentiating these states is the number of water molecules that enter into two separate locations in the active site; the catalytic cavity (between the ligand and the catalytic dyad) and the area around WAT301 (between the ligand and the protease flaps). The presence of water molecules in the latter position is generally associated by Pietrucci *et al.* [366] with states further along the exit pathway. To investigate whether these states were sampled in the simulations presented here the number of water molecules in each location was counted for each snapshot in all simulations. Water molecules within 3 Å of the D25/D125 dyad were counted as within the catalytic cavity, the criteria for those counted as between drug and flaps was within 7 Å of the central carbon of lopinavir and within 5 Å of residues 50 or 150. In the 50 × 1 ns data set no water molecules (other than that inserted pre-simulation in position WAT301) appear in the region between lopinavir and the flaps. Water entry into the catalytic cavity is also infrequent, although it is much more common in the resistant than the susceptible sequences (see Table 5.6). In the 1 × 50 ns data set water entry into both locations

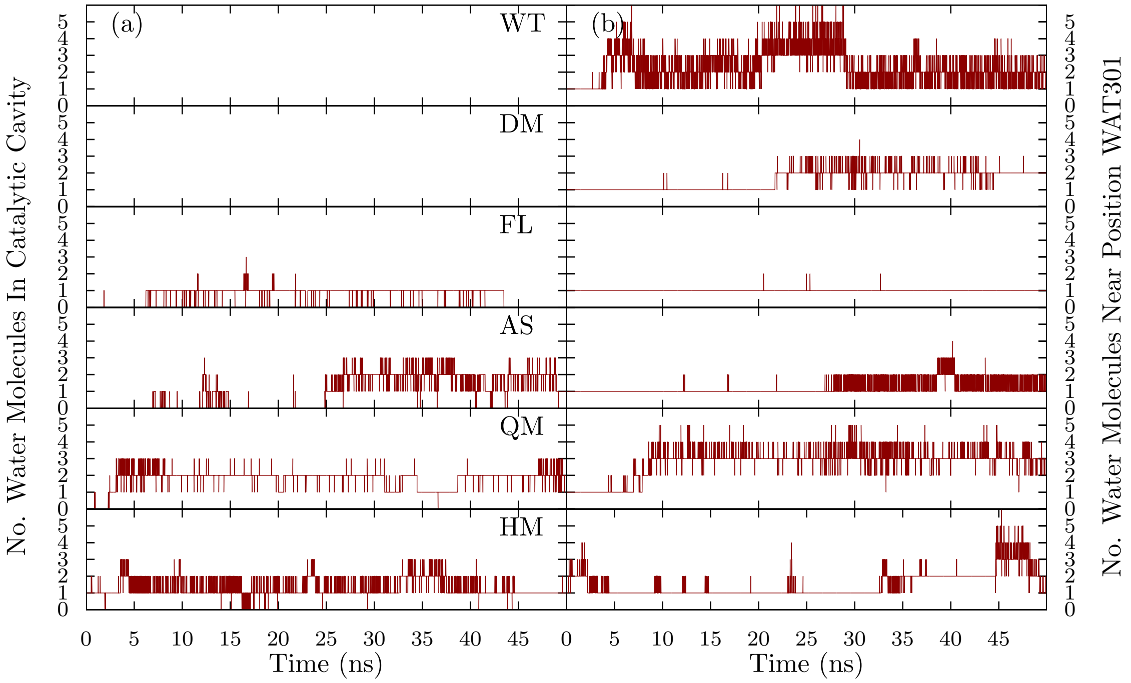


Figure 5.10: The number of water molecules entering the active site during the 1 × 50 ns separated into those which (a) enter within 3 Å of the catalytic dyad and (b) those that enter between the protease flaps and the bound lopinavir. The criteria for the second location was that the molecules were within 7 Å of the oxygen bound carbon of the lopinavir hydroxyethylene moiety and within 5 Å of residues 50 and 150 in the protease flaps.

is observed in several system. The sequences differ in the number of water molecules entering each location and in the duration of their stay once present. Figure 5.10 shows the number of water molecules occupying each location over the course of each 50 ns simulation. The frequent discontinuities in the lines represent the dynamic nature of the configurations, which often change as confined water molecules exchange with those in the free solvent. The QM system, which exhibits the largest deviation between experimental and theoretical binding affinities, shows large numbers of water molecules in both locations. This suggests that when water molecules are present both above and below the inhibitor the simulation is moving into regions of phase space in which the MMPBSA methodology may not be applicable. Similar, if less pronounced, effects can be observed in the other highly resistant mutants (AS and HM). In the FL system one water molecule is present in the catalytic cavity almost constantly between 6 ns and 44 ns into the trajectory. The presence of this water molecule coincides with a reduction in attractiveness in ΔG_{MMPBSA} (see Figure 5.8a). In this system only fleeting entry of an extra water molecule around WAT301 is detected. Other than this exception, in all other systems once water has entered between flap and inhibitor the system remains in a state with additional water present in this region. The failure to convert back to the original state in these circumstances contrasts with the situation for the catalytic cavity water molecule in FL, where the water molecule returns to the solvent and the computed ΔG_{MMPBSA} value returns to a similar average value to before water entry.

Water entry into the catalytic cavity is observed in the 50 × 1 ns ensemble simulations which show strong correlation between $\Delta G_{MMPBSA}$and ΔG_{exp} but water entry between flap and inhibitor is not. This suggests that water entry in the former region can be reasonably modelled within the framework of MMPBSA, whereas the latter is more problematic. This is in line with the evidence from Pietrucci *et al.* [366] indicating that HIV-1 protease-substrate complexes with water molecules beneath the ligand are close in free energy to the conformation seen in crystal structures without one, and that states involving water between flaps and substrate are only observed further along the exit pathway.

5.3.4 Evaluation of Single Trajectory and Ensemble Strategies

Both of the methods employed to investigate the convergence of the free energy calculations presented here indicate that the 50 × 1 ns strategy performs more comprehensive sampling of the minima relevant to the MMPBSA calculations than the 1 × 50 ns approach. Despite the fact that the 1 × 50 ns WT calculation exhibits the smallest deviation from the experimental values, none of the systems using this strategy exhibited convergent ΔG_{MMPBSA} values. This contrasts strongly with the 50 × 1 ns case, where all systems produced values converged to a state where σ_{MMPBSA} was below 1 kcal mol^{-1} at the trajectory mid point and exhibit sampling of a correct Gaussian distribution. Not only do the ensemble ΔG_{MMPBSA} values exhibit superior convergence properties to the single trajectories but they are also much more highly correlated to the experimental free energy values (with a correlation coefficient, κ, of 0.98 compared to 0.62).

Neither strategy, however, can exhibit satisfactory Gaussian sampling of the configurational entropy (despite convergence of σ_{NM} to within 2 kcal mol^{-1} in the 50 × 1 ns case). Furthermore, the absolute free energy differences produced by both strategies are only marginally correlated to the experimental values. A distinction should however be made between the two strategies as the discrepancy between the experimental and computed values observed in the 1 × 50 ns data set are highly variable (ranging from 1.23 to 19.95 kcal mol^{-1}), whereas those for the 50 × 1 ns simulations are more tightly bounded. This makes the existence of a systematic error plausible in the later case but not the former. The improved sampling observed in the ensemble data can, at least in part, be attributed to the fact that the initial randomisation of velocities provides access to a wider range of minima. It is also likely that the shorter simulation length avoids structural drift and exploration of configurations on the drug exit pathway for which the MMPBSA approach is no longer valid.

5.4 Extended Ensemble Evaluation

The superior internal convergence and sampling properties of the 50 × 1 ns strategy along with the increased level of correlation between the computed ΔG_{MMPBSA} values and experiment provided compelling reasons to investigate the effect on the calculations of extending the ensemble runs. In order to do this each of the 50 replicas were extended to 4 ns in duration. This simulation length was chosen as being short enough to minimise the probability of additional water molecules entering the active site between the inhibitor and the protease flaps. This was observed to occur in the 1 × 50 ns data set systems most often after 5 ns (see Figure 5.10b) and to lead to the sampling of conformations that may form part of an exit pathway, where the MMPBSA approach is no longer valid.

5.4.1 Sampling and Convergence

Figure 5.11a shows that the ΔG_{MMPBSA} values for all systems in the MDR Test Set using the 50 × 4 ns ensemble provide excellent agreement with the expected Gaussian distribution. The $-T\Delta S_{NM}$ values for all sequences exhibit improved correspondence with the expected normal distributions compared to the 50 × 1 ns data set with modal values close to the mean in all cases (see Figure 5.11b). Despite this, significant deviations suggesting that the measured values remain grouped around distinct peaks. Considering the level of sampling employed here it is likely that this result is an inherent property of the normal mode methodology. By the mid point of the concatenated trajectories the σ values for both components are below 0.5 kcal mol^{-1}, with the exception of the DM sequence MMPBSA result which is nonetheless below 2 kcal mol^{-1} (see Figure 5.12). These results indicate that the computed free energy differences are excellently converged and that, at least for ΔG_{MMPBSA}, thorough statistical sampling of all minima visited during the simulations has been performed. In both the convergence and sampling, notable improvements have been made over that seen in the 50 × 1 ns ensemble with greater consistency now seen across the different sequences. In particular the sampling of the entropic contribution to the binding free energy is considerably enhanced even if fully Gaussian sampling is not possible using the approach taken in this study.

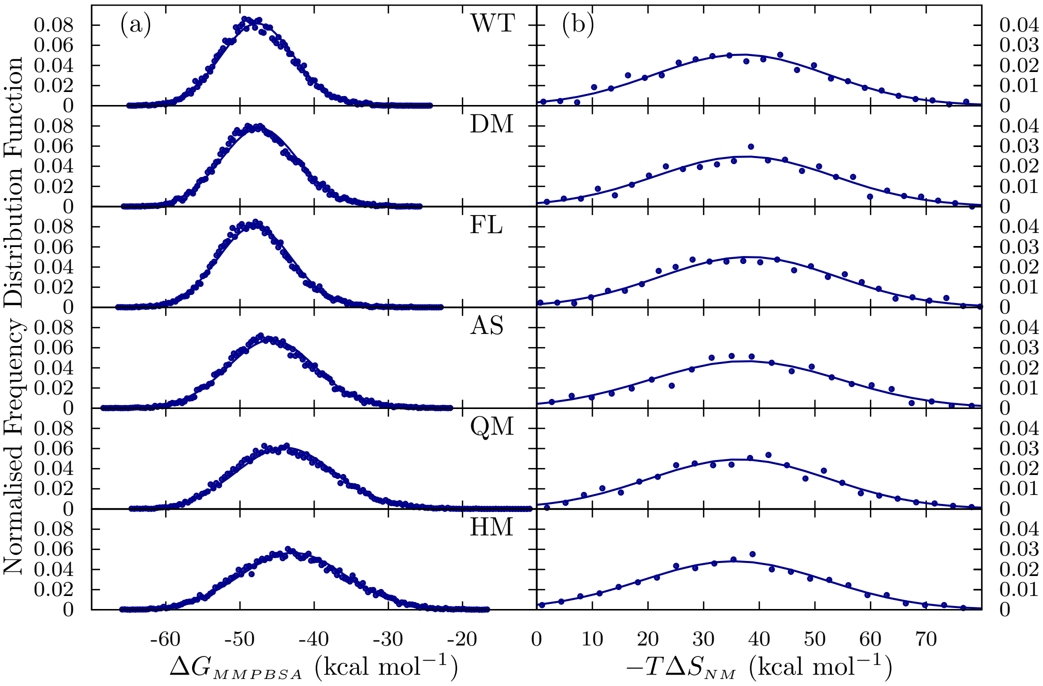


Figure 5.11: Normalised frequency distribution analysis of (a) the mainly enthalpic , ΔG_{MMPBSA}, and (b) the entropic, $-T\Delta S_{NM}$, components of the binding free energy for the 50 × 4 ns trajectories (blue circles) for each system in the MDR Test Set. The expected normal distribution given the same mean and standard deviation for each data set is shown by the blue lines.

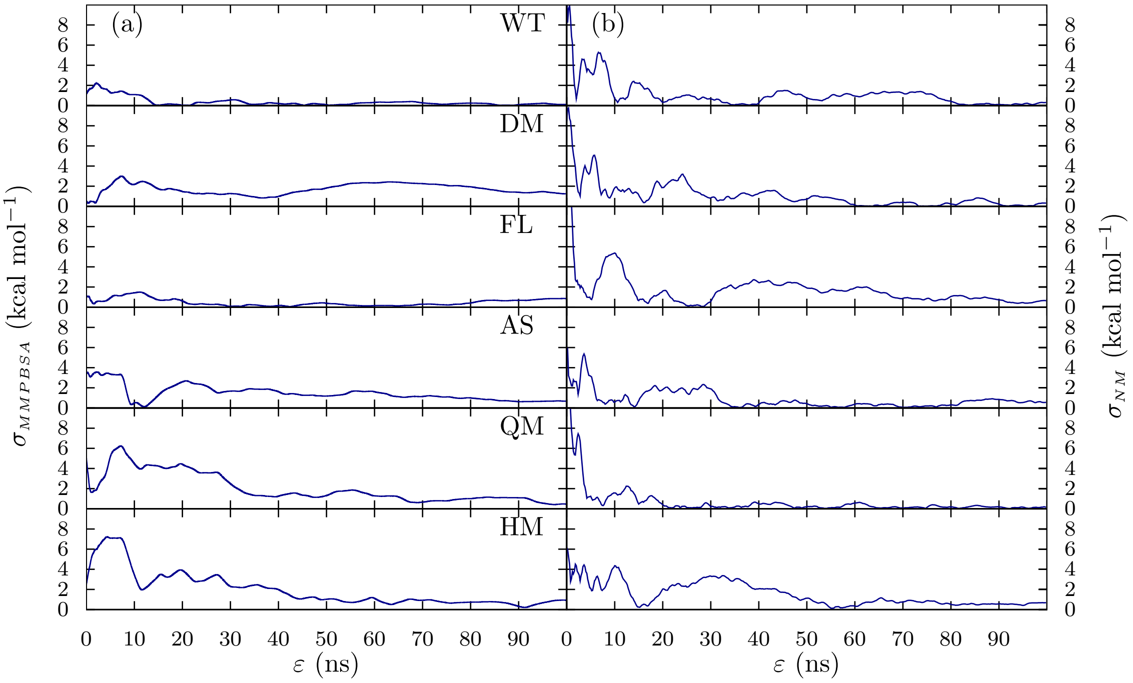


Figure 5.12: RMS difference in cumulative means, $\sigma(\varepsilon)$, of (a) the mainly enthalpic, ΔG_{MMPBSA}, and (b) the entropic, $-T\Delta S_{NM}$, components of the binding free energy for the 50 × 4 ns concatenated trajectories.

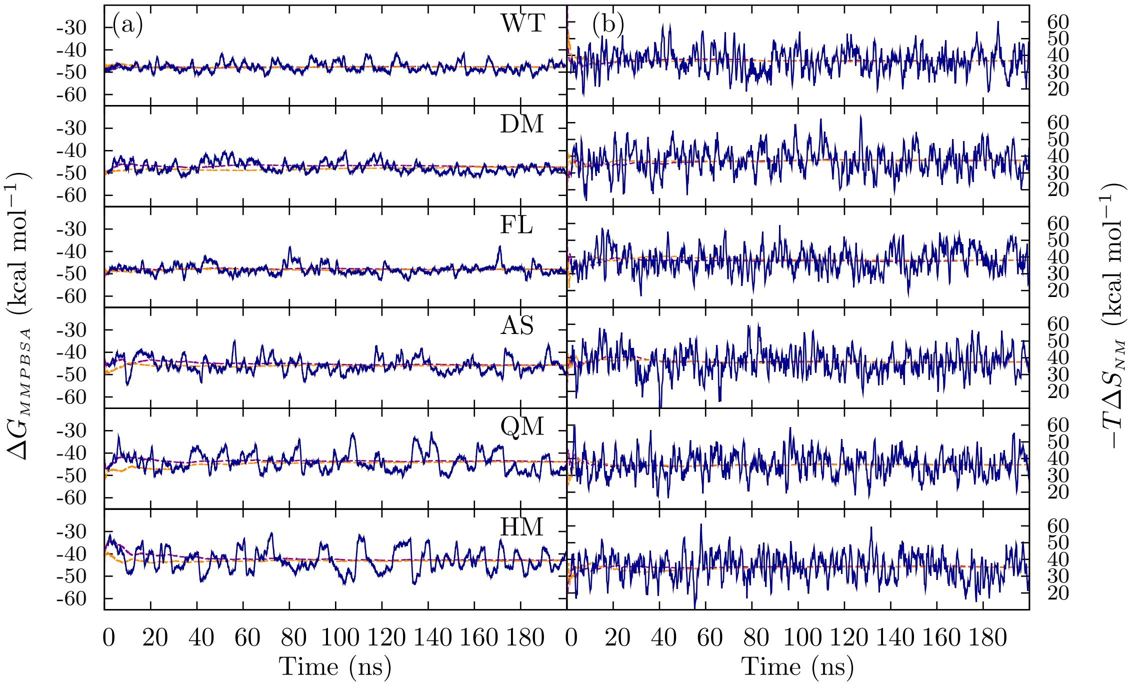


Figure 5.13: Convergence of (a) the enthalpically dominated $\Delta G_{MMPBSA}$and (b) the entropic $-T\Delta S_{NM}$component of the binding free energy difference for the 50 × 4 ns strategy. Forward and reverse cumulative means are shown in orange and magenta respectively, 1ns running means are shown in blue.

5.5 Absolute and Relative Binding Free Energy Rankings

The change from single simulation to ensemble simulations brought a marked improvement in both ranking and the consistency of the deviation from experiment, the extension of the individual replica length within the ensemble produces further, if only incremental, benefits. This improvement can clearly be seen when the theoretical values produced by each strategy are plotted against those obtained experimentally (see Figure 5.14). Unlike those for the shorter 50 × 1 ns ensemble the absolute binding affinity, ΔG_{theor}, values from the 50 × 4 ns dataset reproduce the experimental rank order of the sequences (see Figure 5.14b and Figure 5.14d), which is reflected in the improved correlation coefficient, κ, of 0.89 (compared to 0.55). The increase in replica length has little impact upon the quality of the ΔG_{MMPBSA} rankings (see Figure 5.14a and Figure 5.14c) and the correlation coefficient is in fact unchanged (at 0.99).

The comparison of the relative binding free energy differences, $\Delta\Delta G_{theor}$, in Figure 5.15 shows how the ranking of each mutant relative to the wildtype calculation has been improved by the extension of the replicas within the ensemble to 4 ns. The over estimate of the resistance of the AS, QM and HM seen in the 50 × 1 ns results is reduced and the DM mutant is now found to be less attractive than WT. Interestingly, this improvement is due to the enhanced entropic sampling. The ΔG_{MMPBSA} ranking of these mutants relative to WT is changed little in the longer ensemble compared to the shorter one. The relative ranking of each mutant relative to wildtype is now within approximately 1.1 kcal mol^{-1} of that seen experimentally for all sequences (except AS where the deviation is around 1.52 kcal mol^{-1}) with a mean deviation across all systems of 0.9 kcal mol^{-1} (see Table 5.8). Despite the improvements bought about by the extension of the ensemble they do not reproduce the superadditive behaviour seen in the experimental results. Superadditivity is, however, exhibited by the $\Delta\Delta G_{MMPBSA}$ values. Particular care is required in interpreting the DM and FL results. In the experimental results both are just distinguishable from WT (differing by only 0.1 kcal mol^{-1}) and indistinguishable from one another. The $\Delta\Delta G_{MMPBSA}$ results computed here indicate that both mutants are very close to the WT value (the DM marginally less attractive, the FL slightly more attractive). The disparity in the absolute binding affinities indicate that both sequences are slightly resistant, with The $\Delta\Delta G_{theor}$ of approximately 1 kcal mol^{-1}.

Table 5.7: Computed free energy differences of binding (ΔG_{theor}) compared with experimental results (ΔG_{exp}) for all six HIV-1 protease sequences in the MDR Test Set with LPV and for wildtype binding to saquinavir using 50×4 ns ensemble-trajectory runs. The enthalpically dominated MMPBSA and the normal mode entropic components are also shown. Correlation coefficients, κ, are provided for each theoretically computed data set compared to the experimental data.

Sequence	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{-T\Delta S_{NM}}$	$\mathbf{\Delta G_{theor}}$	$\mathbf{\Delta G^{*}_{exp}}$	$\mathbf{\Delta\Delta G^{*}_{theor-exp}}$
			Ensemble (50 × 4 ns)		
WT	-47.68 (0.03)	36.74 (0.49)	-10.94 (0.52)	-15.1 (0.09)	4.16 (0.61)
HM	-42.86 (0.05)	35.45 (0.52)	-7.41 (0.57)	-11.3 (0.08)	3.89 (0.65)
QM	-43.93 (0.05)	36.35 (0.51)	-7.58 (0.56)	-12.8 (0.04)	5.22 (0.60)
AS	-45.75 (0.04)	37.53 (0.54)	-8.22 (0.58)	-13.9 (0.10)	5.68 (0.68)
FL	-48.01 (0.04)	38.08 (0.50)	-9.93 (0.54)	-14.9 (0.09)	4.97 (0.63)
DM	-47.37 (0.03)	37.39 (0.51)	-9.98 (0.54)	-14.9 (0.05)	4.92 (0.59)
κ	0.98		0.89		
SAQ-WT	-44.20 (0.04)	36.30 (0.53)	-7.90 (0.57)	-13.0 (0.04)	5.10 (0.61)

*Experimental results are taken from Ohtaka *et al.* [92].

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

The full details of the binding affinities computed from the 50 × 4 ns ensemble are shown in Table 5.7. The consistent deviation of approximately 5 kcal mol^{-1} (the RMS deviation from the experimental values is reduced from 5.11 to 4.85 kcal mol^{-1}with the extension of replica length from 1 ns to 4 ns) of all of the ΔG_{theor} values from the experimental values is suggestive of a possible systematic error. The limitations of the end point approximation underlying the MMPBSA method employed here offer a potential source of this discrepancy. This description cannot account for changes in the state of the system upon binding. In the HIV-1 protease system a number of such changes have been identified, and their free energy contributions evaluated, including the closing of the flaps upon ligand binding, alteration of the catalytic dyad protonation state and water mediated interactions. Two studies have investigated the free energy penalty of

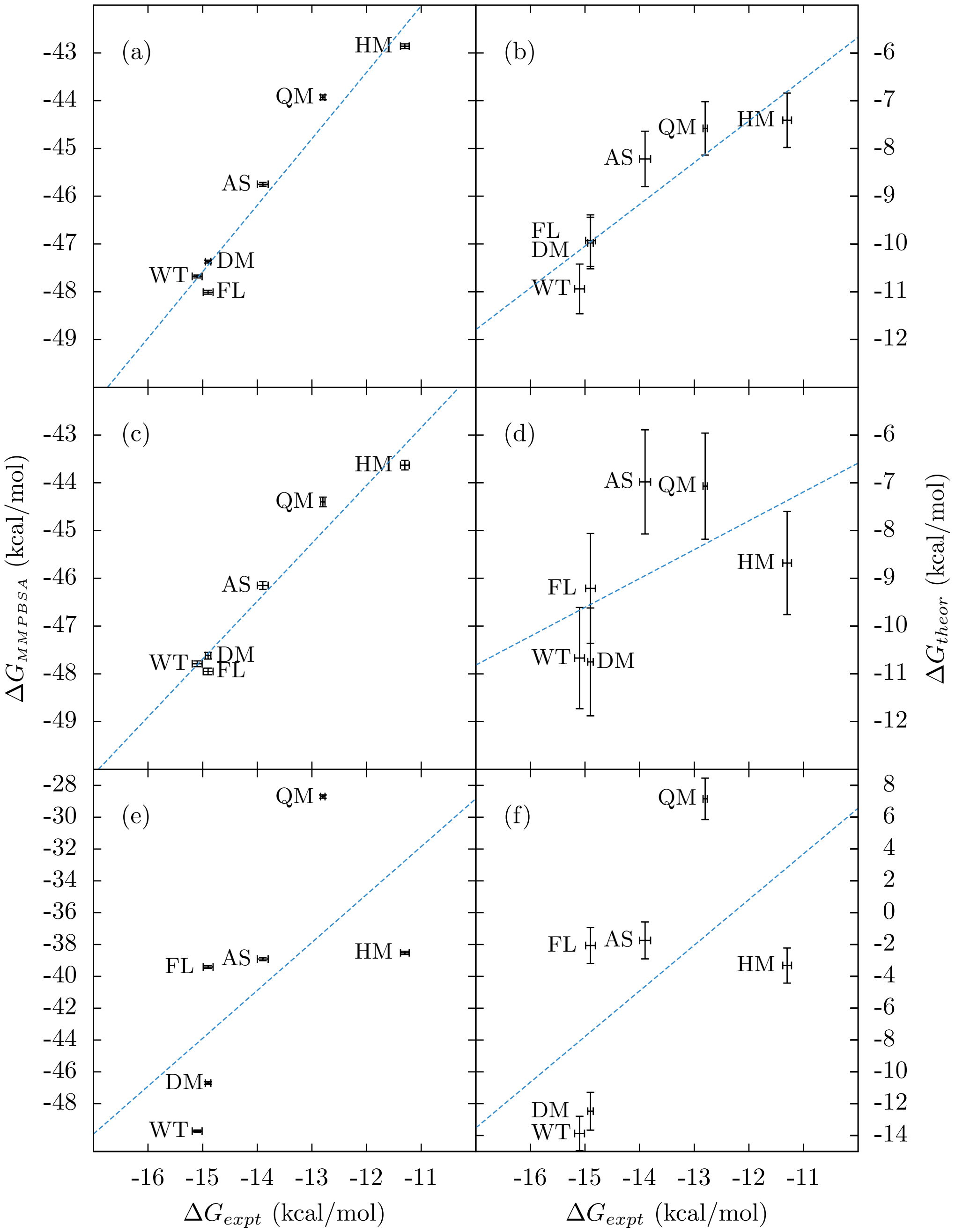


Figure 5.14: Comparison of the theoretical binding free energies from all three simulation strategies with the experimental values for each sequence in the MDR Test Set. The enthalpically dominated ΔG_{MMPBSA} and absolute ΔG_{theor} values are shown for each strategy: (a) & (b) 50 × 4 ns strategy; (c) & (d) 50 × 1 ns and (e) & (f) 1 × 50 ns. The error bars represent standard errors.

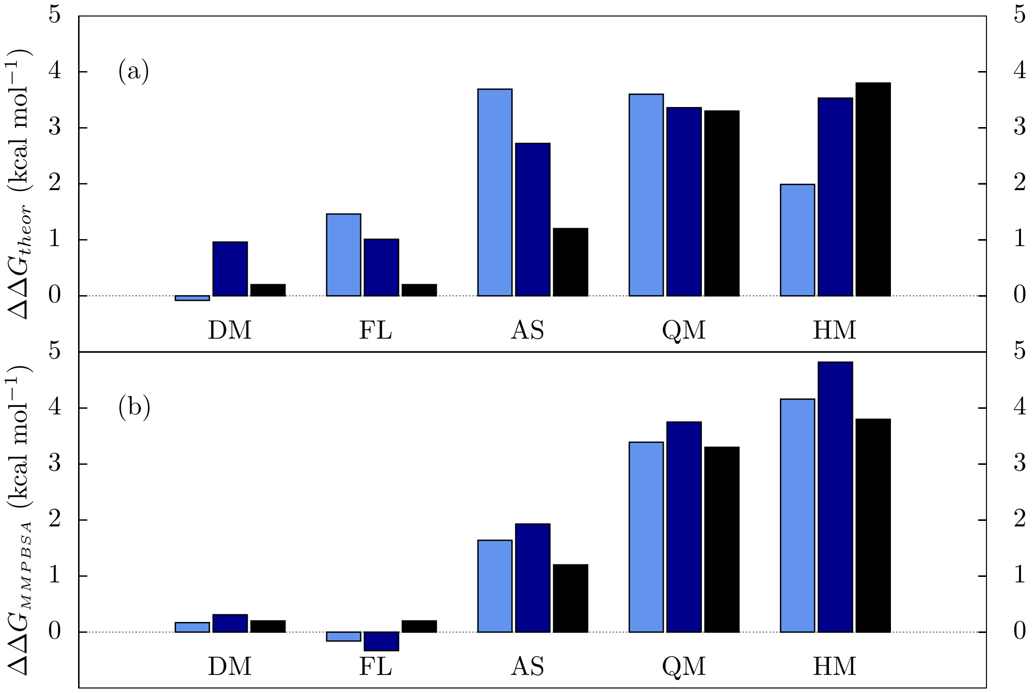


Figure 5.15: Comparison of the theoretical and experimental relative free energy differences for both the (a) the overall binding free energies and (b) the enthalpically dominated free energy calculated using MMPBSA. The values for the 50 × 1 ns and 50 × 4 ns ensembles are shown in light and dark blue respectively with the experimental results in black.

changing the conformation of the protease flaps from semi-open to closed upon ligand binding. The earliest of these calculated the free energy difference between these states from umbrella sampling molecular dynamics simulations using the potential of mean force [367]. Depending upon the choice of reaction path the penalty was calculated as 2 ± 2, 6 ± 6 or 13 ± 5 kcal mol^{-1}. The later estimate in particular appears physically unreasonable, as the conformational change of the glycine rich flaps appears to result in only minor overall reductions of the hydrophobic interactions (the loss of same monomer interactions in the transition is compensated by increases in those between the monomers). A more recent calculation, using μs scale ensemble molecular dynamics simulations, estimates the change at a more feasible 2.4 ± 0.4 kcal mol^{-1}[229] in agreement with the lowest estimate from the earlier study. The protonation state of the catalytic dyad is believed to be highly dependent on the local chemical environment [247–250]. At physiological pH the binding of an inhibitor is thought to cause a change from a dianionic to a monoprotonated state of the catalytic dyad [368]. This change is believed to elicit a favourable free energy change of 1 to 2 kcal mol^{-1}[362]. A final significant contribution comes from the favourable contribution of WAT301 bound between the inhibitor and flaps. Thermodynamic integration studies of this contribution for a variety of inhibitors estimate it between 3 to 3.5 kcal mol^{-1}[369, 370] and structural refinement based studies suggest it is between 4 and 6 kcal mol^{-1}[371]. The sum of these terms gives a correction of between 2 and 5.1 kcal mol^{-1}, in good agreement with the deviation of our results from the experimental values.

5.5.1 Thermodynamic Decomposition

Despite the difficulties in interpreting the results for the mutants with similar binding affinities to wildtype the fact that the distinction between these sequences and the highly resistant mutants (AS, QM and HM) is present in both the ΔG_{MMPBSA} and ΔG_{theor} rankings means that decomposition of the former quantity may still offer insight into the origin of the exhibited resistance. Table 5.9 shows the decomposition of ΔG_{MMPBSA} into van der Waals (ΔG^{MM}_{vdw}), electrostatic (ΔG^{MM}_{ele}), and polar (ΔG^{sol}_{pol}) and non polar (ΔG^{sol}_{nonpol}) solvation terms. Across all sequences a similar breakdown is seen. The binding is primarily driven by highly favourable van der Waals interactions. This is supplemented by the much smaller contribution of the non polar solvation and partially compensated by the net electrostatic repulsion (composed of favourable vacuum electrostatics and an unfavourable polar solvation contribution). All of the highly resistant sequences show a reduction in the attractiveness of the van der Waals component (with the biggest change being of 2.4 kcal mol^{-1} in the HM system). This difference is likely to be predominantly explained by the direct reduction of hydrophobic interactions caused by the V82A and I84V substitutions common to all three sequences. The net electrostatic component also becomes more repulsive in the HM and QM systems.

5.5.2 Reproducibility

The subtle differences in the binding affinities of several of the mutants in the MDR Test Set highlights the importance of assessments of the accuracy in both parts of the calculation. The convergence analysis already presented can only assess the internal consistency of the data. Confidence in the comparison of values computed from different data sets, such as those for different sequences, requires an understanding of the accuracy of the method employed for a given sample size. In order to gain some quantitative insight into this, a reproducibility analysis was performed upon the WT and HM systems.

The results of the reproducibility analysis are presented in Table 5.10. In the 50 × 1 ns ensemble the variance in the MMPBSA derived binding affinity for both the WT and HM systems was around 5 kcal mol^{-1}, there was a large discrepancy in the reproducibility of the entropic component however, with a difference of 0.05 for the WT compared to 2.03 for HM. Overall the error in the absolute free energy differences was 0.43 kcal mol^{-1}for WT and 1.50 kcal mol^{-1}for the HM system. The increase in replica length actually

Table 5.8: Computed relative binding free energy differences ($\Delta\Delta G_{theor}$) compared with the experimental results ($\Delta\Delta G_{exp}$) for all HIV-1 protease sequences in the MDR Test Set compared to wildtype, using the 1 × 50 ns, 50 × 1 ns and 50 × 4 ns data sets. Relative free energy differences from the enthalpically dominated MMPBSA calculation ($\Delta\Delta G_{MMPBSA}$) and the correlation coefficients (κ) of each theoretical data set compared to experiment are also shown.

		1 × 50 ns		**50 × 1 ns**		**50 × 4 ns**	
Sequence	$\mathbf{\Delta\Delta G^{*}_{exp}}$	$\mathbf{\Delta\Delta G_{MMPBSA}}$	$\mathbf{\Delta\Delta G_{theor}}$	$\mathbf{\Delta\Delta G_{MMPBSA}}$	$\mathbf{\Delta\Delta G_{theor}}$	$\mathbf{\Delta\Delta G_{MMPBSA}}$	$\mathbf{\Delta\Delta G_{theor}}$
HM	3.8 (0.17)	11.2 (0.17)	10.55 (2.18)	4.16 (0.16)	1.99 (1.18)	4.82 (0.08)	3.53 (1.09)
QM	2.3 (0.13)	21.01 (0.17)	21.02 (2.38)	3.39 (0.16)	3.60 (1.21)	3.75 (0.08)	3.36 (1.08)
AS	1.2 (1.19)	10.81 (0.17)	12.13 (2.24)	1.64 (0.14)	3.69 (1.17)	1.93 (0.07)	2.72 (1.10)
FL	0.2 (0.18)	10.31 (0.16)	11.81 (2.21)	-0.16 (0.13)	1.46 (1.22)	-0.33 (0.06)	0.96 (1.06)
DM	0.2 (0.14)	3.02 (0.15)	1.40 (2.27)	0.17 (0.13)	-0.08 (1.20)	0.31 (0.06)	0.96 (1.06)
κ	-	0.62	0.56	0.98	0.55	0.98	0.89

*Experimental results are taken from Ohtaka *et al.* [92].

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

Table 5.9: Decomposed contributions to the free energy of binding for wildtype and all MDR proteases with lopinavir as well as wildtype with saquinavir using 50 × 4 ns ensemble-trajectory runs.

Sequence	$\mathbf{\Delta G^{MM}_{vdw}}$	$\mathbf{\Delta G^{MM}_{ele}}$	$\mathbf{\Delta G^{sol}_{pol}}$	$\mathbf{\Delta G^{sol}_{nonpol}}$	$\mathbf{\Delta G^{tot}_{ele}}$	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{-T\Delta S_{NM}}$
WT	-72.82 (0.03)	-51.29 (0.05)	84.56 (0.05)	-8.13 (0.00)	33.27 (0.04)	-47.68 (0.03)	36.74 (0.49)
HM	-70.43 (0.04)	-50.51 (0.08)	86.17 (0.06)	-8.10 (0.00)	35.66 (0.05)	-42.86 (0.05)	35.45 (0.52)
QM	-71.62 (0.03)	-52.03 (0.08)	87.88 (0.06)	-8.15 (0.00)	35.85 (0.05)	-43.93 (0.05)	36.35 (0.51)
AS	-71.02 (0.03)	-51.97 (0.07)	85.38 (0.05)	-8.13 (0.00)	33.41 (0.04)	-45.75 (0.04)	37.53 (0.54)
FL	-73.11 (0.03)	-52.02 (0.06)	85.26 (0.05)	-8.14 (0.00)	33.24 (0.04)	-48.01 (0.04)	38.08 (0.50)
DM	-72.71 (0.03)	-52.29 (0.05)	85.76 (0.05)	-8.13 (0.00)	33.47 (0.04)	-47.37 (0.03)	37.39 (0.51)
SAQ-WT	-75.79 (0.03)	-49.18 (0.06)	89.08 (0.06)	-8.32 (0.00)	39.91 (0.04)	-44.20 (0.04)	36.30 (0.53)

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

produces an small increase in the error, to 0.50 kcal mol^{-1}, for the WT system (with the majority of the discrepancy again to be found in ΔG_{MMPBSA}). In contrast for the HM sequence both the MMPBSA and configurational entropy components are found to be more reproducible in the 50 × 4 ns ensemble, resulting in an reduced error of 0.82 kcal mol^{-1}in ΔG_{theor}. The net effect of this variation is to produce an uncertainty in the relative binding affinities of the two systems of approximately 1.3 kcal mol^{-1}. This result suggests that in order for us to have confidence in the relative ranking of two mutants the computed binding affinities must vary above this threshold.

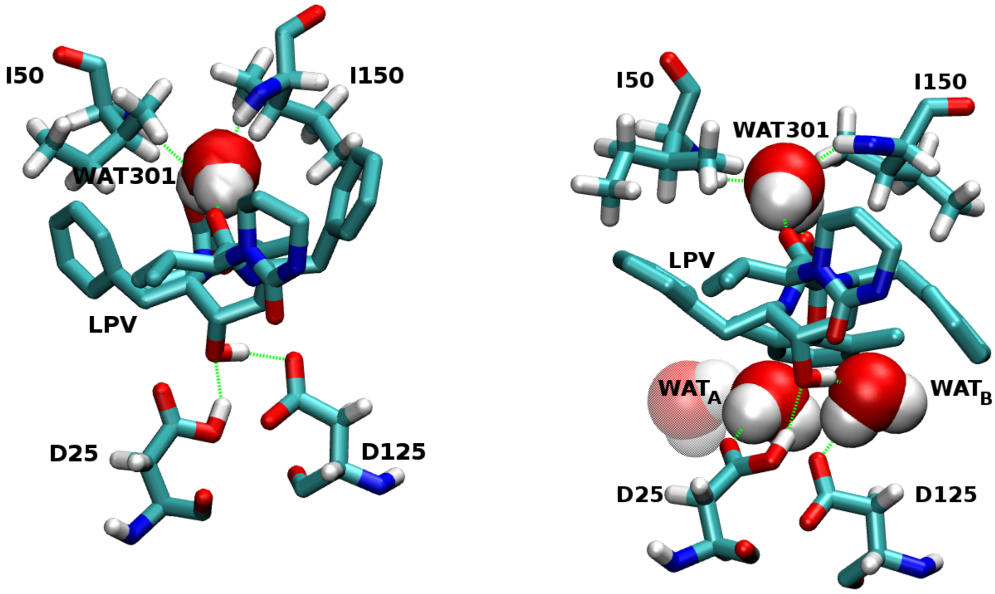
Table 5.10: Reproducibility of ensemble calculations for 50×1 ns and 50×4 ns ensemble-trajectory runs. Two ensemble simulations, labelled I and II, were performed for both the WT and HM sequences. Ensemble I is that used in the intra-sequence comparisons presented elsewhere in this chapter.

Sequence	**Sample**	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{-T\Delta S_{NM}}$	$\mathbf{\Delta G_{theor}}$	$\mathbf{\Delta\Delta G_{II-I}}$
Ensemble (50 × 1 ns)					
WT	I	-47.85 (0.05)	36.61 (0.68)	-11.24 (0.73)	-
	II	-48.33 (0.05)	36.66 (0.71)	-11.67 (0.76)	-0.43 (0.76)
HM	I	-43.63 (0.10)	34.95 (0.98)	-8.68 (1.08)	-
	II	-44.16 (0.07)	36.98 (0.74)	-7.18 (0.81)	1.50 (1.89)
Ensemble (50 × 4 ns)					
WT	I	-47.68 (0.03)	36.74 (0.50)	-10.94 (0.53)	-
	II	-48.38 (0.03)	36.94 (0.51)	-11.44 (0.54)	-0.50 (1.07)
HM	I	-42.86 (0.05)	35.45 (0.52)	-7.41 (0.57)	-
	II	-43.36 (0.05)	36.77 (0.52)	-6.59 (0.57)	0.82 (1.14)

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

5.6 Structural Correlates of Resistance

The observation of the impact of water ingress into the active site of the HIV-1 protease in the 1 × 50 ns simulations suggests that the frequency of such events may play an important role in determining the free energy differences observed in the ensemble simulations too. As noted previously no additional water molecules enter the active site in the vicinity of the conserved WAT301 in any of the 50 × 1 ns ensemble simulations. This remains the case in the extended ensemble. However, water entry is observed in many trajectories in the region between the hydroxyethylene moiety of lopinavir and the catalytic dyad (close to the position posited for the entry of a water molecule during the lysis of natural polypeptide substrates). Figure 5.16a shows the minimised initial structure where a single water molecule in the WAT301 position and the catalytic dyad both form hydrogen bonds with the bound lopinavir. The presence of water molecules in this region between the catalytic dyad disrupts the hydrogen bonding networks between the drug and D25 and D125. The occupancy of the hydrogen bonds formed between lopinavir and the protease in each system is shown in Table 5.11. The influence of water



(a) (b)

Figure 5.16: Conformations of the catalytic cavity of the HM mutant bound to lopinavir for (a) the energy minimised initial structure and (b) a representative snapshot showing water mediated alteration of the hydrogen bond network in the 1 × 4 ns ensemble. The conserved water molecule which mediates hydrogen bonds between the flaps and ligand is labelled WAT301. Two other water molecules mediating protein-inhibitor interactions in the catalytic cavity are labelled WAT$_A$ and WAT$_B$. Water molecules shown in transparent representations are present but are not involved directly in any interactions between drug and enzyme. Potential hydrogen bonds are shown as green lines.

entry in the more resistant mutants is shown in the reduced occupancy of the bonds between D125, G27 and the oxygen in the hydroxyethylene moiety of lopinavir. In the mutant systems, in particular the AS and QM sequences, this change results in D125 being free to more frequently form bonds with the nearby lopinavir backbone nitrogen (N4). Away from the active site, and the direct influence of water molecules in the catalytic cavity, the bonds between residues 29 and 30 in the P2′ subsite and the inhibitor are also altered, reduced in the resistant mutant systems in the former case and increased in the latter.

A typical conformation of the catalytic cavity after water entry is shown in Figure 5.16b. In this conformation D25 maintains it's hydrogen bond to the central oxygen of lopinavir but WAT$_B$ disrupts the bond between D125 and this moiety of the inhibitor. WAT$_B$ now mediates this interaction forming hydrogen bonds with both the D125 and lopinavir. An additional water molecule, WAT$_A$ in the picture, mediates interactions between D25 and the backbone N4 nitrogen of lopinavir. The figure also shows a further water molecule present in the catalytic cavity but which despite being within 3 Å of both protein and ligand does not obviously appear to be involved in any specific interaction between the two.

The population density of water molecules within 3 Å of the catalytic dyad was calculated for each of the systems in the MDR Test Set. Comparison with the theoretical and experimental binding free energy differences yields excellent correlation coefficients of 0.89 and 0.98 respectively (see Table 5.12). This is indicative of water mediated interactions playing a significant role in the causation of drug resistance. The rate of water entry is also a likely consequence of alterations of the size and shape of the active site in the MDR mutants relative to wildtype. In a recent study [372] a similar mutant to the AS sequence in this study, containing A82F and I84V substitutions, was seen to exhibit significant deformation of the active site geometry. In the case of the MDR mutants structural changes in the active site volume caused or induced by the mutations may explain both the increased accessibility of the catalytic cavity to water molecules and the progressive decrease in inhibitor binding. The disruption of protein-ligand interactions resulting from the presence of water molecules may also help to explain the increase in the net electrostatic repulsion highlighted by the thermodynamic decomposition of the binding free energy.

5.7 Cross Drug Thermodynamic Ranking

A further test of the ensemble methodology was made by simulating the first generation inhibitor saquinavir (SAQ) bound to the WT sequence. Experimentally saquinavir is seen to be bound 2.1 kcal mol^{-1} less strongly than lopinavir(LPV), a difference similar to that between WT and AS bound to LPV. The theoretical binding free energy difference, shown in Table 5.7, is -7.9 kcal mol^{-1} which is 5.1 kcal mol^{-1} less attractive than the experimental value. This difference is consistent with the differences seen in the other systems reported here. The difference between the value for SAQ and LPV bound to WT is 3.04 kcal mol^{-1} which is in good agreement with experiment and is similar to the difference between WT and AS in the theoretical LPV data set. These findings suggest that the methodology presented here can distinguish between drugs with a similar accuracy, of approximately 1 kcal mol^{-1}, to that demonstrated between different sequences bound to the same inhibitor.

Decomposition of ΔG_{theor} (shown in Table 5.9) suggests that the difference in binding affinity is primarily due to changes in the enthalpically dominated ΔG_{MMPBSA} component, with the configurational entropy changes within a standard error of one another

Table 5.11: The frequency of occupation of putative hydrogen bonds between the inhibitor lopinavir (LPV) and the HIV-1 protease sequences in the MDR Test Set calculated over the production simulations of the 50 × 4 ns dataset. A hydrogen bond is identified as being formed when a donor acceptor pair are within 3.5 Å of one another and the donor-hydrogen-acceptor angle is less that 120 degrees. The oxygen and nitrogen atoms of LPV are numbered from the left of the schematic shown in Figure 5.2a. Only bonds which were occupied for more than 10% of at least one studied sequence are listed.

	Donor	**LPV O3**	**LPV N1**	**LPV N3**	**LPV N4**	**LPV O3**	**LPV N1**	**29 N**	**30 N**
Sequence	**Acceptor**	**125 OD2**	**29 OD1**	**27 O**	**125 OD2**	**125 OD1**	**29 OD2**	**LPV O1**	**LPV O1**
WT		0.945	0.772	0.496	0.205	0.179	0.134	0.999	0.051
HM		0.819	0.258	0.198	0.408	0.247	0.199	0.960	0.115
QM		0.846	0.240	0.105	0.567	0.245	0.286	0.985	0.177
AS		0.931	0.418	0.234	0.535	0.212	0.235	0.986	0.160
FL		0.935	0.737	0.472	0.246	0.210	0.138	0.999	0.054
DM		0.948	0.655	0.479	0.320	0.235	0.179	0.992	0.094

Table 5.12: Comparison of the computed relative free energy differences of binding (ΔG_{theor}) and experimental results (ΔG_{expt}) with the population density of catalytic water occupancy (ρ) for all sequences within the MDR Test Set with lopinavir using 50 × 4 ns ensemble trajectory runs. The values for both the main 50 × 4 ns data set (labelled Ensemble I) and those run for the reproducibility analysis (Ensemble II) are given for the WT and HM sequences. Correlation coefficients of the water occupancy relative to the theoretical (κ_{theor}) and experimental (κ_{exp}) free energy values are also provided (the coefficients listed for Ensemble II are calculated with the values for FL, DM, AS and QM taken from Ensemble I).

		50 × 4 ns - Ensemble I		**50 × 4 ns - Ensemble II**	
Sequence	$\mathbf{\Delta G^*_{exp}}$	$\mathbf{\Delta G_{theor}}$	ρ	$\mathbf{\Delta G_{theor}}$	ρ
WT	-15.1 (0.09)	-10.94 (0.52)	0.057	-11.44 (0.54)	0.020
HM	-11.3 (0.08)	-7.41 (0.57)	0.484	-6.59 (0.57)	0.335
QM	-12.8 (0.04)	-7.58 (0.56)	0.361	-	-
AS	-13.9 (0.10)	-8.22 (0.58)	0.167	-	-
FL	-14.9 (0.09)	-9.93 (0.54)	0.042	-	-
DM	-14.9 (0.05)	-9.98 (0.54)	0.039	-	-
κ_{exp}			0.99		0.93
κ_{theor}			0.89		0.92

*Experimental results are taken from Ohtaka *et al.* [92].
Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

(suggesting similar levels of flexibility change upon binding). The ΔG_{MMPBSA} component is 3.48 kcal mol^{-1} less attractive for SAQ bound to WT compared to LPV. The origin of this difference is a greater polar solvation penalty (89.08 compared to 84.56 kcal mol^{-1}) which is only partially compensated by more favourable van der Waals interactions (-75.79 compared to -77.82 kcal mol^{-1}).

5.8 Conclusions

This study has explored the effectiveness of the approximate MMPBSA and normal mode methodology to reproduce the experimental binding affinities of six multi-drug resistant (MDR) HIV-1 protease mutants. The correct ranking was obtained with a correction coefficient of 0.89 and a mean deviation in the relative ranking of only 0.9 kcal mol^{-1}. The theoretical absolute binding affinities exhibit a systematically less attractive binding free energy by approximately 5 kcal mol^{-1} compared to the experimental results. This can be explained as originating from a combination of contributions not accounted for in the MMPBSA approach; conformational changes of the flaps between semi-open and closed positions, alteration of the catalytic dyad protonation state and the contribution of a water molecule bound between the inhibitor and flaps (known as WAT301). Summing the estimates of these contributions from other studies provides a correction of between 2 and 5.2 kcal mol^{-1} in good agreement with the deviation of the theoretical results calculated in the study presented here.

In order to obtain these results an ensemble of 50 replica simulations, each of 4 nanoseconds duration (varying only in the initial velocities assigned to each atom), was used. A comparison of ensemble and single long trajectory simulations producing the same length of production simulation showed that the ensemble approach samples relevant areas of phase space more efficiently. The single trajectories are frequently trapped in single minima resulting in inconsistent sampling and poor convergence. The ensemble strategy, however, obtained accurate and converged results not possible with the single trajectory approach. Reproducibility analysis of the 50 × 4 ns ΔG_{theor} values for the susceptible WT and highly resistant HM sequences suggests that the ranking of systems can reliably be made as long as the difference in binding free energies is greater than 1.3 kcal mol^{-1}.

A cross drug ranking of the wildtype sequence bound to lopinavir and the less potent first generation inhibitor saquinavir was also successfully performed. In both the cross drug and mutant ranking the majority of the inter-system differences are attributable to the enthalpically dominated ΔG_{MMPBSA} component of the calculation. The reduced affinity of saquinavir is a consequence of a higher polar solvation penalty. The most significant contributions to MDR resistance was attributable to direct reduction of the van der Waals interactions by the V82A/I84V mutational pair (which has an impact of up to 2.5 kcal mol^{-1}) and increased electrostatic repulsion induced by water mediated alteration of the hydrogen bond network in the catalytic cavity (creating a maximum change of 2.5kcal mol^{-1}). The convergence analysis of the configurational entropy component computed using normal mode analysis suggests that systems where this is the differentiating factor may be less accurately calculated using this methodology. Even the extensive sampling performed in this study (using 1000 snapshots) could not fully converge the normal mode calculation, with grouped peaks still apparent. It appears that this is an inherent limitation of the methodology.

In order to obtain the sampling required to gain the accurate, converged results presented here it was necessary to perform 200 ns of fully atomistic simulation. It is likely that these computations represent the computational limit of approximate thermodynamic end point methods such as MMPBSA. At this level we can discriminate systems with binding affinities within 1 kcal mol^{-1} of one another. In order to get beyond the accuracy presented here it is probable that more exact methods (such as thermodynamic integration) are required. The advantage of the MMPBSA methodology is, however, that theoretically the turn around time for the entire study presented here could be three days (assuming that 9600 cores on a modern supercomputer such as Ranger were available simultaneously). This rapid execution time highlights the potential of these methods to play a role in patient specific clinical decision support systems designed to optimise treatment choice in situations in which mutations to the target protein affects intervention efficacy.

Chapter 6

Virtual Patient Experiment

6.1 Introduction

The rapid acquisition of mutations conferring resistance to particular drugs remains a significant problem in the treatment of HIV infection, potentially decreasing both the magnitude and duration of the response to treatment [373]. Even for expert clinicians, it is frequently impossible to identify straightforward relationships between genotype and drug response. This has resulted in the production and use of computer based clinical decision support systems (CDSS). Amongst the most popular of these are those produced by the Stanford HIVdb[1], ANRS[2] and RegaDB[3]. These systems use data collected from patient databases and the published literature in order to give resistance scores to individual mutations which can be combined additively to assess the resistance levels of complete sequences. Clinicians then use the assigned levels of resistance to different drugs to select a drug cocktail suitable for the viral sequence present in each individual patient. Several studies have shown that the use of such genotypic resistance analysis to guide the selection of drugs within a HAART regimen improves virological outcomes [207, 208, 374–376].

The ANRS and Rega systems are rule based algorithms that both report three levels of resistance: susceptible, resistant, and an intermediate level (the definition of which is different for each algorithm). The HIVdb algorithm assigns a drug penalty score for each drug resistance mutation. Summing the contributions from each mutation provides an overall score which is then converted into one of the following levels of inferred drug resistance: susceptible, potential low-level resistance, low-level resistance, intermediate resistance, and high-level resistance. A variety of studies have compared the performance

[1] HIVdb: hivdb.stanford.edu

[2] ANRS: www.anrs.fr

[3] RegaDB: www.rega.kuleuven.be/cev/regadb/

of these popular, freely available, systems at predicting *in vivo* virological response and found them to be generally reliable [209–213]. Recently, a large study (looking at over 3000 treatment change episodes) by Frentz *et al.* [377] found little difference between the performance of HIVdb, ANRS and RegaDB in predicting undetectable viral load at a variety of time points after treatment initiation. This finding is perhaps unsurprising, given that all of these systems rely on the same literature as the source from which their rules are derived. However, detailed analysis of four widely used prediction systems (those previously mentioned and the less commonly studied Visible Genetics version 6[4] [378]) reported that for a significant number of sequences the systems disagree on the level of resistance to be expected [379]. The majority of tested sequences, 66.4%, produced concordant results from all systems, whilst results for 4.4% of sequences resulted in complete discordance (with at least one system suggesting susceptibility and another resistance) and the remaining 29.2% showed partial discordance (i.e. minor differences are present in the level of resistance assigned by each system). Whilst, the overall success of the systems indicates that these discordances are likely to be rare nonetheless they represent a potential cause of suboptimal treatment choice for individual patients.

One factor complicating the assignment of resistance scores to sequences is the presence of interactions between mutations that cause non-additive effects on phenotype and fitness. Such interactions between mutations are termed ‘epistasis’. The term is often used in the context of interactions between whole genes whose phenotype is altered by mutation; in cases such as this where the relationship is between point mutations in the same gene the term ‘intragenic epistasis’ is used to prevent ambiguity. Intragenic epistatic effects are likely to play an important role in determining the level of viral resistance [269, 280, 380]. The inclusion of insight into combinatorial mutational effects from a broader range of sources, including computational modelling, in decision support software offers the potential to further improve automated treatment [381]. Motivated by this belief the EU ViroLab project[5] created a virtual laboratory consisting of complementary, multilevel computational tools aimed at comparing and enhancing the existing repertoire of decision support approaches [382–384]. One of the simulation techniques chosen to supplement the traditional resistance assignment tools was molecular dynamics (MD). As seen in Chapter 5, molecular dynamics offers the ability to derive quantitative, as well as qualitative, insight into the interplay of resistance causing mutations. The facility to investigate the drug resistance phenotype of the particular strain of HIV infecting a patient without the need for costly experiments by using *in*

[4] Since the publication of the study Visible Genetics has been bought by Bayer Diagnostics (`http://www.bayerhealthcare.com`) and their resistance assessment software incorporated into the TRUGENE HIV-1 genotyping test and OpenGene sequencing systems.

[5] ViroLab: `www.virolab.org`

silico homology modelling, simulation and free energy calculations offers an attractive option to supplement existing CDSS.

In this chapter, we describe the use of MD simulations and free energy calculations to investigate an instance where the HIVdb, ANRS and Rega prediction systems disagree on the resistance levels produced by an HIV-1 protease derived from a real patient. The performance of these simulations requires the use of substantial computational resources and the management of large amounts of data, these requirements have prompted the development of an automated simulation pipeline called the Binding Affinity Calculator (BAC) [355]. The work flow from the creation of the system to be simulated to the final analysis can be executed across multiple computational resources by making use of two further tools developed within the ViroLab project, the Application Hosting Environment (AHE) [385] and GridSpace Engine [386]. Molecular level simulations are not only vastly cheaper than wet lab alternatives but potentially offer the ability to generate atomistic understanding of the causes of resistance. This represents the lowest level of the ViroLab philosophy of creating a multiscale holistic approach to decision support, “from molecule to man” [384].

6.1.1 Virtual Patient Experiment

The ViroLab virtual laboratory (VL) contains a wealth of tools for investigating the relationship between HIV genomic sequence and the level of resistance to anti retroviral drugs [382]. To show the potential of integrating diverse systems such as traditional drug ranking systems, literature mining, patient data and molecular simulations into a single interface the Virtual Patient Experiment (VPE) was designed. The aim of the VPE was to take a patient sequence for which the Virolab comparative drug ranking system (cDRS) provided discordant results for one of the available protease inhibitors and to use the other tools within the virtual laboratory to produce the sort of insight that could help a clinician who was to be facing a decision on how to treat this virtual patient. The cDRS allows the user to simultaneously obtain drug resistance rankings (susceptible, intermediate or resistant) for an input sequence or set of mutations from three well established drug ranking systems: Stanford HIVdb, ANRS and RegaDB.[6] The Virolab and EuResist[7] databases were queried to find patient sequences that met these criteria resulting in the choice of a sequence containing the mutations L10I, I13V, K14KR, I15V, K20T, L63P, A71IV, V77IV, L90M, I93L in combination with the drug lopinavir. This sequence was deemed to be susceptible by HIVdb but displayed intermediate resistance according to ANRS and Rega. In instances such as this, the VL provides a tool which

[6] The following versions of the drug ranking systems rule sets were used in determining the sequence used in the VPE: HIVdb 5.1.2, ANRS 17 and Rega 8.0.1.

[7] EuResist: www.euresist.org

Table 6.1: Percentage of patient sequences within the HIVdb database containing the mutations identified as potentially resistance causing in the VPE for both PI treatment naive and experienced individuals. The figure in brackets is the percentage for those undergoing LPV monotherapy.

Mutation	**Treatment Naive**	**PI Treated**
L10I	7.3	41.0 (12.5)
A71I	0.0	3.3 (0.0)
A71V	4.6	38.0 (7.0)
L90M	0.0	43.0 (5.2)

allows a clinician or researcher to investigate the cause of the discordance by inspecting the rules used to determine the ranking by each system. The only mutations within this set which influenced the ranking were L10I, A71IV and L90M.

In order to investigate whether these mutations caused resistance and if so how they interacted to do so, all possible combinations were simulated. In addition the full patient sequence was simulated (using the altered residues at positions where polymorphisms were detected) with both valine and isoleucene present at position 71. For simplicity the two full patient sequences shall be referred to as VPE-A71V and VPE-A71I, depending on the amino acid present at position 71.

Table 6.1 shows that the mutations identified for study in the VPE occur with differing frequencies in both naive and treated patients. L10I and A71V are infrequent polymorphisms in naive patients, which are strongly selected for under protease inhibitor (PI) treatement. Neither A71I or L90M are observed in naive patients but are selected for by treatment to very different degrees, with A71I only occurring in 3.3% of treated patients and L90M in 43.0%. These figures are for treatment with either one or more PI. The HIVdb only contains 57 sequences from patients undergoing lopinavir monotherapy. The HIVdb only contains 57 sequences from patients undergoing lopinavir monotherapy, analysis of this, limited data set, suggests that selection for the resistance linked mutations in the VPE is less strong for lopinavir than when other PIs are used. Additionally, a study of 1313 HIV infected individuals in Spain investigated the effects of therapies containing LPV [387]. This study found similar occurrence frequencies to HIVdb amongst naive patients (L10I, A71I, A71V and L90M were found in 7.7%, 0.0%, 5.4% and 2.9% of individuals respectively) but found higher levels of selection under LPV treatment for all mutations under consideration (L10I, A71I, A71V and L90M were found in 27.0%, 1.1%, 20.1% and 24.3% of patients respectively) although this was less than that found in the HIVdb for general PI treatment.

In order to assess the impact of these mutations on the free energy of binding (and hence upon resistance) we require comparison systems, that determine the binding affinity for

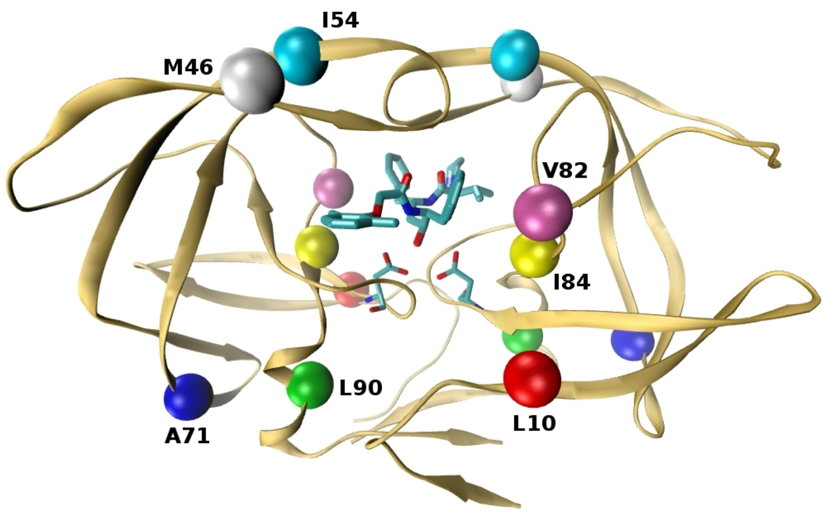


Figure 6.1: HIV-1 protease (backbone shown in ribbon representation) bound to the inhibitor lopinavir (shown in chemical structure representation along with the catalytic dyad at position 25 of each protease monomer). The locations of the mutations found in the multi drug resistant (MDR) mutants (described in Table 5.1) used for the benchmark simulations and residue A71 are highlighted and labelled. Protease is a homodimer and the location of each mutation is given the same color on both monomers.

susceptible and resistant sequences. The binding affinity values presented in Chapter 5 for a series of mutants (referred to as the MDR Test Set) provide and ideal set of benchmark values, as they contain both the susceptible HXB2 wildtype sequence, labelled WT, and several mutants which are known clinically, as well as experimentally, to cause resistance. The sequence containing the mutations L10I, M46I, I54V, V82A, I84V and L90M (and labelled HM) was chosen to provide the resistant benchmark for the VPE as the most conclusively resistance sequence in the MDR Test Set. The double mutant containing V82A and I84V (known as AS) was also used as it provides a useful comparison for systems of intermediate, yet clinically relevant resistance. The location within the protease structure of all mutations with known clinical relevance in the sequences to be studied are shown in Figure 6.1. All three mutations which impacted the resistance scoring of the CDSS are located more than 10 Å from the active site.

6.2 Methods and Analysis

The simulations and free energy calculations were performed using the automated BAC tool[355]. The protocol used for structure preparation, simulation and analysis was the same as that described in Chapter 5. The following is a brief overview of this process.

6.2.1 Simulation and Free Energy Calculation Protocol

The required protein sequences were created from the 1MUI crystal structure using the *in silico* mutational algorithms of the program VMD[315]. Each protein system was solvated in a cuboid box of TIP3P water molecules [359], with a minimum 14Å buffering distance in all three orthogonal dimensions. The system was then minimised with all protein and ligand heavy atoms constrained to their positions in the initial structure. Each system heated from 50 to 300 K over 50 ps after which the system was maintained at a temperature of 300 K. Once the system had been heated to the correct temperature in all subsequent simulation steps the pressure is maintained at 1 bar. This results in the system sampling an isothermal isobaric (NPT) ensemble. Simulation proceeded for 200 ps before a mutation relaxation protocol was enacted in which each mutated residue and residues within 5 Å were released in turn from the constraints for 50 ps. After the 50 ps relaxation period the restraints were reapplied to each region. The final equilibration stage was the gradual reduction of the restraining force on the complex from 4 to 0 kcal mol^{-1}Å^{-2} during a 350 ps period. Following this the systems were allowed to evolve freely. The entire equilibration stage was designed to take 2 ns for all systems meaning that this final stage varied in length according to the number of mutations which required relaxation in the previous stages. After the equilibration is complete structures are output for analysis every 10 ps. Every output snapshot was post processed using MMPBSA, meaning that a hundred sets of coordinates were analysed for each nanosecond of simulation. The more computationally expensive normal mode analysis was performed on every 20 snapshots, producing five entropy estimates per nanosecond of simulation.

6.2.2 Principal Component Analysis

Principal component analysis (PCA) is a dimensional reduction technique that allows the isolation of the most significant conformational differences between a set of structures. Here the structures are provided by snapshots from the molecular dynamics trajectory. The correlation matrix is calculated from an aligned molecular dynamics trajectory and then diagonalised. This provides an orthogonal set of eigenvectors representing linearly independent modes of conformational change called principal components. The eigenvalues associated with each principal component are a measure of the variance in the original dataset described by that component. The principal component analysis presented here was performed on the backbone coordinates of a concatenated trajectory of all mutant sequences under investigation along with the WT and HM benchmark sequences. All structures used for MMPBSA calculations were included in the PCA.

The trajectory concatenation and sidechain atom stripping were performed using VMD [315] and the PCA conducted using the ptraj module of AMBER9[70]

6.3 Results

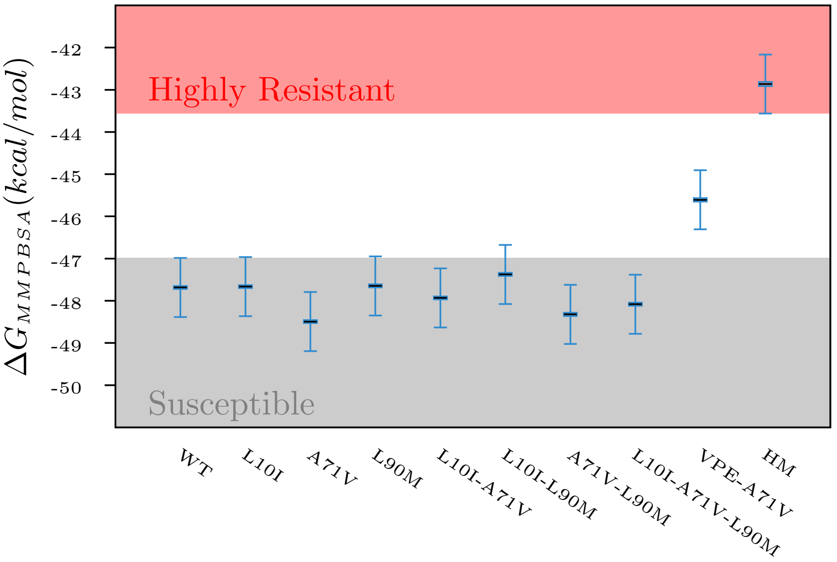
In order to simplify the presentation of the binding affinity results, the full set of sequences has been split into two sets; those containing alanine or valine as residue 71 and those containing alanine or isoleucene at this position. As part of the discussion of these results and as a natural extension of the consideration of the entropic component of the binding free energy, protein flexibility will be analysed. Following this a comparison of the structural differences observed across the entire dataset will be described.

6.3.1 Binding Affinity and Protein Flexibility

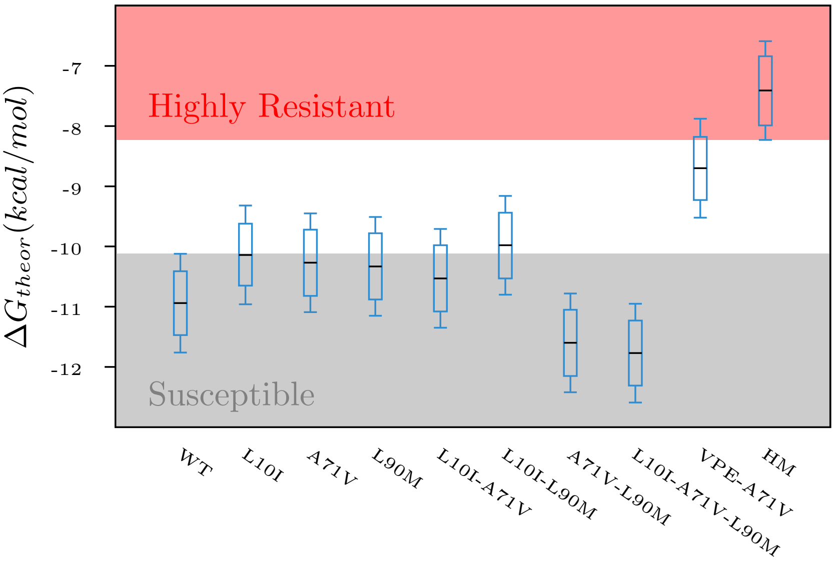
6.3.1.1 A71V

Figure 6.2 shows a comparison of the mutants within the VPE (considering only those with alanine or valine at position 71) with the WT and HM benchmark systems using both the absolute binding affinity, ΔG_{theor}, and the MMPBSA calculated value, $\Delta G_{_{MMPBSA}}$. Whilst, the A71V mutation alone, or in any combination with the L10I and L90M mutations, inserted into the HXB2 wild type sequence, induces no significant reduction in affinity for lopinavir a significant level of resistance is exhibited by the full patient sequence. A particularly striking difference is exhibited between the triple mutant containing all mutations identified by the cDRS as potentially causing resistance, L10I-A71V-L90M, and the same mutations incorporated in the full patient sequence (VPE-A71V). The full patient sequence is 3.06 kcal mol^{-1} more resistance using ΔG_{theor} and 2.47 kcal mol^{-1} according to $\Delta G_{_{MMPBSA}}$.

The VPE-A71V binding affinity, using either metric, is considerably less resistant than the HM sequence. The change from WT is comparable with the AS mutant values obtained in Chapter 5. Table 6.2 shows the detailed comparison and thermodynamic decomposition of all sequences under investigation from the VPE , demonstrating that in fact the $\Delta\Delta G_{_{MMPBSA}}$ value for the VPE-A71V sequence is marginally less attractive (hence more resistant) than that for AS mutant, while the $\Delta\Delta G_{theor}$ value is 0.48 kcal mol^{-1} more attractive than this known resistance causing sequence. The decomposition of the enthalpically dominated $\Delta G_{_{MMPBSA}}$ component of the binding free energy show in Figure 6.2b indicates that the origin of the difference from wild type is the polar solvation term, ΔG^{sol}_{pol}, which is 3 kcal mol^{-1} more repulsive than in the WT case, with both the



(a)



(b)

Figure 6.2: A comparison of the binding free energies computed for all mutants containing either alanine or valine at position 71 studied in the virtual patient experiment bound to LPV. The values for the known susceptible WT sequence and known resistant HM sequence are also shown for comparison. a) shows the binding affinity value calculated using MMPBSA, ΔG_{MMPBSA}, and b) the absolute binding affinity, ΔG_{theor}. The black lines show the mean, the candle stick the standard error and the whiskers the error based on the WT and HM reproducibility for each system. The grey and red shaded region show the range of values deemed susceptible and resistant defined using the WT and HM benchmark values.

Table 6.2: Comparison of the free energy of binding to LPV for the benchmark sequences (susceptible WT and resistant AS and HM) and the subset of VPE sequences containing either alanine or valine at position 71. a) Decomposition of the absolute binding free energy, ΔG_{theor}, into the enthalpic and entropic contributions ($\Delta G_{_{MMPBSA}}$ and $-T\Delta S_{_{NM}}$respectively) and the relative free energy differences ($\Delta\Delta G$) between the mutant and wild type values. b) Decomposed contributions to the enthalpically dominated contribution to the free energy of binding, $\Delta G_{_{MMPBSA}}$, computed using MMPBSA.

(a)

Sequence	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{-T\Delta S_{NM}}$	$\mathbf{\Delta G_{theor}}$	$\mathbf{\Delta\Delta G_{MMPBSA}}$	$\mathbf{\Delta\Delta G_{theor}}$
WT	-47.68 (0.03)	36.74 (0.50)	-10.94 (0.53)	-	-
L10I	-47.66 (0.03)	37.53 (0.48)	-10.13 (0.51)	0.02 (0.06)	0.80 (1.05)
A71V	-48.49 (0.03)	38.23 (0.52)	-10.26 (0.55)	-0.81 (0.06)	0.67 (1.08)
L90M	-47.65 (0.04)	37.32 (0.51)	-10.33 (0.55)	0.03 (0.07)	0.61 (1.08)
L10I-A71V	-47.93 (0.03)	37.40 (0.52)	-10.53 (0.55)	-0.25 (0.06)	0.41 (1.08)
L10I-L90M	-47.38 (0.04)	37.39 (0.51)	-9.99 (0.55)	0.30 (0.07)	0.96 (1.08)
A71V-L90M	-48.32 (0.04)	36.72 (0.51)	-11.60 (0.55)	-0.64 (0.07)	-0.66 (1.08)
L10I-A71V-L90M	-48.08 (0.03)	36.31 (0.50)	-11.77 (0.53)	-0.40 (0.06)	-0.83 (1.06)
VPE-A71V	-45.61 (0.04)	36.90 (0.48)	-8.71 (0.52)	2.07 (0.07)	2.24 (1.06)
AS	-45.75 (0.04)	37.53 (0.54)	-8.22 (0.58)	1.93 (0.07)	2.72 (1.10)
HM	-42.86 (0.05)	35.45 (0.52)	-7.41 (0.57)	4.82 (0.08)	3.53 (1.11)

Mean energies are in kcal/mol. Standard errors are shown in parentheses.

(b)

Sequence	$\mathbf{\Delta G^{MM}_{vdw}}$	$\mathbf{\Delta G^{MM}_{ele}}$	$\mathbf{\Delta G^{sol}_{pol}}$	$\mathbf{\Delta G^{sol}_{nonpol}}$	$\mathbf{\Delta G^{tot}_{ele}}$	$\mathbf{\Delta G_{MMPBSA}}$
WT	-72.82 (0.03)	-51.29 (0.05)	84.56 (0.05)	-8.13 (0.00)	33.27 (0.04)	-47.68 (0.03)
L10I	-72.91 (0.03)	-49.67 (0.05)	83.05 (0.04)	-8.14 (0.00)	33.38 (0.04)	-47.66 (0.03)
A71V	-72.97 (0.03)	-51.04 (0.05)	83.65 (0.05)	-8.13 (0.00)	32.61 (0.04)	-48.49 (0.03)
L90M	-72.40 (0.03)	-51.68 (0.05)	84.57 (0.05)	-8.14 (0.00)	32.89 (0.04)	-47.65 (0.04)
L10I-A71V	-72.65 (0.03)	-50.20 (0.05)	83.04 (0.04)	-8.12 (0.00)	32.84 (0.04)	-47.93 (0.03)
L10I-L90M	-72.71 (0.03)	-52.29 (0.05)	85.76 (0.05)	-8.13 (0.00)	33.47 (0.04)	-47.37(0.03)
A71V-L90M	-73.41 (0.03)	-53.14 (0.06)	86.36 (0.05)	-8.13 (0.00)	33.22 (0.05)	-48.32 (0.04)
L10I-A71V-L90M	-73.31 (0.03)	-52.45 (0.06)	85.81 (0.05)	-8.14 (0.00)	33.36 (0.05)	-48.08 (0.04)
VPE-A71V	-73.32 (0.03)	-51.65 (0.06)	87.56 (0.06)	-8.20 (0.00)	35.91 (0.05)	-45.61 (0.04)
AS	-71.02 (0.03)	-51.97 (0.07)	85.38 (0.05)	-8.13 (0.00)	33.41 (0.04)	-45.75 (0.04)
HM	-70.43 (0.04)	-50.51 (0.08)	86.17 (0.06)	-8.10 (0.00)	35.66 (0.05)	-42.86 (0.05)

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

Table 6.3: The frequency of occupation of putative hydrogen bonds between the inhibitor lopinavir (LPV) and the HIV-1 protease sequences in the VPE calculated over the production MD simulations (consisting of 50 replicas producing 4ns trajectories) for each system. A hydrogen bond is identified as being formed when a donor-acceptor pair is within 3.5 Å of one another and the donor-hydrogen-acceptor angle is less that 120 degrees. The oxygen and nitrogen atoms of LPV are numbered from the left of the schematic shown in Figure 5.2a. Only bonds which were occupied for more than 10% of at least one studied sequence are listed.

	Donor	**LPV O3**	**LPV N1**	**LPV N3**	**LPV N4**	**LPV O3**	**LPV N1**	**29 N**	**30 N**
Sequence	**Acceptor**	**125 OD2**	**29 OD1**	**27 O**	**125 OD2**	**125 OD1**	**29 OD2**	**LPV O1**	**LPV O1**
WT		0.945	0.772	0.496	0.205	0.179	0.134	0.999	0.051
L10I		0.958	0.715	0.548	0.169	0.166	0.162	0.998	0.038
A71V		0.991	0.811	0.534	0.175	0.137	0.108	0.999	0.044
L90M		0.944	0.773	0.600	0.161	0.204	0.097	0.989	0.052
L10I-A71V		0.962	0.705	0.547	0.179	0.158	0.142	0.998	0.039
L10I-L90M		0.948	0.655	0.479	0.320	0.235	0.179	0.992	0.090
A71V-L90M		0.960	0.703	0.431	0.346	0.210	0.185	0.999	0.060
L10I-A71V-L90M		0.987	0.709	0.407	0.347	0.193	0.174	0.991	0.057
VPE-A71V		0.977	0.543	0.331	0.382	0.161	0.127	0.996	0.209
AS		0.931	0.418	0.234	0.535	0.212	0.235	0.986	0.160
HM		0.819	0.258	0.198	0.408	0.247	0.199	0.960	0.121

van der Waals and electrostatic contributions (ΔG_{vdw}^{MM} and ΔG_{ele}^{MM} respectively) being marginally more attractive. This is a different pattern to that seen in the AS and HM sequences where small changes in polar solvation energy are augmented by considerable reduction in the attractiveness of the van der Waals interactions (directly caused by the mutation of the active site residues 82 and 84). Further differentiation of the resistant VPE-A71V sequence from the resistant mutants in the MDR test set comes from the observation that it (along with all other sequences containing A71V) maintains hydrogen bonds between the hydroxyethylene moiety (labelled O3, nitrogen and oxygen atoms are separately numbered from the left of the schematic shown in Figure 5.2a) and the catalytic ASP 125 more frequently than the WT (see Table 6.3 for a list of hydrogen bond frequencies between LPV and the protease systems under investigation). The VPE-A71V sequence does, however, exhibit reduced frequency of bonding with residues 27 and 29 in a similar fashion to the known resistant mutants from the MDR test set. In common with the AS and HM systems, but unlike any other we have investigated, the VPE-A71V sequence also has substantial hydrogen bonds between the oxygen O1 and the backbone nitrogen of residue 30 in the P2$'$ subsite.

In the study presented in Chapter 5, the loss of bonding with residues 27 and other changes in the hydrogen bonding network (involving residues 29 and 30), which is similar to that observed in the VPE-A71V system, was correlated with both resistance and the entry of water molecules into the catalytic site. Table 6.4 shows the population density of water molecules within 3 Å of the catalytic dyad (ρ) for the sequences under consideration here. Only the virtual patient sequence, VPE-A71V, shows substantial water occupation in this area, with a frequency 0.324 compared to 0.057 for WT. This is between the levels observed for the AS and HM known resistant mutants. The A71V single mutant shows the opposite change in active site accessibility, with water ingress exhibited by less than one percent of snapshots (a fifth of that seen for the WT sequence).

Figure 6.2a indicates that, according to the ΔG_{MMPBSA} metric, three sequences (A71V, A71V-L90M and L10I-A71V-L90M) may bind more strongly than the wild type. The largest change in ΔG_{MMPBSA} (of 0.81 kcal mol^{-1}) appears in the A71V single mutant but this increase in attraction is counteracted by a large change in entropic contribution ($-T\Delta S_{NM}$) which results in an absolute binding affinity, ΔG_{theor}, value which is less negative than that computed for the WT (see Figure 6.2b and Table 6.2a). The increase in binding affinity for the A71V-L90M and L10I-A71V-L90M persists even when entropy is accounted for in ΔG_{theor}, indicating that at least these two mutants containing A71V may be hyper-susceptible to LPV. The phenomenon of hyper-susceptibility to LPV has been observed in a range of protease sequences (particularly those of subtype C viruses) experimentally although the clinical impact remains uncertain[296, 388, 389].

Table 6.4: Population density, ρ, of water within 3 Å of the catalytic dyad exhibited by each of the sequences studied as part of the Virtual Patient Experiment for which position 71 is occupied by either alanine or valine.

Sequence	ρ
WT	0.057
L10I	0.042
A71V	0.009
L90M	0.035
L10I-A71V	0.053
L10I-L90M	0.039
A71V-L90M	0.084
L10I-A71V-L90M	0.031
VPE-A71V	0.324
AS	0.167
HM	0.484

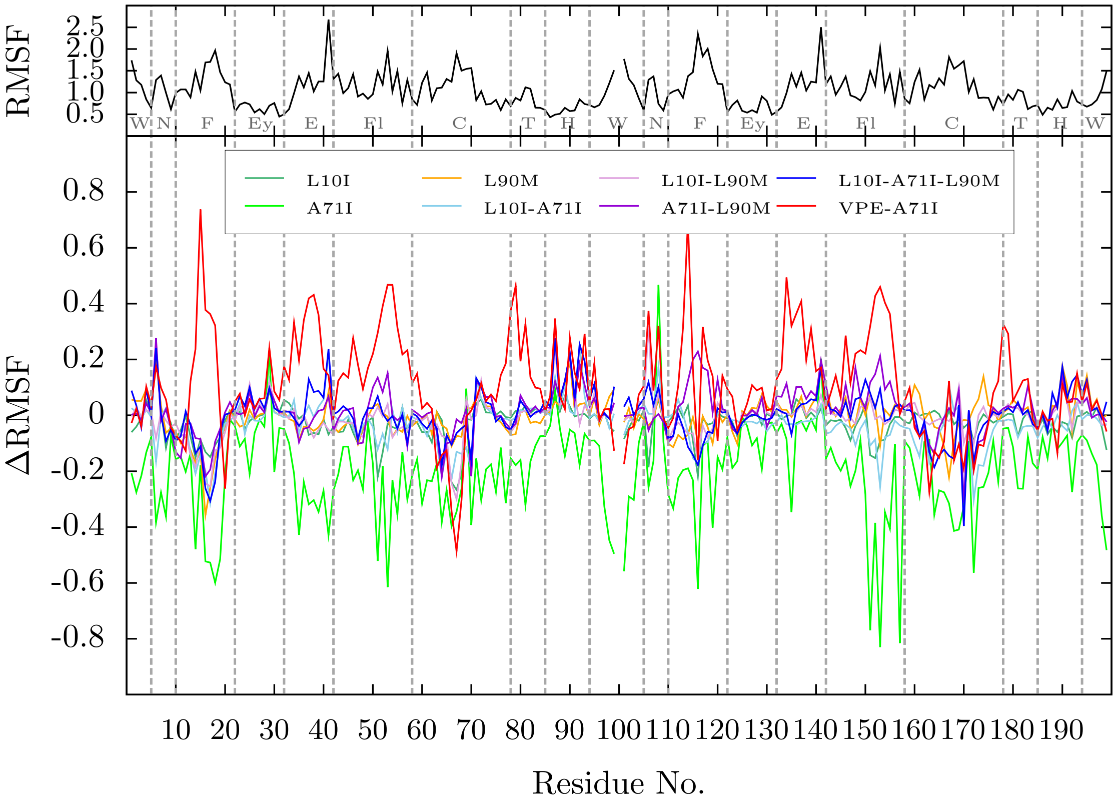
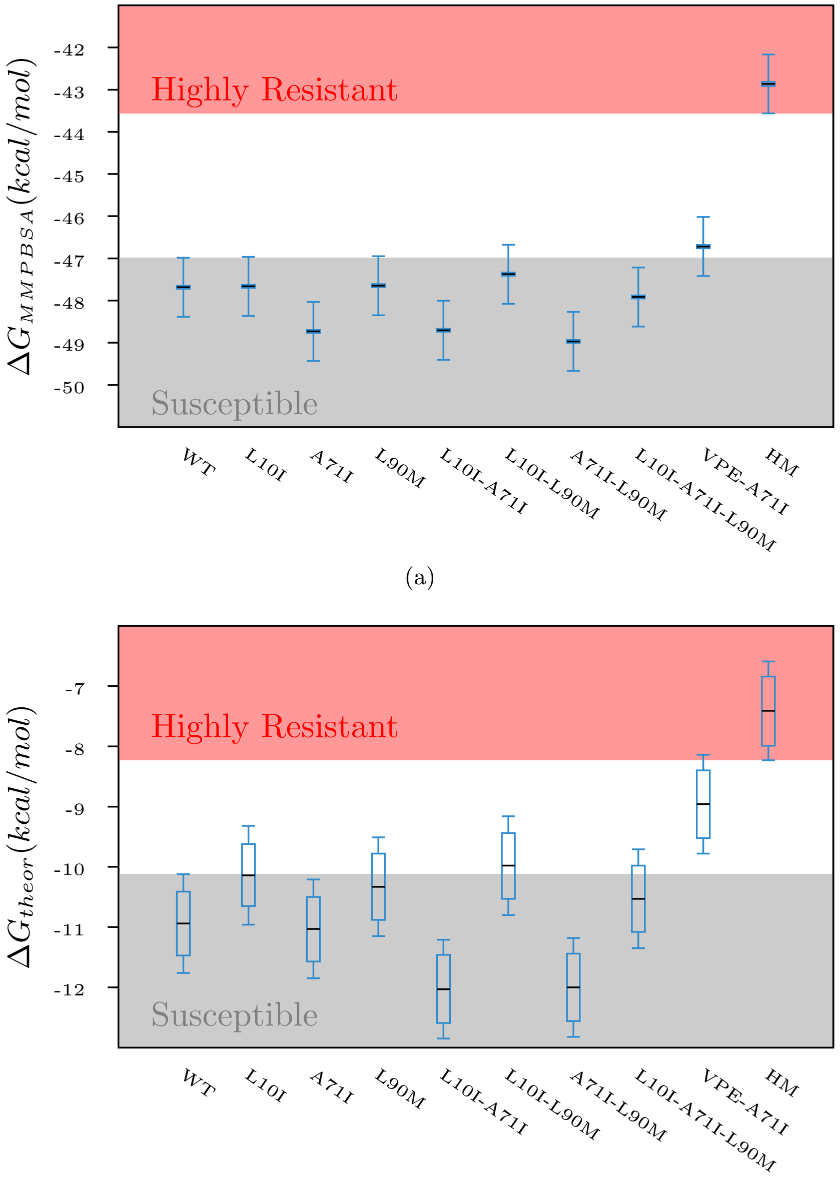


Figure 6.3: The root mean square fluctuation, RMSF, relative to the average structure is shown for each residue, in number order, of the WT sequence. Beneath this is the per residue difference in RMSF, ΔRMSF, exhibited by each of the sequences studied as part of the Virtual Patient Experiment for which position 71 is occupied by either alanine or valine. A positive ΔRMSF value indicates that the fluctautions observed at that position in the mutant system are greater than those at the corresponding location in WT.

An assessment of the flexibility of a protein bound to LPV during a simulation can be made using the root mean square fluctuation (RMSF) of the structures explored during each ensemble relative to the average structure for that sequence. Figure 6.3 shows the RMSF of each residue within the WT structure and the differences (ΔRMSF) for all positions in each of the sequences in the A71V related subset of the VPE relative to this. In the WT there are two very stable regions in both monomers, one around the catalytic ASP 25 (coinciding with the 'eyes' between residues 22 and 32) and the other the helix formed by residues 86 and 94. The most flexible regions in that system are the 'fulcrum' (between residues 10 and 22), residue 41 and parts of the 'cheek sheet' preceding residue 70 (this applies to both monomers). With the exception of the A71V and VPE-A71V sequences, all of the systems exhibit broadly similar flexibility to WT, some additional flexibility is shown in the helix and whiskers particularly around residues 87, 90 and 93. In line with the increase entropic barrier to LPV binding, the A71V single mutant exhibits reduced flexibility across the entire structure with an average ΔRMSF of -0.2 Å. Notable exceptions to this pattern are residues 29 and 108. The loss of mobility is particularly large in the area of the fulcrum of both chains and the flaps of the second chain where losses of up to 0.8 Å are observed. A noticeable reduction is also seen in the 'whiskers' (residues 95 to 99) which are involved in the dimer β sheet and the 'elbow' (between residues 32 and 42). Many of the regions seen to be stabalised in the A71V mutant are more flexible than WT in the VPE-A71V system. It is worth noting that for both of these systems changes observed in one monomer are almost invariably present also in the other. The region with the most prominently increased fluctuations is the fulcrum where ΔRMSF values of 0.75 Å are observed but the flaps, elbow and wall turn (containing residues 79 to 86) also gain flexibility. The wall turn contains key hydrophobic residues involved in interactions with the drug such as V82 and I84. The added flexibility in this region may allow the protein to gain more favourable interactions with LPV partially explaining the more attractive ΔG^{MM}_{vdw} component of the binding affinity. The region of the cheek sheet around residue 69 in contrast to much of the rest of the structure is seen to strongly stabalised with a ΔRMSF of under 0.5 Å. The observed overall gain in flexibility of the VPE-A71V system is not reflected in a higher entropic barrier to binding, suggesting that this flexibility must also be present in the free enzyme as well as the drug bound complex.

6.3.1.2 A71I

The binding affinity results for the VPE mutation set including A71I, shown in Figure 6.4, exhibit differences in response between the triple mutant and full patient



(a)

(b)

Figure 6.4: A comparison of the binding free energies computed for all mutants containing either alanine or isoleucine at position 71 studied in the virtual patient experiment bound to LPV. The values for the known susceptible WT sequence and known resistant HM sequence are also shown for comparison. a) shows the binding affinity value calculated using MMPBSA, ΔG_{MMPBSA}, and b) the absolute binding affinity, ΔG_{theor}. The black lines show the mean, the candle stick the standard error and the whiskers the error based on the WT and HM reproducibility for each system. The grey and red shaded region show the range of values deemed susceptible and resistant defined using the WT and HM benchmark values.

sequence as observed for A71V, although the change is of considerably lower magnitude. The difference in absolute binding affinity compared to WT, $\Delta\Delta G_{theor}$, for the VPE-A71I sequence is 1.98 kcal mol^{-1}, comparable with that seen for VPE-A71V (where $\Delta\Delta G_{theor}$ was 2.07 kcal mol^{-1}) and indicative of at least intermediate resistance. Whereas the change in enthalpically dominated $\Delta\Delta G_{MMPBSA}$ is only 0.96 kcal mol^{-1}. The $\Delta\Delta G_{MMPBSA}$ value if accurate would represent a low level of resistance (approximately equivalent to a 5 fold change in K_d/IC$_{50}$) but is at the limit of the ability of our method to distinguish systems. The reduced difference between the patient sequence containing A71I and the L10I-A71I-L90M system compared to that seen when valine is present at position 71 is also in part due to the fact that the triple mutant binds with almost identical strength to the wildtype, using both the ΔG_{MMPBSA} and ΔG_{theor} measures (the L10I-A71V-L90M bind more tightly than WT). Similarly to the A71V data set none of the sequences investigated here display any resistance other than VPE-A71I. Again, as in the A71V data set, some sequences exhibit signs of hypersusceptibility.

Table 6.5b shows the thermodynamic decomposition of the ΔG_{MMPBSA} values. The resistance of VPE-A71I, like that of VPE-A71V, is primarily caused by changes in the electrostatic components of binding. However, where the majority of the change in the A71V containing variant was in ΔG^{pol}_{sol} in this case ΔG^{ele}_{MM} was more significant (with changes of 0.25 and 0.58 kcal mol^{-1} compared to WT respectively). This is reflected in slightly different changes in the active site hydrogen bonding networks between the two patient sequence derived systems. Unlike VPE-A71V the bond between catalytic ASP 125 and the lopinavir hydroxyethylene moiety is reduced in occupation compared to WT in VPE-A71I. As with all other resistant mutants under consideration there is a shift away from bonds with residues 29 and 27 and towards those with 30. The increase in bonding frequency with residue 30 is particularly pronounced in VPE-A71I with occupation levels over 60% (compared to less than 30% in VPE-A71V). As with the other resistant mutants investigated here the change in hydrogen bonding within the active site is accompanied by water entry into the catalytic cavity. Water appears within 3 Å of the catalytic dyad in 25% of snapshots (see Table 6.7). This is less common than in the HM or VPE-A71V systems but more so than the AS mutant. The increase in water population density in this region is not shared by any of the other systems containing A71I.

The A71I, L10I-A71I and A71I-L90M systems have $\Delta\Delta G_{MMPBSA}$ values considerably more negative than the WT (-1.05, -1.02 and -1.29 kcal mol^{-1} respectively). While these values are close to the resolution limit of our method they are greater than the reproducibility variability seen in the WT and HM systems. This increase in the strength of binding is conserved for the L10I-A71I and A71I-L90M systems when the entropic contribution is included in the results (they have $\Delta\Delta G_{theor}$ values of -1.09 and -1.06 kcal

Table 6.5: Comparison of the free energy of binding to LPV for the benchmark sequences (susceptible WT and resistant AS and HM) and the subset of VPE sequences containing either alanine or isoleucine at position 71. a) Decomposition of the absolute binding free energy, ΔG_{theor}, into the enthalpic and entropic contributions (ΔG_{MMPBSA} and $-T\Delta S_{NM}$respectively) and the relative free energy differences ($\Delta\Delta G$) between the mutant and wild type values. b) Decomposed contributions to the enthalpically dominated contribution to the free energy of binding, ΔG_{MMPBSA}, computed using MMPBSA.Decomposed contributions to the free energy of binding for wildtype and all MDR proteases with lopinavir as well as wildtype with saquinavir using 50 × 4 ns ensemble-trajectory runs.

(a)

Sequence	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{-T\Delta S_{NM}}$	$\mathbf{\Delta G_{theor}}$	$\mathbf{\Delta\Delta G_{MMPBSA}}$	$\mathbf{\Delta\Delta G_{theor}}$
WT	-47.68 (0.03)	36.74 (0.50)	-10.94 (0.53)	-	-
L10I	-47.66 (0.03)	37.53 (0.48)	-10.13 (0.51)	0.02 (0.06)	0.80 (1.05)
A71I	-48.73 (0.03)	37.70 (0.50)	-11.03 (0.53)	-1.05 (0.06)	-0.09 (1.07)
L90M	-47.65 (0.04)	37.32 (0.51)	-10.33 (0.55)	0.03 (0.07)	0.61 (1.08)
L10I-A71I	-48.70 (0.04)	36.68 (0.53)	-12.02 (0.57)	-1.02 (0.07)	-1.09 (1.10)
L10I-L90M	-47.38 (0.04)	37.39 (0.51)	-9.99 (0.55)	0.30 (0.07)	0.96 (1.08)
A71I-L90M	-48.97 (0.03)	36.97 (0.52)	-12.00 (0.55)	-1.29 (0.06)	-1.06 (1.09)
L10I-A71I-L90M	-47.92 (0.04)	37.39 (0.51)	-10.53 (0.55)	-0.24 (0.07)	0.41 (1.08)
VPE-A71I	-46.72 (0.04)	37.76 (0.52)	-8.96 (0.56)	0.96 (0.07)	1.98 (1.09)
AS	-45.75 (0.04)	37.53 (0.54)	-8.22 (0.58)	1.93 (0.07)	2.72 (1.10)
HM	-42.86 (0.05)	35.45 (0.52)	-7.41 (0.57)	4.82 (0.08)	3.53 (1.11)

Mean energies are in kcal/mol. Standard errors are shown in parentheses.

(b)

Sequence	$\mathbf{\Delta G^{MM}_{vdw}}$	$\mathbf{\Delta G^{MM}_{ele}}$	$\mathbf{\Delta G^{sol}_{pol}}$	$\mathbf{\Delta G^{sol}_{nonpol}}$	$\mathbf{\Delta G^{tot}_{ele}}$	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{-T\Delta S_{NM}}$
WT	-72.82 (0.03)	-51.29 (0.05)	84.56 (0.05)	-8.13 (0.00)	33.27 (0.04)	-47.68 (0.03)	36.74 (0.49)
L10I	-72.91 (0.03)	-49.67 (0.05)	83.05 (0.04)	-8.14 (0.00)	33.38 (0.04)	-47.66 (0.03)	37.53 (0.48)
A71I	-73.33 (0.03)	-50.07 (0.05)	82.79 (0.04)	-8.13 (0.00)	32.72 (0.04)	-48.73 (0.03)	37.70 (0.50)
L90M	-72.40 (0.03)	-51.68 (0.05)	84.57 (0.05)	-8.14 (0.00)	32.89 (0.04)	-47.65 (0.04)	37.32 (0.51)
L10I-A71I	-73.19 (0.03)	-50.12 (0.05)	82.72 (0.04)	-8.11 (0.00)	32.60 (0.03)	-48.70 (0.04)	36.68 (0.53)
L10I-L90M	-72.71 (0.03)	-52.29 (0.05)	85.76 (0.05)	-8.13 (0.00)	33.47 (0.04)	-47.37(0.03)	37.39 (0.51)
A71I-L90M	-73.14 (0.03)	-51.65 (0.05)	83.95 (0.05)	-8.13 (0.00)	32.30 (0.04)	-48.97 (0.03)	36.97 (0.52)
L10I-A71I-L90M	-72.55 (0.03)	-50.70 (0.05)	83.44 (0.04)	-8.10 (0.00)	32.74 (0.04)	-47.92 (0.04)	37.39 (0.51)
VPE-A71I	-72.63 (0.03)	-50.71 (0.06)	84.81 (0.05)	-8.19 (0.00)	34.10 (0.05)	-46.72 (0.04)	37.76 (0.52)
AS	-71.02 (0.03)	-51.97 (0.07)	85.38 (0.05)	-8.13 (0.00)	33.41 (0.04)	-45.75 (0.04)	35.57 (0.54)
HM	-70.43 (0.04)	-50.51 (0.08)	86.17 (0.06)	-8.10 (0.00)	35.66 (0.05)	-42.86 (0.05)	35.45 (0.52)

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

Table 6.6: The frequency of occupation of putative hydrogen bonds between the inhibitor lopinavir (LPV) and the HIV-1 protease sequences in the VPE calculated over the production simulations (consisting of 50 replicas producing 4ns trajectories). A hydrogen bond is identified as being formed when a donor-acceptor pair is within 3.5 Å of one another and the donor-hydrogen-acceptor angle is less that 120 degrees. The oxygen and nitrogen atoms of LPV are numbered from the left of the schematic shown in Figure 5.2a. Only bonds which were occupied for more than 10% of at least one studied sequence are listed.

	Donor	**LPV O3**	**LPV N1**	**LPV N3**	**LPV N4**	**LPV O3**	**LPV N1**	**29 N**	**30 N**
Sequence	**Acceptor**	**125 OD2**	**29 OD1**	**27 O**	**125 OD2**	**125 OD1**	**29 OD2**	**LPV O1**	**LPV O1**
WT		0.945	0.772	0.496	0.205	0.179	0.134	0.999	0.051
L10I		0.958	0.715	0.548	0.169	0.166	0.162	0.998	0.038
A71I		0.986	0.832	0.536	0.155	0.134	0.065	0.999	0.025
L90M		0.944	0.773	0.600	0.161	0.204	0.097	0.989	0.052
L10I-A71I		0.987	0.671	0.530	0.192	0.136	0.155	0.999	0.039
L10I-L90M		0.948	0.655	0.479	0.320	0.235	0.179	0.992	0.090
A71I-L90M		0.946	0.798	0.547	0.203	0.198	0.106	0.998	0.012
L10I-A71I-L90M		0.968	0.704	0.510	0.243	0.172	0.114	0.992	0.019
VPE-A71I		0.939	0.676	0.414	0.291	0.191	0.122	0.986	0.697
AS		0.931	0.418	0.234	0.535	0.212	0.235	0.986	0.160
HM		0.819	0.258	0.198	0.408	0.247	0.199	0.960	0.121

Table 6.7: Population density, ρ, of water within 3 Å of the catalytic dyad exhibited by each of the sequences studied as part of the Virtual Patient Experiment for which position 71 is occupied by either alanine or valine.

Sequence	ρ
WT	0.057
L10I	0.042
A71I	0.025
L90M	0.035
L10I-A71I	0.037
L10I-L90M	0.039
A71I-L90M	0.058
L10I-A71I-L90M	0.064
VPE-A71I	0.253
AS	0.167
HM	0.484

mol^{-1} respectively) but not for the A71I single mutant (which is almost indistinguishable from WT with a $\Delta\Delta G_{theor}$ of -0.09 kcal mol^{-1}). Unlike the A71V system, which also exhibits an increased entropic barrier to binding, no change in the flexibility of the protein within the complex can be detected in the residue RMSF values calculated from the simulation (see Figure 6.5).

With the exception of the A71I single mutant the changes in RMSF relative to that seen in WT observed for all systems containing isoleucene at position 71 follow similar patterns to their valine containing counterparts. Only minor deviations from wildtype flexibility are observed for any of the systems except for VPE-A71I, where substantial increases in flexibility are seen in the fulcrum, flaps, elbow and wall turn. The gain in flexibility in these regions is shared with the VPE-A71V system, but the magnitude of the change is reduced.

6.3.2 Structural Changes

The changes in binding affinity and flexibility between systems described in this chapter are accompanied by changes in the conformations explored. PCA is a useful tool for gaining insight into which changes are most significant in a specific data set. The key questions we wish to answer are which structural changes are associated with resistance and which are associated with deformation caused by accommodation of mutated amino acids and the general flexibility of the protease. In order to focus on the resistance associated changes a combined trajectory of the backbone of the WT, L10I-A71I-L90M, L10I-A71V-L90M, VPE-A71I and VPE-A71V systems was created and analysed to produce principal components (PCs) which capture the most significant variance between

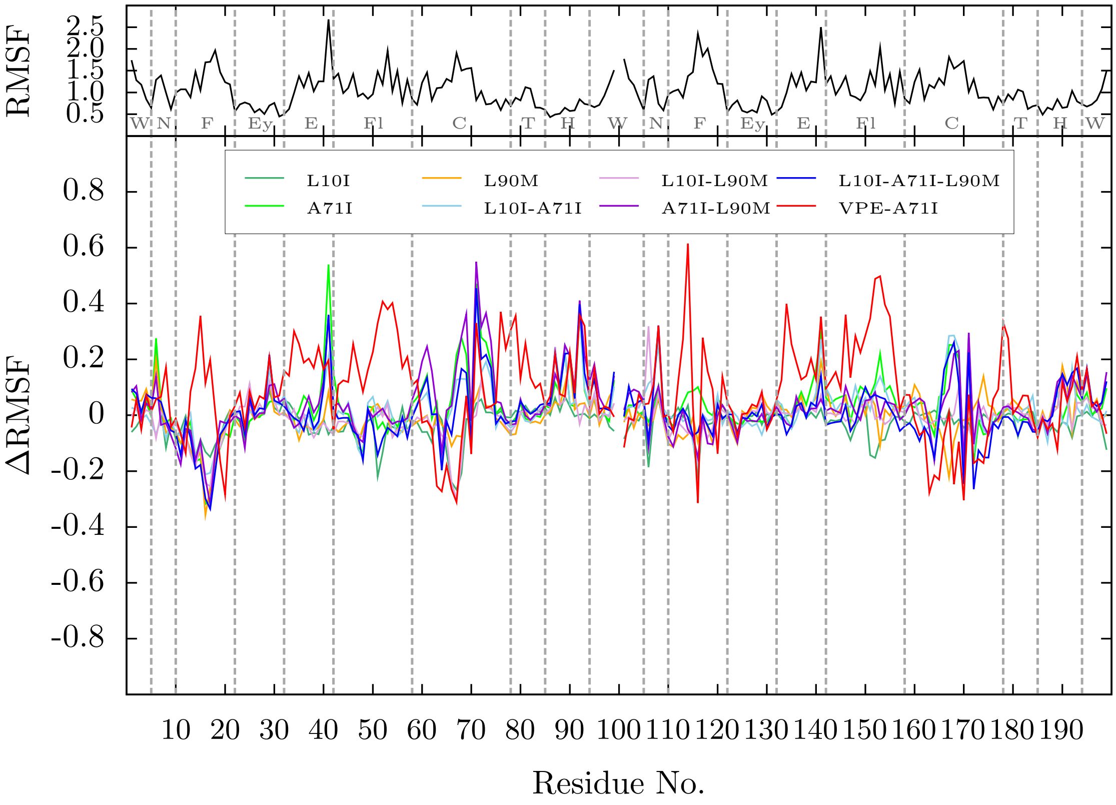


Figure 6.5: The root mean square fluctuation, RMSF, relative to the average structure is shown for each residue, in number order, of the WT sequence. Beneath this is the per residue difference in RMSF, ΔRMSF, exhibited by each of the sequences studied as part of the Virtual Patient Experiment for which position 71 is occupied by either alanine or isoleucene. A positive ΔRMSF value indicates that the fluctuations observed at that position in the mutant system are greater than those at the corresponding location in WT.

snapshots. The PCs can then be used to identify the most significant changes that distinguish the systems which are susceptible and resistant to LPV according to the binding affinity calculations presented above.

6.3.2.1 Principal Component Analysis

Figure 6.6 shows the level of variation captured by each of the first ten PCs. Only the first three describe greater than 5% of the observed differences in structure between snapshots. The first two components account for 46% of the total variation and the focus of this section will primarily be on what they can tell us about the differences between systems. In the context of protein conformational changes, the assumption of PCA that the observed data set (in this case the coordinates of each protein atom) is best expressed as a linear combination of certain basis vectors is significant. It is likely that the real conformational changes are not optimally described using such an assumption and hence PCA is better employed as an investigative tool to identify parts

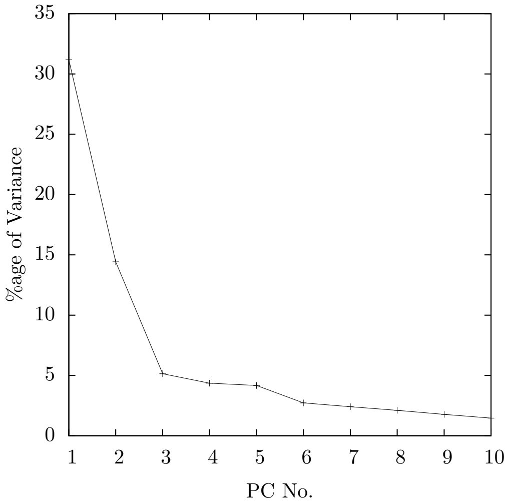
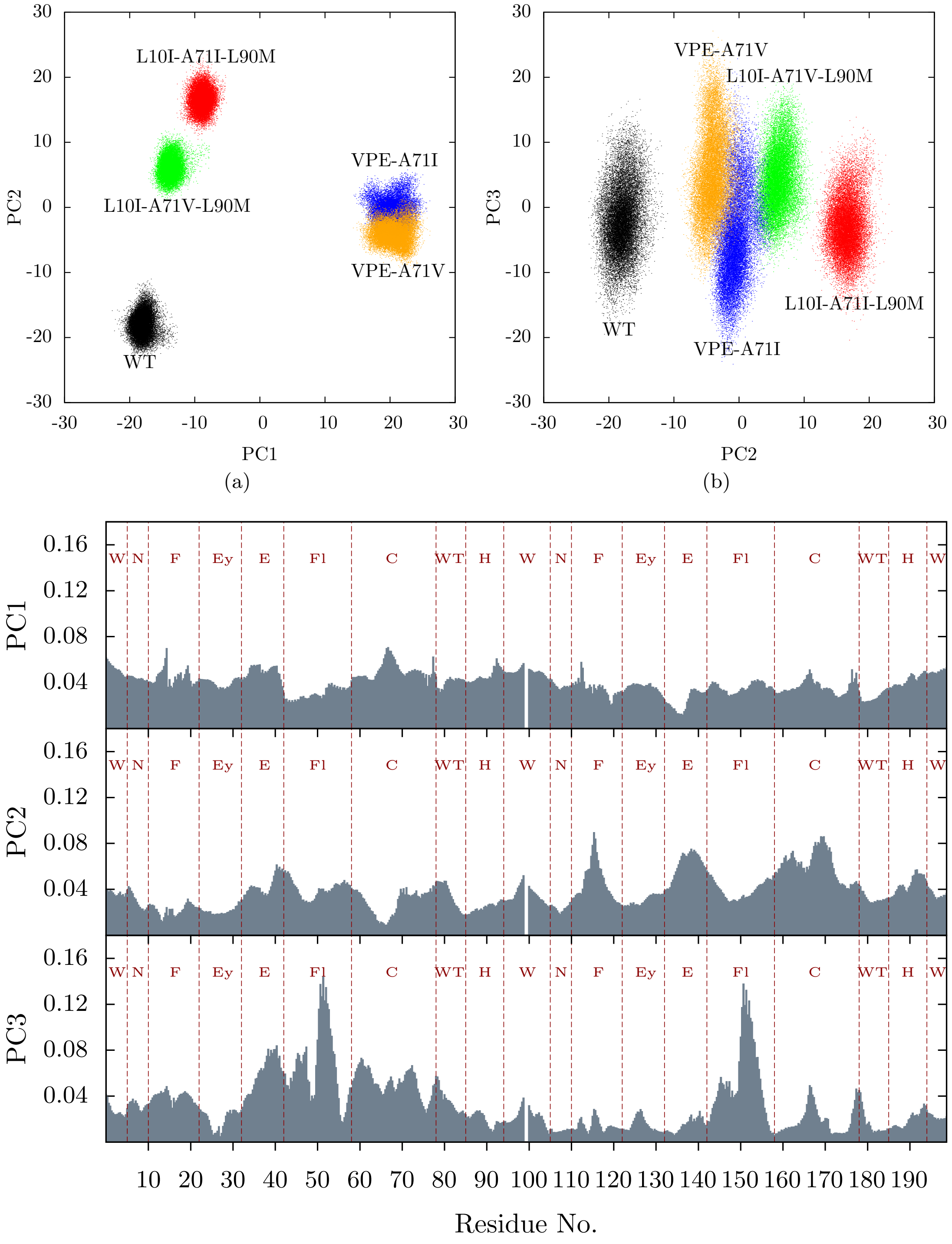


Figure 6.6: The percentage of the variation observed over the concatenated trajectory of WT, L10I-A71I-L90M, L10I-A71V-L90M, VPE-A71I and VPE-A71V which is captured by each principal component.

of the structure and general trends that can be explored further. Here the results of PCA are used to identify metrics which help describe conformational changes associated with resistance. Once such measurements have been identified they are applied to the larger dataset of all the sequences studied within the VPE to ensure that the differences can still be used to distinguish the resistant systems. If this is the case we can have confidence that we have identified structural changes that are linked to the changes in binding affinity.

The projections of the snapshots contained within the combined trajectory onto the first two principal components shown in Figure 6.7a indicates that each of the systems can be easily split into three groups using PC1 and PC2; the WT, two triple mutants and two patient sequences form three well separated groups. This means that conformational changes described by both PCs differentiate the susceptible and resistant systems and may give information about the structural origin of resistance in the VPE sequences. Figure 6.7b shows that PC3 has limited ability to distinguish the different systems and consequently represents fluctuations to the overall structure which are present in all of the systems used to generate the combined trajectory.

The changes described by PC1 across the combined trajectory are represented in Figure 6.8. The structure of the most negative projection from the trajectory is seen in blue, that of the most positive in grey (this convention will be used in all other structural figures from the PCA). The most pronounced change occurs as a global expansion of the structure along the axis of the active site cavity (corresponding to the fact that the PC1 shows variation for all residues in Figure 6.7c). This change is particularly visible in the change in separation of residue 35 (in the elbow) and 45 (in the flaps) in both

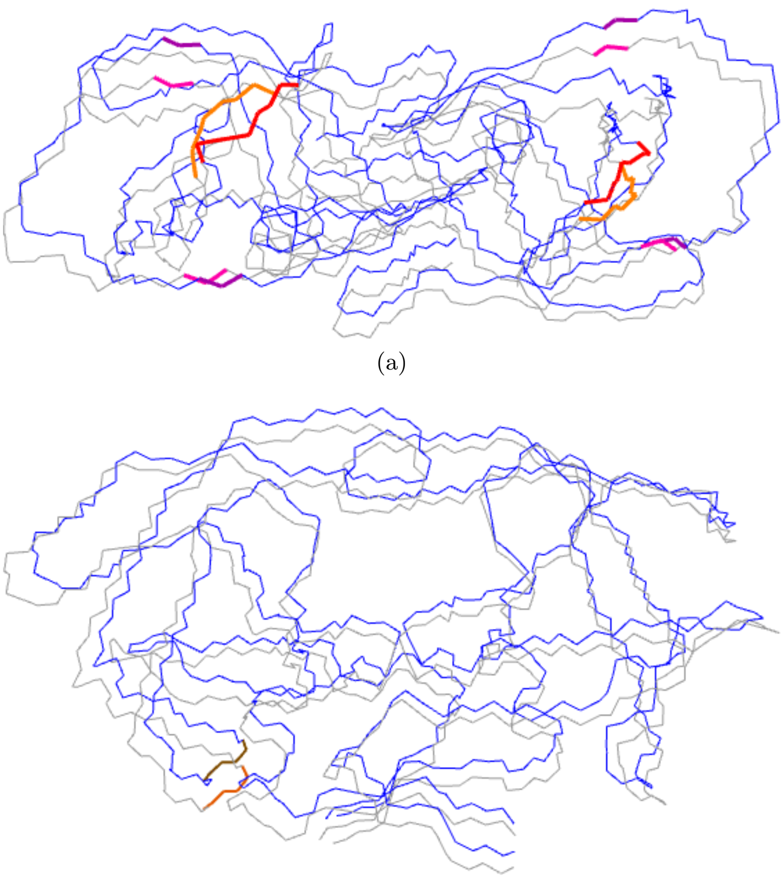


(a)

(b)

(c)

Figure 6.7: The description of the concatenated trajectory (consisting of tall frames produced by the WT, L10I-A71I-L90M, L10I-A71V-L90M, VPE-A71I and VPE-A71V production simulations) produced by PCA. (a) & (b) show the projections for each snapshot of the trajectory along the PCs 1 and 2, and 2 and 3 respectively. All three systems can easily be separated using PCs 1 and 2 but share a similar range of values along PC3. (c) shows the magnitude of the variation described by each of the first three principal components at each position along the backbone of the protease structure.



(a)

(b)

Figure 6.8: The structural variation of the protease described by PC1. The most negative projection observed is shown in blue, the most positive in grey. (a) shows a view down from the PR flaps and (b) along the active site cavity. Residues which undergo significant changes along PC1 are highlighted with lighter shades used for the positions at the negative extrema, darker for the positive. In (a) red highlights the 77 to 79 loop and purple residues 35 and 45 and in (b) residues 68 and 69 are depicted in brown.

monomers highlighted in blue in Figure 6.8a. The region around residue 79, between the cheek sheet and wall turn, however, is seen to change conformation and move towards the active site (this change is highlighted in red in Figure 6.8a). The separation of residues 35 and 45, and 25 and 79 were chosen as metrics to investigate in the full selection of sequences. Figure 6.7a shows that the projections corresponding to the WT and two triple mutant systems have negative values whilst the VPE sequences have positive values. This gives us an expectation of structural expansion, including increased separation between residues 35 and 45 in the VPE structures, and a reduced distance between 79 and the catalytic ASP 25 relative to the WT. Figure 6.9 shows the averages of the distance between the C_α atoms of residues 35 and 45, and residues 135 and 145. In all cases except the two patient derived sequences the average distance between the two pairs of residues were around 14 Å and 14.2 Å (the averages are slightly higher for

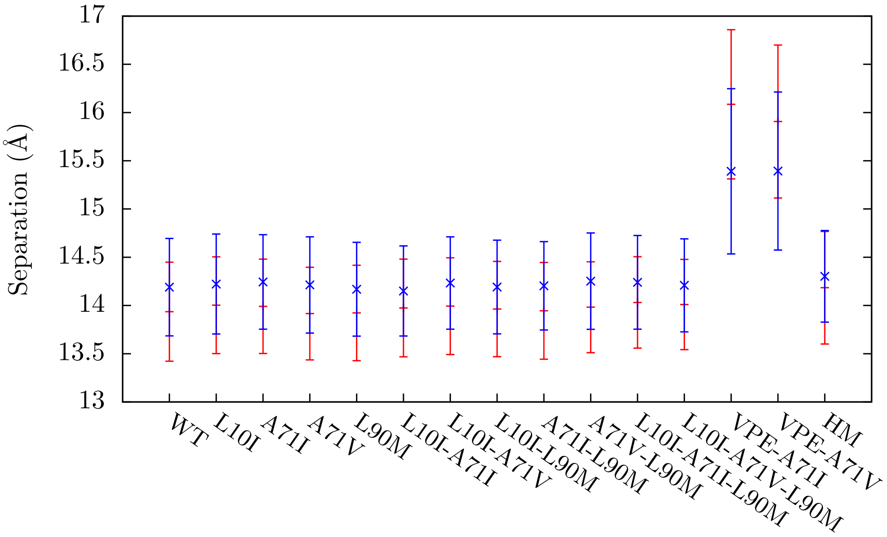


Figure 6.9: Separation of residues 35 and 45, and residues 135 and 145 for all systems under study, shown in red and blue respectively. Distances are measured between the C_α atoms of both residues. The error bars indicate the standard deviation.

HM but the standard deviations on the measurements make it hard to evaluate the significance of this change). The VPE systems, however, have separations of 16 Å and 15.4 Å for the first and second monomers respectively. This indicates that both systems do indeed enlarge in this dimension and also that the change is asymmetric with the change in the first monomer being around 2.0 Å and only 1.2 Å in the second. The predicted movement of residues 79 and 179 towards the active site ASPs in the VPE sequences is shown in Figure 6.10, the changes are relatively small however with a differences of only 0.8 Å and 0.4 Å for the first and second monomers respectively. The second monomer of HM undergoes a similar change to that seen in the patient sequence based systems but all other measurements in the figure exhibit no change from those of WT. The fact that neither the enlargement of the structure nor the alteration of the position of residue 79 seen in the VPE systems are replicated in HM indicate that a different mechanism is causing the resistance in the two cases.

Figure 6.11 shows the structural variations described by PC2. The elbow, fulcrum and cheek sheet of the second monomer move as a unit in accordance with the large magnitude variation seen in this area in Figure 6.7c. These sections of the protease have previously been observed to move as a rigid unit and are believed to facilitate flap opening and impact upon substrate specificity [390, 391]. Another potentially significant shift occurs in the separation between the residues around 71 (in the cheek sheet) and those around 93 (in the loop between the helix and whiskers), and those around 171 and 193. This rearrangement is of particular interest as it involves the area surrounding the mutated residue in position 71. The separation between 71 and 93 increases along with the value of the projection. This means that according to Figure 6.7b the two triple

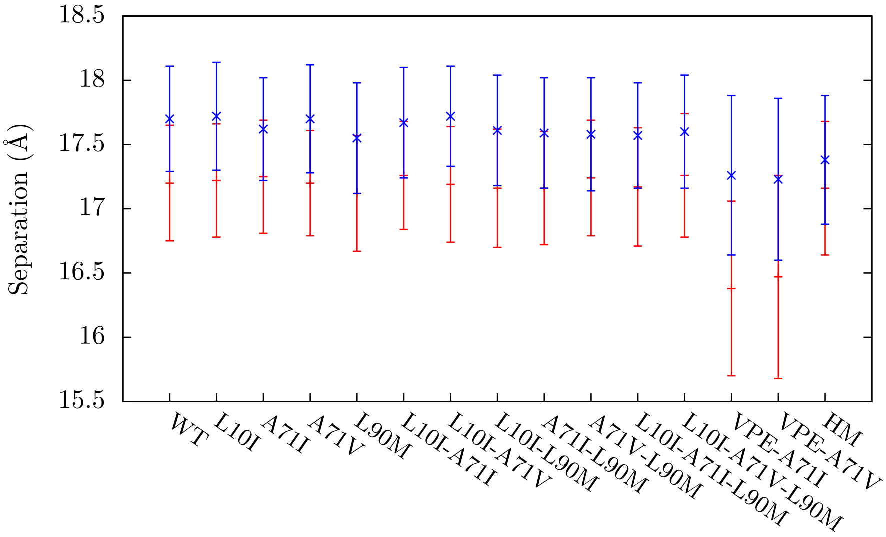


Figure 6.10: Separation of residues 25 and 79, and residues 125 and 179 for all systems under study, shown in red and blue respectively. Distances are measured between the C_α atoms of both residues. The error bars indicate the standard deviation.

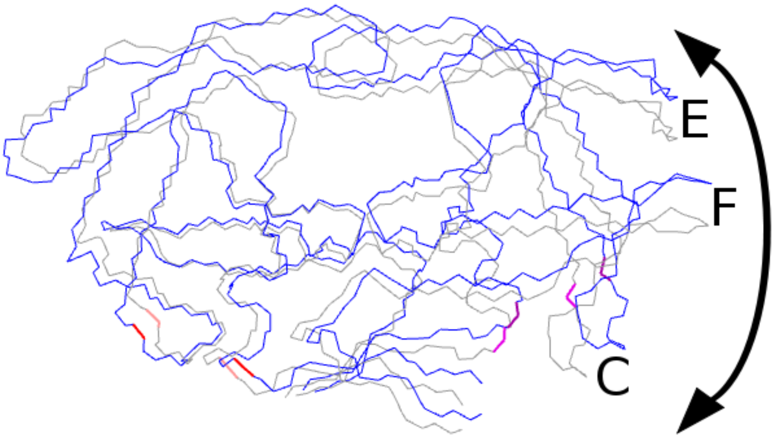


Figure 6.11: The structural variation of the protease described by PC2. The most negative projection observed is shown in blue, the most positive in grey. The (E)lbow, (F)ulcrum and (C)heek sheet of the second monomer move as a unit. Red and purple are used to highlight residues 71 and 93, and 171 and 193 respectively which undergo significant shifts along PC2 (lighter shades are used for the positions at the negative extrema, darker for the positive).

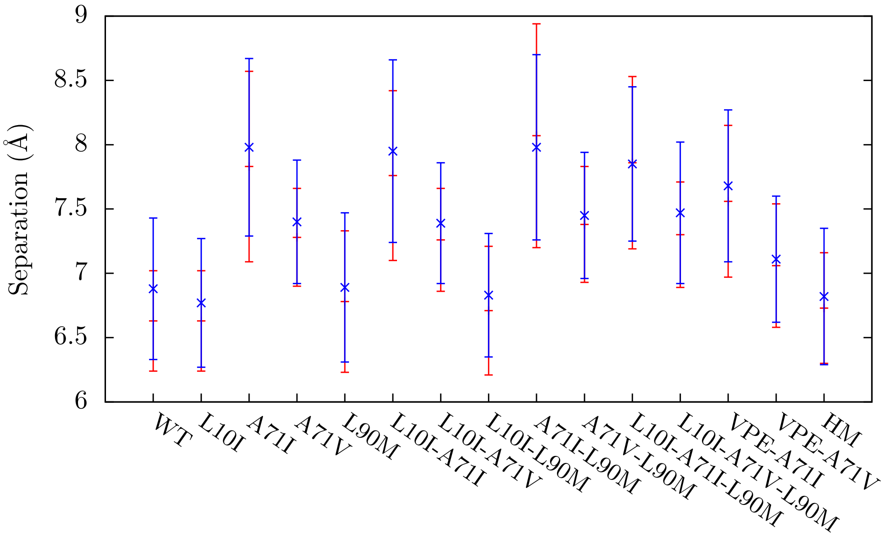


Figure 6.12: Separation of residues 71 and 93, and residues 171 and 193 for all systems under study, shown in red and blue respectively. Distances are measured between the C_α atoms of both residues. The error bars indicate the standard deviation.

mutants (L10I-A71I-L90M and L10I-A71I-L90M) should have the largest value, then the patient sequences and then the WT. In both cases when the sequences vary only by a different residue at position 71 the system containing isoleucine has a higher projection value than that containing valine. This observation is in line with these changes being caused by the structure accommodating the mutation at position 71 as isoleucine is bulkier than valine. Measurements of the separation of residues 71 and 93, and residues 171 and 193 for all systems under study are shown in Figure 6.12. These data show the expected pattern of differing separation between 71 and 93 depending on the size of the residue at position 71 (the wildtype contains alanine which is smaller than either of the mutant residues). In both cases the VPE sequences have lower separation than the other mutants containing the same residue at position 71 (in the first monomer the differences are 0.3 Å for both sequences in the second they are are 0.2 and 0.4 Å for VPE-A71I and VPE-A71V respectively).

Figure 6.13 and Figure 6.7c both show that PC3 largely describes fluctuations of the flaps which are known to be highly flexible. The flap conformational changes are similar to those seen in other studies of bound protease [392]. Only slight shifts in the projection values are seen between the different systems studies in the PCA (see Figure 6.7b) and as such the motions seem likely to be part of the fluctuations experienced by all structures. The projections for the two VPE sequences show a slightly larger spread, reflecting the added flexibility observed in this region in Figure 6.5 and Figure 6.3.

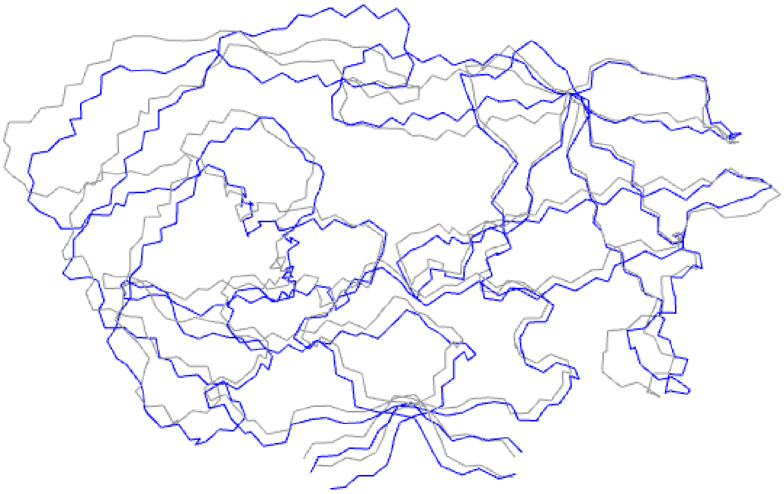


Figure 6.13: The structural variation of the protease described by PC3. The most negative projection observed is shown in blue, the most positive in grey. The most significant changes occur in the flaps and nearby cheek sheet of the first (leftmost) monomer.

6.3.2.2 Dimer Interface Conformation

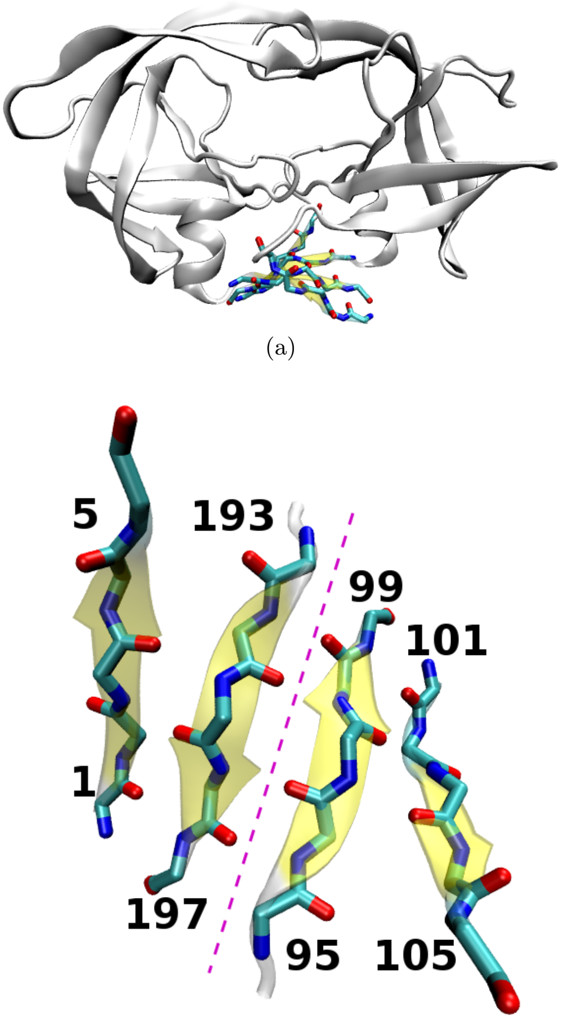
Based on crystal structure evidence, Skalova *et al.* [393] suggested that one effect of the A71V mutation was the alteration of the conformation of the dimer interface. They claimed that this change is then communicated to the elbow, fulcrum and cheek sheet regions. These are the same regions which exhibit conformational changes, correlated with deformations around residue 71, in PC2 in the aforementioned principal component analysis. Along with the evidence of changes induced in static structures, NMR evidence suggests that the four stranded dimer β sheet is divided into two sections, allowing bending about the centre line (shown in Figure 6.14) [394]. Changes in the relative orientation of these sheets are hard to detect in the PCA analysis, so direct analysis was applied to see if such changes are apparent in the simulations presented here.

The cross product of the vectors running between the first and last residue of each strand, as identified in Figure 6.14b, was used to define the normal of a plane representing each half of the β sheet. The dot product of the two normals was then used to calculate the angle, θ, between the two planes. The average angle of each system under investigation is shown in Table 6.8. The two VPE mutants increase θ and exhibit the largest change from the WT of the other systems. The change produced by the VPE-A71V is 1° larger than that in the VPE-A71I sequence, however, the behaviour in the latter system is particularly notable as all other A71I containing systems decrease θ. The change in the angle is not reproduced in the resistant HM system, again suggesting that the mechanism of resistance encountered in the VPE sequences is of a different character to that of systems where direct active site mutations are involved.

Table 6.8: The average of the angle, θ, between the planes formed by strands of the dimer β sheet either side of the dashed line indicated in Figure 6.14b. The standard deviation is given in brackets and the difference between each system and the WT value is also shown.

Sequence	$\langle\theta\rangle$	$\boldsymbol{\Delta}\langle\theta\rangle$
WT	146.83 (7.04)	-
L10I	147.69 (6.99)	0.85
A71I	145.77 (7.53)	-1.06
A71V	147.28. (6.87)	0.44
L90M	146.11 (7.66)	-0.72
L10I-A71I	146.09 (6.96)	-0.74
L10I-A71V	147.04 (7.28)	0.20
L10I-L90M	146.87 (7.87)	0.03
A71I-L90M	145.28 (7.84)	-1.55
A71V-L90M	147.15 (7.67)	0.31
L10I-A71I-L90M	145.32 (8.16)	-1.51
L10I-A71V-L90M	147.15 (8.47)	0.32
VPE-A71I	149.19 (7.75)	2.36
VPE-A71V	150.18 (7.55)	3.35
HM	146.38 (9.63)	-0.45

The difference in conformation observed in the VPE structures is suggestive of an effect being mediated by the mutation of the nearby residue 93 (from isoleucine to leucine). Mutations at position 93 are strongly associated with substrate recognition [395]. If this conjecture is correct, then a plausible hypothesis to explain the relative rarity of the A71I at position 71 is that, without other mutations in the sequence, it distorts the dimer interface in a way that diminishes the ability of the enzyme to discriminate natural substrates, reducing viral fitness. The VPE sequences must be viable as they are derived from patient data, this would suggest that the positive changes in θ which they exhibit do not hinder recognition of natural substrate. The structural changes induced in the A71V containing variants also result in positive changes in θ, in contrast to the negative changes observed in the A71I containing sequences (except VPE-A71I). It is plausible that this difference (allied to the smaller distortion of the region measured by the separation between residues 71 and 93 presented in Figure 6.12) explains the lower fitness penalty apparently associated with the introduction of A71V compared to A71I, in the absence of compensatory mutations. Some credence is lent to this idea by *in vitro* experiments that show HIV-1 sequences containing protease with A71V have increased replicative capacity [292].



(a)

(b)

Figure 6.14: The dimer β sheet location and conformation within the HIV-1 protease. (a) shows the dimer β sheet in the context of the overall structure of the HIV-1 protease. (b) shows the strands of the sheet with the first and last residues of each labelled. The sections either side bend about the dashed line.

6.4 Conclusions

The aim of the Virtual Patient Experiment was to show the potential of molecular simulations to enhance or allow the assessment of predictions produced by existing clinical decision support systems (CDSS). A patient derived sequence for which three existing CDSS systems were found to give discordant resistance rankings for the drug lopinavir was identified using the ViroLab comparative drug ranking system (cDRS). The Virolab virtual laboratory was also used to identify the HIV-1 protease mutations that were considered when producing these predictions. The list of mutations (L10I, A71IV and L90M) were simulated in the context of the HXB2 wildtype and in all possible combinations along with the full patient sequence (with both isoleucine and valine in position

71). Using the BAC [355] tool to automate the system setup, data transfer, simulation and analysis allowed us to easily run simulations across a variety of resources on both the US Teragrid (Ranger and Kraken) and the EU DEISA network (HECToR, SARA and LRZ). To simplify the analysis of the theoretical minimum turn around time (TAT) of BAC orchestrated free energy calculations consider the case when only the Ranger machine (with 62976 cores) at the Texas Advanced Computing Center (TACC) is used for the production simulations. The optimally scaled rate of computation was approximately 4 h/ns on 64 Opteron cores/replica. The theoretical minimum turn around time (TAT) for the 13 sequences investigated for this study (using 50 replicas each producing 4 ns trajectories) using 41600 cores (simulating 650 replicas simultaneously) could be as short as 16 hours. In practice peak performance was around 300 ns/day and generally far lower. MMPBSA and normal-mode post processing took 3 and 20 h/ns, respectively, but was run in parallel for each nanosecond using the Leeds node (256 cores) of the U.K. National Grid Service and the local Mavrino cluster (96 cores). The theoretical minimum total turn-around time (simulation + free energy calculation) using this approach was thus approximately a week. This TAT would allow such simulations to be provided on a clinically relevant timescale assuming suitable levels of resources were available. The BAC simulation workflow evaluated in Chapter 5 was used to simulate and provide binding affinity calculations for each of these sequences determined that other than the two full patient sequences (VPE-A71I and VPE-A71V) all were susceptible to lopinavir (i.e. had a binding free energy as attractive, or more attractive than, the HXB2 wild type within the resolution of our calculations)[8]. The full patient sequences, however, would be ranked as having intermediate levels of resistance. Using the absolute binding energy, ΔG_{theor}, VPE-A71I and VPE-A71V have changes in binding affinity comparative to WT of 1.98 and 2.07 kcal mol^{-1} respectively. These values are close to that shown by the known resistance mutant AS, of 2.7 kcal mol^{-1}, in the multiple drug resistant mutants used to validate the ability of our simulation and free energy calculation protocol to reproduce *in vitro* experimental results. Excluding the entropic component of the calculation (which is known to have less reliable convergence properties compared to the MMPBSA part of the free energy calculation) both systems remain resistant but the VPE-A71I system has a $\Delta\Delta G_{MMPBSA}$ value of only 0.98 kcal mol^{-1} whilst VPE-A71V value remains high at 2.07 kcal mol^{-1}. This ranking provides evidence that in many circumstances the entire sequence may need to be considered in order to gain an accurate assessment of the drug resistance of the virus infecting a particular patient. In the case of the specific sequence considered here (containing L10I, I13V, K14KR, I15V, K20T, L63P, A71IV, V77IV, L90M and I93L relative to HXB2 wildtype) the binding

[8] This is now the consensus result of the systems with their current rule sets, updated since the instigation of this study. The rules of all CDSS referred to in this chapter are constantly revised as new patient data and literature becomes available. The results of this study were not considered by any of the CDSS.

affinity calculation presented here is not the only evidence available as a correlational study of resistance to patient genotype has identified a pattern of mutations at positions 10, 63, 71, 90 and 93 as being associated with protease inhibitor resistance (albeit few patients treated with LPV were involved in the cohort) [264].

In addition to the binding affinity results, molecular simulation also allows us to gain mechanistic insight into patterns of resistance. Here we have shown that the full patient sequences both containing both A71I and A71V adopt substantially different conformations to those of other resistant mutants such as the hexamutant (HM) system used as a high resistance benchmark. In the VPE mutants the overall enzyme conformation is expanded compared to WT (and all other mutants investigated in this study) with the distances between residues 35 and 45 on both monomers increased by 2.0 Å for the first and 1.2 Å for the second. Contrary to this overall movement, residue 79 on both monomers bends in towards the active site. The structural deformation caused by replacing alanine with the bulkier valine or isoleucine and seen in all other structures containing A71I or A71V is reduced. Some correspondences in the mode of resistance compared to AS and HM do exist, the active site hydrogen bond network is similarly perturbed and water molecules more commonly occupy the catalytic cavity in all resistant systems we have simulated.

Further structural changes can be observed in the dimer β sheet. The four strands of which form two sub sheets which can bend relative to one another. The angle between the two pairs of strands is substantially increased in the resistant VPE mutants compared to WT. This is notable in the case of VPE-A71I as all other mutants studied containing the A71I substitution have a decreased angle. It is possible that such deformations may impact upon the enzymatic fitness of the protease sequence and hence at least partially explain the rarity of the A71I mutation compared to A71V which is known to increase the replicative capacity of HIV-1 *in vitro* [292].

Chapter 7

Towards an Understanding of NNRTI Binding in HIV-1 Reverse Transcriptase

7.1 Introduction

As discussed in Chapter 4, the most common drug cocktails used in clinical treatment of HIV-1 infection contain inhibitors of two enzymes which play vital roles in the HIV-1 life cycle: protease (PR) and reverse transcriptase (RT). In the last two chapters we have detailed the application of molecular simulation and the MMPBSA method of free energy estimation to the investigation of drug resistant variants of the HIV-1 protease. An obvious question is whether this simulation and analysis protocol can be adapted to successfully assess free energy differences in HIV-1 RT. The Binding Affinity Calculator (BAC) tool, developed to automate protease simulations, was enhanced to allow the construction, simulation and analysis of RT models.

HIV-1 RT is approximately five times the size of HIV-1 PR and the consequent increase in computational effort required to simulate it has led to it being a much less popular target for such investigations. In addition, the complexity and flexibility of many regions of the structure (as described in Chapter 4) make the challenge of investigating the RT more daunting, albeit potentially more rewarding.

An important question, that remains a topic of debate, is the precise method of inhibition of the RT polymerase activity by the Non-Nucleoside RT Inhibitor (NNRTI) class of drugs. Molecular dynamics simulations offer the possibility of gaining insight into this mechanism, an enticing prospect motivating our study of both the free energy of binding of this class of drugs and their impact on the structural and dynamic properties of HIV-1 RT.

Although far fewer molecular simulation studies of HIV-1 RT have been conducted compared to protease, some insights into the system have been produced in this fashion. One of the earliest MD studies of RT, conducted by Madrid *et al.* [312] investigated the stability of the open conformation of the apo enzyme. A structure bound to double stranded DNA (PDB structure 2HMI) with the template and primer removed was used to create a series of eight one nanosecond simulations in implicit solvent. Six of these showed the p66 thumb moving towards a configuration similar to that seen in the closed crystal structures where it stabilises. In the remaining two simulations the thumb was seen to move further from the active site before stabilising. These results show reasonable agreement with the experimental results of Kensch *et al.* [317], where it was observed that 65% of RT was in the closed form at 273 K and 95% at 313 K (the simulation was run at 298 K). The conformational change, however, happens on an unrealistically fast timescale of 30 ps to 120 ps due to the use of an implicit solvent. A more recent study by Carvalho *et al.* [396] used a range of crystal structures to create a homology model of the unliganded enzyme with the p66 conformation largely based on that of the 1DLO structure. Simulations in explicit solvent indicated that the closed form was stable through out the 1 ns run. Despite this stability, few contacts between the thumb and fingers were observed. Whilst overall the structure showed only small fluctuation the fingers between residues 120 and 150 were seen to be flexible.

Another study by Madrid *et al.* [397] compared the motions of the unliganded system to that of the DNA liganded system. They found that the DNA bound system was more flexible and that the binding of the ligand affected the way the different domains interacted (as measured by the cross correlation of the constituent residues). Major changes were observed in the p66 connection, which was anticorrelated to the other p66 domains and correlated to the p51 thumb in the unliganded case. In the DNA case, the connection shows no strong correlations at all, however strong anticorrelations between the p66 thumb and RNaseH domain, which are not seen in the unliganded case, are present. The unliganded form simulation saw the RNaseH as anticorrelated with the p66 fingers, palm and connection. These flexibility results qualitatively agree with Gaussian network models of the same systems [398].

While in the DNA case the motions seen in MD simulations agree with those of network models, a different picture is seen in the comparison of MD simulations of the NNRTI liganded system and the network models of Temiz & Bahar [47]. Two MD simulations which look at the comparative dynamics of unliganded and NNRTI bound RT are available, one performed by Shen *et al.* [333] and one by Zhou *et al.* [399]. The former is a steered molecular dynamics study which showed that the motility of the thumb is severely reduced by the binding of an NNRTI, although the simulation did not include the p51 subunit or the RNaseH domain. This does not agree with the network model findings, which show that the thumb retains motility while the interdomain correlations are altered. The network models of Temiz & Bahar [47] consider the apo RT structure and two examples of the NNRTI bound enzyme (nevirapine and efavirenz were the drugs involved). In the NNRTI bound forms the motion of the p66 fingers and thumb were seen to decouple but the size of motions were altered only slightly. The efavirenz model did actually exhibit slightly lowered thumb motility. It has been posited that this may be a partial explanation for the greater effectiveness of efavirenz as an inhibitor [47]. There are reasons to be cautious about the findings of both of these models. The lack of inclusion of much of the enzyme by Shen *et al.* [333] is a major weakness, considering the fact that the NNRTIBP involves residue 138 of p51 and the interdependency of the domain motions seen in the earlier work by Madrid [397]. The network model, by its nature, is a course grained model and may not include important interactions, which may determine the interaction and motility of the enzyme. The results shown in [399] however come from simulations involving the entire enzyme in atomistic detail. This study shows a reduced motility of the p66 thumb when Nevirapine binds, but that some of this motion is restored by in a triple mutant (V106A, Y181C, Y188C). This simulation was, however, only 1 ns long and the level of motion was not large over this timescale, meaning that motility was judged by the spread of points in a principal component analysis of the systems motion.

The motions of mutant forms of RT have been little studied in comparison with those of PR. There is one major exception to this, Rodríguez-Barrios & Gago [342], Rodríguez-Barrios *et al.* [344] have used targeted molecular dynamics, in which forces are imposed on a structure to guide it into a target conformation, to suggest that the K103N may not be related to the ability of NNRTIs to remain bound to the NNRTIBP but instead that it creates a greater energy penalty to the creation of the binding pocket in the first place. In the earlier of the two studies [342] the sidechain of Y188 was seen to rotate before that of Y181 when the NNRTIBP was created. This preference disappeared in the second study when NNRTIs were placed at the putative entrance to the binding pocket, with the order of change seemingly being defined by the plasticity of the incoming drug [344].

In addition to the dynamic properties previously discussed, Zhou *et al.* [399] report that using MMPBSA they see a reduction in binding affinity in the mutant system, although this is unsurprising considering that this triple mutant replaces a number of the bulkier NNRTIBP residues with smaller ones. There is, however, evidence from earlier work [120] which showed that the MMPBSA in conjunction with molecular docking techniques could predict the lowest energy binding mode of efavirenz using simulations of only the 20 Å area of RT surrounding the binding pocket. The same group have also attempted to use the technique as part of a screening protocol for drug candidates [400]. The simulations used in that case, however, completely neglected the motion of the enzyme concentrating on simulating the ligand motion alone. The results showed some ability to discriminate drug candidates but failed to identify several known NNRTIs within the test set and the ranking between those that could be identified was only moderate. This may well reflect the combination of using a single simulation and the failure to describe the motion of the binding pocket. A more recent study has indicated that short simulations combined with the MMPBSA free energy can correctly reproduce the experimental relative ranking of different NNRTI binding affinities [401]. Modelling studies of the NVP bound RT (with the system truncated to contain only the protein and water molecules within 10 Å of the NVP drug) have indicated that water bound along side the inhibitor can increase ΔG by as much as -5.45 kcal mol^{-1}[402]. All of these studies have been based on short simulations (of less than 2 ns) and provided no indication of the level of convergence of the reported free energy values.

The longest RT molecular simulations to date were performed by Ivetac & McCammon [311]. This study looked at the dynamics of the apo and NNRTI bound enzyme and a structure in which an NNRTI had been removed. Their work ran 4 copies of each system, with every run producing 30 ns of simulation. They found that half of the copies of the closed system opened to a similar binding cleft width as that in DNA bound crystal structures. Four simulations of the system with NVP removed explored a similar conformation after 10 ns of simulation. They did not report any binding energy data.

The results reported in this chapter aim to bring together an understanding of the domain scale motions of HIV-1 RT with that of the binding free energies of NNRTIs. It is hoped that such insights will provide the foundations for the assessment of the impact of mutations on inhibitor binding in this system. The first section of this chapter focuses on a comparison of the conformational exploration of HIV-1 RT in its apo form, in complex with a natural DNA substrate and when bound to two widely used NNRTIs (efavirenz, EFZ, and nevirapine, NVP). This study culminates in a comparison of the binding energies of the two inhibitors and a determination of the impact of the larger subdomain motions of RT upon them. The second part of this chapter is concerned with establishing the feasibility of using the BAC workflow to assess the resistance level of mutant HIV-1 RT strains.

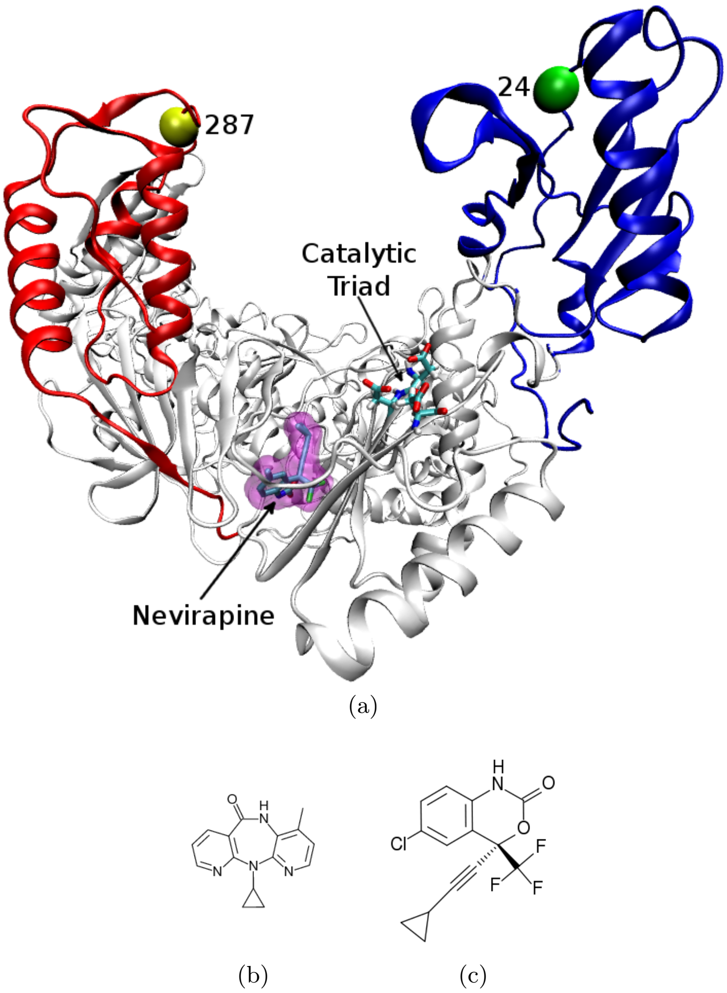
Table 7.1: The names of the HIV-1 RT systems simulated listed with their bound ligands and the PDB entry of the crystallographic structures used to create them. A structure is designated as open if a separation of greater than 15 Å exists between residue 24 in the p66 fingers and residue 287 in the p66 thumb (the locations of these residues are indicated in Figure 7.1a). The number of atoms in the fully solvated model used for our molecular dynamics simulations is also shown.

System	Conformation	Ligand Type	PDB	No. Atoms
CSD	Closed	None	1DLO [323]	164,486
OPN	Open	None	2HMI [403]	182,668
DNA	Open	Double stranded DNA	2HMI [403]	196,721
RMD	Open	None	1IKW [93]	186,096
EFZ	Open	NNRTI	1IKW [93]	186,126
NVP	Open	NNRTI	3HVT [404]	181,197

7.2 Investigation of the Impact of NNRTI Binding on HIV-1 RT Structure and Dynamics

The focus of this part of the chapter is twofold: firstly, to gain a qualitative understanding of the changes in dynamics that affect the RT enzyme under a variety of ligation states in order to inform future studies and, secondly, to determine whether MMPBSA can differentiate the binding energies of different NNRTIs. The inhibitors efavirenz (EFZ) and nevirapine (NVP) are chosen to represent the NNRTI class of drugs (the chemical structures of both drugs are shown in Figure 7.1). The intention is to develop insights which will facilitate the use of the same approach which has been successfully applied to the HIV-1 protease system where qualitative studies [286] have informed quantitative studies of the binding affinity of inhibitory drugs [288].

In this study six different HIV-1 RT systems representing the unliganded, drug bound and natural double stranded DNA substrate bound enzyme were simulated for a total of 26 nanoseconds each. The ligands bound and the crystallographic structures upon which the simulated systems are based are listed in Table 7.1. The available crystal structures, discussed in Chapter 4, can be divided into two broad conformational classes, one of which is called 'open' and the other 'closed'. In this study an 'open' conformation is defined as one in which there is a separation of greater than 15 Å between residue 24 in the p66 fingers and residue 287 in the p66 thumb; below this, a structure is designated as 'closed'. The locations of residues 287 and 24 in the open conformation are shown in Figure 7.1a.



(a)

(b)

(c)

Figure 7.1: The structure of HIV-1 RT bound to the drug nevirapine is shown in a) with the inhibitor shown in pink surface representation, the fingers in blue and the thumb in red. The polymerase catalytic triad is depicted in chemical structure and the positions of residues 24 (green ball) and 287 (yellow ball) in the p66 subunit, the separation of which is used in this study to define open and closed structures, are also shown. (b) and (c) show the chemical structures of the two NNRTIs used in this study which are nevirapine and efavirenz, respectively.

7.2.1 Methods

The simulations and free energy calculations were performed using the automated BAC tool [355]. The protocol used for structure preparation, simulation and analysis was derived from that described in Chapter 5. The following is a brief overview of this process, highlighting the changes made to the protocol for use with HIV-1 RT.

7.2.1.1 Model Creation

Unfortunately, all available HIV-1 RT crystal structures are incomplete and a number of loop residues in the p51 subunit of the NNRTI bound structures are missing (residues 217 to 231 in 1IKW, 225 to 260 in 3HVT and 365 to 352 in both structures). The models were completed by copying in the coordinates from 1HQU [343] (this structure was chosen due to its high resolution (2.7 Å) and the fact it was bound to an NNRTI) after alignment of the surrounding residues using VMD [315]. In each case the final model contains 556 residues in the first (p66) chain and 427 in the second (p51) chain for a total of 983 residues. Once the manual editing of structures was complete the rest of the simulation workflow was automated using the Binding Affinity Calculator (BAC) scripts created to automate simulations and free energy calculations for the HIV-1 protease [355]. Each system was solvated using a cubic box of TIP3P water molecules [359] with at least 14 Å distance around the protein. The systems were neutralised by the addition of Cl^- ions in the non-DNA bound systems (8, 8, 12, 12 and 10 Cl^- were required respectively for the closed apo system, open apo system, DNA bound, EFZ bound, EFZ removed and NVP bound systems) and (22) Na^+ ions in the DNA bound case.

Inhibitor potential parameterisation was performed by extracting the drug coordinates into separate files, using the PRODRG tool [357] to insert missing hydrogens. The geometries were then optimised using Gaussian 98 [358] (with the 6-31G** basis functions). The Restrained Electrostatic Potential (RESP) procedure, part of the AMBER package [70], was used to calculate the partial charges. The force field parameters for the inhibitors were described using the General AMBER Force Field (GAFF) [53]. The protein and DNA elements of all systems were described by the standard AMBER force field (ff03) [360] which is parameterised for bio-organic molecules and including DNA in particular. The default variants (such as protonation states) for amino acids in physiological conditions were used for all residues.

7.2.1.2 Molecular Dynamics

The molecular dynamics package NAMD2 [50] was used throughout the minimisation, equilibration and production stages of the simulations. Electrostatic interactions were treated using the particle mesh Ewald (PME) [405] method and SHAKE [66] constraints were applied to all bonds involving hydrogen atoms in order to employ a 2 fs integration time step. Minimisation was conducted using the conjugate gradient and line search algorithms for 2000 iterations of each system. During this process all heavy atoms were restrained using a force constant of 5 kcal $mol^{-1}Å^2$.

The next stage of the equilibration process was a mutational relaxation protocol in which each mutated residue and residues within 5 Å are released in turn from the restraints for 50 ps. This allowed the residues to reorientate into more favourable conformations if necessary. After the 50 ps relaxation period the restraints are reapplied to each region.

The equilibration phase anneals the system taking the temperature from 50 K to 300 K in 50 ps. Once achieved, the final temperature was maintained using a Langevin thermostat with a coupling coefficient of 5 ps^{-1}. This was followed by completely isothermal equilibration for 200 ps in the canonical (NVT) ensemble. In both of these stages the restraints imposed during minimisation were retained. The restraints were then gradually reduced in four steps of 1 kcal mol^{-1}Å^{2}, each step running for 50 ps. The restraints applied are weaker than in the protease case as no regions are known to suffer solvation induced deformities unlike the flap region of the protease. After this, the restraints were removed completely and the systems allowed to evolve under isothermal-isobaric (NPT) conditions using a Berendsen barostat [59] with a target pressure of 1 bar and a pressure coupling constant of 0.1 ps. Coordinate trajectories were recorded every 1 ps throughout all equilibration and production runs.

The simulations presented here were run using 512 cores on the Intrepid[1] BlueGene P machine at the Argonne National Laboratory, achieving a rate of 2 h/ns. Additional simulations were run utilising both the 62,976 core Ranger machine[2] at the Texas Advanced Computing Centre (TACC), part of the US Teragrid, and the 3,328 core Huygens system[3], administered by SARA in the Netherlands, which is part of the EU DEISA grid. A simulation rate of approximately 3.5 h/ns was achieved on 256 cores per system. These last two resources are comparable to those used to perform the protease simulations described in Chapter 5 and Chapter 6 where only 64 processors provided optimal processing speed.

7.2.2 Analysis

The analysis presented here has two main purposes; firstly to explore the changes to enzyme dynamics made by the binding of NNRTIs to HIV-1 RT and, secondly, to calculate the binding affinities of the drugs NVP and EFZ to the enzyme. To investigate the former problem two techniques are used: cross-correlation matrices and principal component analysis. The reason for employing both related methodologies, is that cross-correlation matrices provide a compact way in which to investigate changes in the relative subdomain motions between the very different closed and open overall enzyme conformations

[1] Intrepid: http://www.alcf.anl.gov/resources/storage.php

[2] Ranger: http://www.tacc.utexas.edu/resources/hpcsystems/

[3] Huygens: www.sara.nl/userinfo/huygens/index.html

without the gross structural variance dominating the analysis. PCA is used to analyse the structures explored by the open conformation systems and the differences between them.

7.2.2.1 Cross Correlation Matrices

Cross correlation matrices are used to identify concerted motions seen in simulation trajectories. The matrices were constructed for the various systems by superimposing the C_α coordinates of snapshots from the trajectory on an average structure using the ptraj program which is part of the AMBER package[70]. The elements of the matrix are given by

$$C(i,j) = \frac{< \Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j >}{< \Delta\mathbf{r}_i >^{\frac{1}{2}} < \Delta\mathbf{r}_j >^{\frac{1}{2}}}, \tag{7.1}$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the displacement vectors of the i^{th} and j^{th} atoms respectively. Providing that the angle between two such vectors is reasonably close to 0 or 180°, then this value will identify pairs of atoms whose motion is correlated[406]. Complete correlation is signified by a $C(i,j)$ value of 1, anti-correlation by -1. The matrices described in this chapter were calculated using the entire 20 ns post-equilibration stage simulations for each system, although similar results were obtained on each individual nanoseconds of simulation and with subsets of data down to 200 ps.

7.2.2.2 Principal Component Analysis

In this study, dimensional reduction via principal component analysis (PCA) is performed upon a concatenated trajectory of the production phase of all open conformation systems (OPN, DNA, RMD, EFZ and NVP) in order to isolate the most significant conformational differences between the structures explored. The correlation matrix is calculated from the molecular dynamics trajectory (after alignment with the average structure) and then diagonalised. This provides an orthogonal set of eigenvectors representing linearly independent modes of conformational change which are the principal components. The eigenvalues associated with each principal component are a measure of the variance in the original dataset described by that component. The principal component analysis presented here was performed on the backbone coordinates of the concatenated trajectory using the bio3d package [407], in order to elucidate how binding DNA or an NNRTI alters the conformations explored by HIV-1 RT in our simulations. Structures were sampled from the combined trajectory every 10 ps.

7.2.2.3 Free Energy Calculations

After the equilibration is complete, structures output every 10 ps were post-processed using MMPBSA, meaning that a hundred sets of coordinates were analysed for each nanosecond of simulation. Unlike the HIV-1 protease cases described in Chapter 5 and Chapter 6 the free energy calculations presented here neglect the conformational entropy component of the binding free energy and are produced from the MMPBSA methodology alone. The increased size of HIV-1 RT renders normal mode calculations on the entire protein impractical, due to both the high computational cost and the slow convergence of the method. Other studies have applied normal mode analysis to truncated sections of the protein but no difference was produced in the ranking of systems [401]. In addition there are problems with such an approach as the physical interpretation of the results is not clear and the truncation may prevent access to significant modes. As a consequence of this it is not possible to reproduce absolute binding affinity values, ΔG, from experiment but differences in binding free energy changes may be compared.

7.2.3 Equilibration

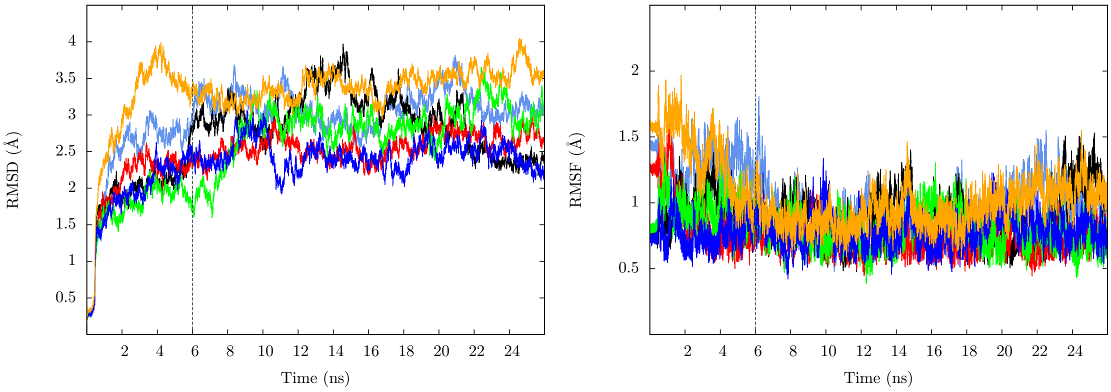
Physical properties can only be reliably calculated from systems which have been adjudged properly equilibrated. For all systems simulated here the minimisation applied is sufficient to remove all bad contacts as measured by the decrease in potential energy which, after the heating phase, remains stable with a standard deviation of less than 450 kcal mol^{-1} in all cases. A further test of whether the systems under study have equilibrated can be made by investigating the structural variation seen over the simulation.

7.2.3.1 Structural Equilibration

In order to assess whether the simulations have equilibrated, the root mean squared deviations of the systems from their initial configurations have been calculated (see Figure 7.2a). The deviations from the initial structure stabilise after approximately 6 ns, with fluctuations in all systems after this point being less than 1.5 Å. The HIV-1 RT is known to be a flexible protein, with a number of loop regions for which conformational changes might be expected throughout even equilibrated simulations and indeed some potentially substantial deviations are observed after the 6 ns cut off (particularly those approximately 8 ns into the simulation of the RMD system, from which the drug EFZ has been removed). Using difference distance matrices, Keller *et al.* [316] determined a set of residues which varied in relative position by less than 2 Å in a wide range of HIV-1

Chain	Residues
p66	4-6, 95-107, 162-163, 180-181, 188-200, 202-205, 226, 234-235, 237-239, 317, 319, 323, 339-345, 349-353, 365-366, 368-402, 405-419, 428-436, 439, 493, 530
p51	6-7, 18-45, 54-64 71-84, 97-111, 113-117, 121, 123-138, 140-174, 176-184, 186-192, 197-198, 201-202, 208, 252, 254-264, 267, 274, 277, 280-282, 284, 296, 298-300, 303-307, 320, 322, 329, 331, 333-335, 364-393, 397-417

Table 7.2: Residues determined to vary relative positions by less than 2 Å in a survey of HIV-1 RT crystal structures by Keller *et al.* [316].

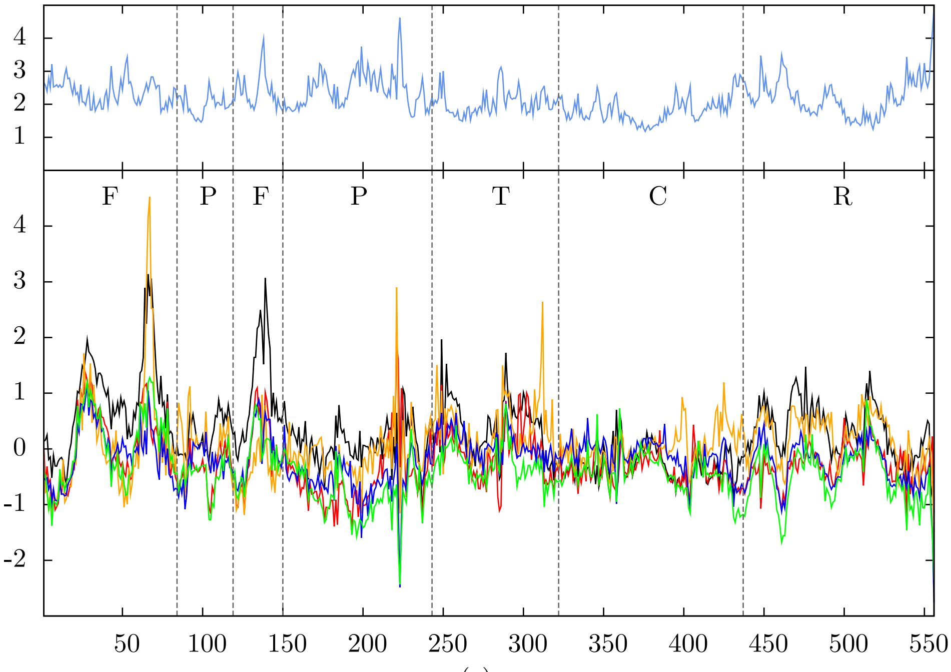


(a) (b)

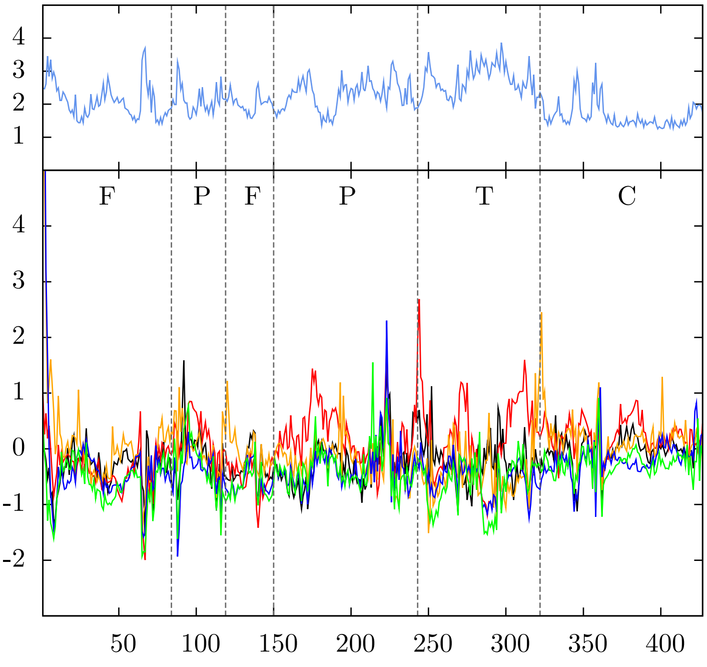
Figure 7.2: Structural variation seen over the 25 ns of simulation performed on each HIV-1 RT system under investigation measured by (a) RMSD of the protein backbone relative to the initial structure and (b) the RMSF of the most structurally stable residues (identified by Keller *et al.* [316] and listed in Table 7.2) compared to an average structure generate from the full simulation trajectories. Values for the CSD system are shown in light blue, OPN in black, DNA in red, RMD in green, EFZ in dark blue and NVP in orange.

RT crystal structures. These residues are assumed to represent the most structurally stable regions of the protein and are listed in Table 7.2. To investigate whether the motions observed after the 6 ns estimated equilibration window are confined to regions for which conformational changes are expected the RMSF of each system was calculated relative to the average structure using only the residues identified by Keller *et al.* [316] as not undergoing large structural rearrangements. Figure 7.2b shows that the RMSF of the simulations is reduced to 1.5 Å at around 6 ns and continues at or below this level throughout all the simulations. Consequently the trajectory between the start of the simulation and 6 ns in is defined as being the equilibration phase and all subsequent parts comprise the production phase.

The fluctuations of individual residues during the equilibration phase are shown in Figure 7.3 for the CSD system, together with the differences for each of the open conformation systems. In all systems the fluctuations in the p51 domain (shown in Figure 7.3a)



(a)



(b)

Figure 7.3: RMSF of each residue of the CSD system is shown above the difference in RMSF for each of the other (open conformation) systems under investigation. The residues are divided into those in (a) p66 and (b) p51. Values for the CSD system are shown in light blue, OPN in black, DNA in red, RMD in green, EFZ in dark blue and NVP in orange. The subdomains are labelled F(ingers), P(alm), T(humb), C(onnection) and R(Nase H). The p66 fingers domain is observed to undergo considerably larger conformational change in the open conformation systems; the difference is particularly pronounced in the OPN and NVP systems.

are similar with the exception of the DNA system where regions of the palm and thumb show larger deviations. The fluctuations in the CSD p66 chain (shown in Figure 7.3a) peak in the fingers, RNaseH and at residue 223. This residue is part of the loop running across the front of the palm region. The open conformation systems show larger deviations in the p66 fingers and thumb since, as expected, they lack the hydrogen bonds which link these regions in the closed system. In the NVP system there is also greater flexibility between the large peaks at residue 218 and residue 225. This difference is associated with the opening of a channel which allows water molecules to enter the NNRTI binding pocket. The relative flexibility of different regions seen here is in line with both the experimental temperature factors and other simulation results [397], giving us confidence that our models are behaving correctly and that no anomalies have been introduced during the editing of the structures.

7.2.3.2 NVP Water Entry

The additional flexibility, indicated by the increased RMSF values between residues 218 and 225, in the NVP system corresponds to deviations of a loop close to the region containing the NNRTI binding pocket. It is associated with the formation of a channel between residues 105, 106, 225, 227 and 236 during the period 0.2 to 0.7 ns into the simulation. This is coincident with the entrance to the binding pocket posited by Esnouf *et al.* [336] on the basis of the protrusion of the NNRTI delavirdine from the pocket. In the simulation, opening of the channel is induced by a water molecule which subsequently enters the NNRTI binding pocket (see Figure 7.4). The distances between the two water molecules which enter the binding pocket during the equilibration phase and the oxygen of NVP are shown in Figure 7.5. The first water entry occurs approximately 0.2 ns into the simulation before the restraints on the protein are fully removed. Only one water molecule occupies the binding pocket at a time with an exchange made between the pocket and solvent (approximately 1.4 ns into the simulation) through the same channel used for the initial water ingress. A water molecule is present in the NNRTI binding pocket throughout the remainder of the simulation. After 0.4 ns the water molecule occupying this position forms hydrogen bonds (defined as remaining within 3.5Å and with an O-H-O angle of at most 30°) with the backbone oxygen of L234 and the oxygen of NVP. The 3HVT crystal structure upon which this simulation is based does not contain any water molecules; however, one is present in this bridging location between drug and protein in the 1VRT structure of the NVP bound HIV-1 RT [408].

The distance track of the first water molecule indicates that it moves positions from approximately 7 Å to around 3 Å from the NVP oxygen between 0.2 ns and 0.4 ns into the simulation. Coincident with this movement two loops in the fingers domain,

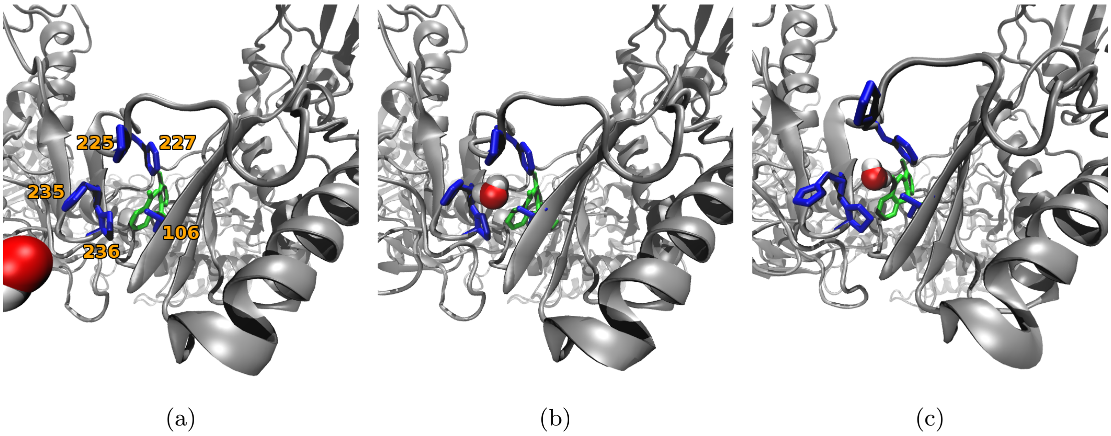


Figure 7.4: The channel through which a water molecule enters the NNRTI binding pocket is shown (a) at the start of the simulation, and after the water molecule enters at (b) 0.2 ns and (c) 0.7 ns into the simulation of the NVP bound system. NVP is shown in green with the residues which rearrange to allow water entry shown surrounding the NNRTI binding pocket in blue.

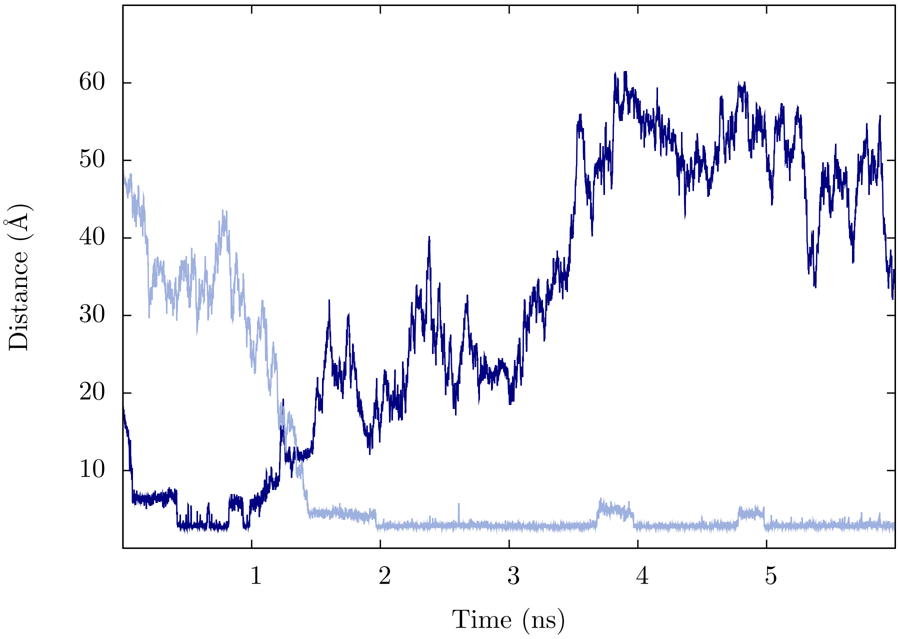


Figure 7.5: The distance between the two water molecules which enter the NNRTI binding pocket of the NVP bound HIV-1 RT during the equilibration phase of simulation and the oxygen molecule of the drug. The first water molecule (represented by the dark blue line) enters after 0.2 ns (before constraints are fully removed) and moves further into the pocket after 0.4 ns before exiting approximately 1 ns into the simulation. The second water molecule (represented by the light blue line) replaces the first after 1.4 ns.

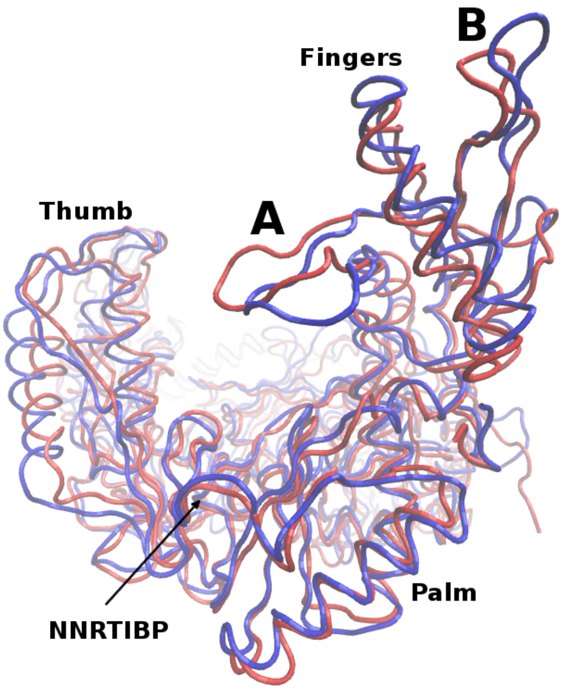


Figure 7.6: The largest conformational changes undergone by the NVP system during equilibration occur in the fingers domain. The initial structure is shown in red, that 0.7 ns into the simulation in blue. The loop containing the residues 65 and 72, implicated in dNTP binding and labelled A, and that between residues 135 and 140 (labelled B) bend away from the binding cleft. The movement is towards the viewer of the figure and from left to right, respectively.

the first between p66 residues 65 and 72 and the second running from p66 residues 135 and 140, undergo significant conformational rearrangements. Both loops move away from the substrate binding cleft as shown in Figure 7.5. These motions account for the additional flexibility observed in the per residue RMSF for the NVP system (see Figure 7.3). Whilst it is tempting to conclude that the changes in fingers conformation are induced by the changes in the NNRTI binding pocket produced by water entry, caution should be exercised as both changes occur soon after the release of constraints upon the system.

7.2.4 Production Phase Structural Analysis

Following the equilibration phase, 20 ns of simulation forms the production phase. The motions of the enzyme during this stage of the simulation were analysed to investigate how they are changed by NNRTI binding.

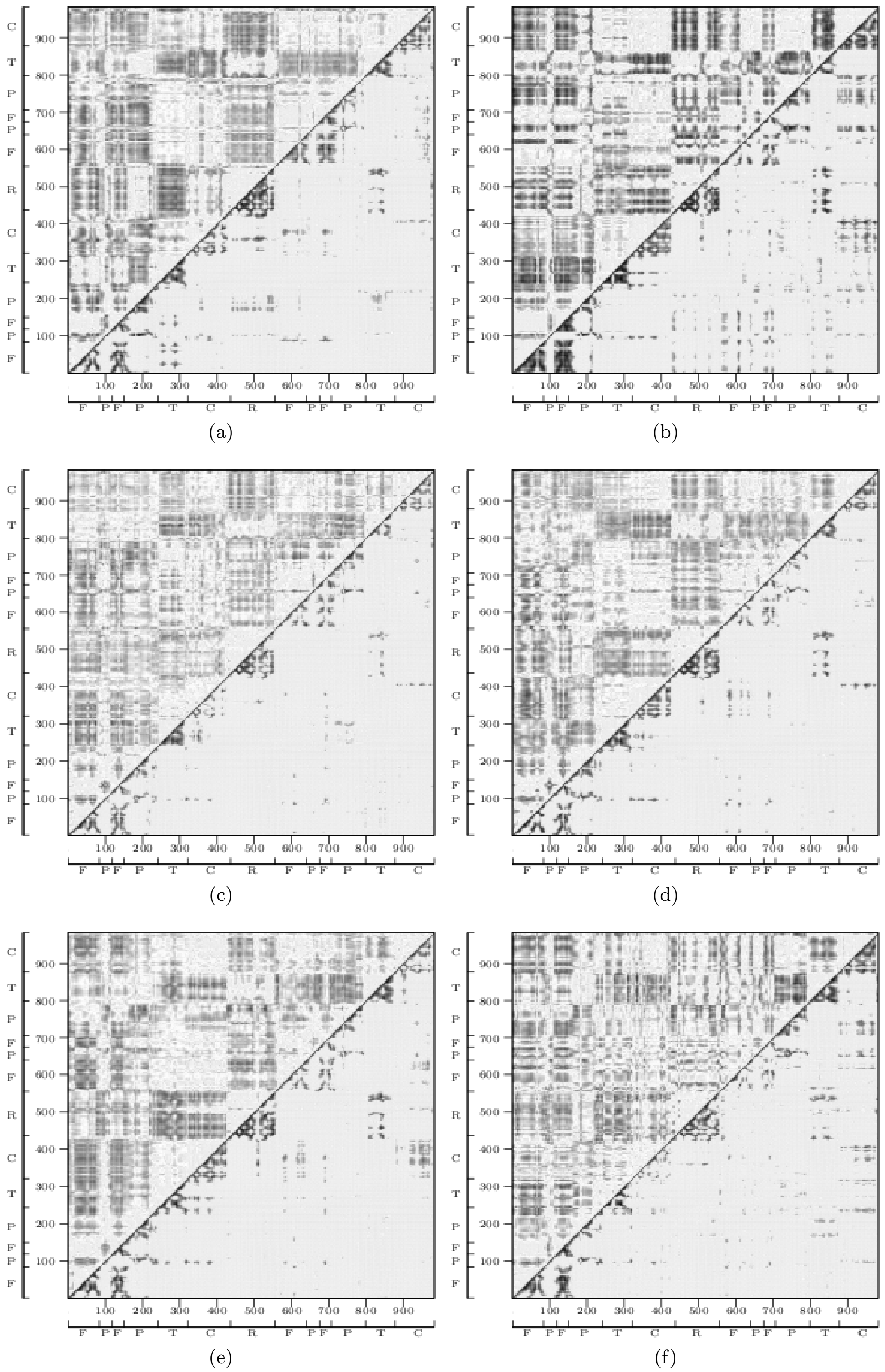
7.2.4.1 Cross Correlation Matrices

The cross-correlation matrices (Figure 7.7) show strong correlations in all the intra-domain regions, indicating that the domains in all of the systems retain their integrity. The areas showing the strongest concerted motion are in broad agreement with both previous molecular dynamics simulations [397] and network models of the unliganded enzyme [398]. The concerted motions are most pronounced in the OPN system, perhaps as a consequence of the larger motions of key subdomains observed over the trajectory (see results presented later), and are weakest in the DNA model. There is no significant evidence of NNRTI binding altering the relationship of the thumb to other domains in either of the systems reported here, with the thumb correlations closely resembling both the open unliganded and DNA bound behaviour (except with respect to the end of p51 where anti-correlations are increased in the DNA simulation). There is variation, however, between the closed conformation system where the p66 thumb is seen to correlate with the p66 fingers and the open conformation systems in which it is not. This would suggest that the correlation may be a function of the specific conformational state of the system (and the hydrogen bonding between domains in the closed system as suggested by crystallographic evidence [323]) rather than the binding of either the natural substrate or an inhibitor. This contradicts the findings of a previous study by Madrid *et al.* [397] which investigated the correlated motions of the open conformation of the apo and DNA bound enzyme and found that the p66 thumb did correlate with the fingers in the open unliganded form of HIV-1 RT. The differences between the two studies may reflect improvements in the AMBER force field used (the previous study was implemented using ff94) and/or the fact that the simulations presented here are substantially longer than those in the previous study (where only 1 ns trajectories were produced) and consequently sampling will have been improved.

In all of the systems except DNA, the RNaseH domain is correlated with the thumb and connection of the p51 subunit. A change in the relationship of this region's behaviour is in line with the fact that residues in both of these domains are seen to interact with both DNA and RNA templates in the reported crystal structures [409].

7.2.4.2 Principal Component Analysis

Figure 7.8 shows the level of variation captured by each of the first ten PCs. Only the first four describe greater than 5% of the observed differences in structure between snapshots. The first two components account for 55% of the total variation and the focus of this section will primarily be on what they can tell us about the differences between systems but the third and fourth PC will also be considered. The results of PCA are used



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7.7: Cross-correlation matrices for the simulations of the HIV-1 RT: (a) CSD, (b) OPN, (c) DNA, (d) RMD (e) EFZ and (f) NVP. Residues 1-556 represent the p66 chain and the remaining residues (557-983) the p51 chain. The subdomains are labelled F(ingers), P(alm), T(humb), C(onnection) and R(NaseH). The intensity of each position represents the magnitude of the correlations; only those with a magnitude greater than 0.3 are shown in the cross-correlation matrices. Correlations are shown below the diagonal, anti-correlations above.

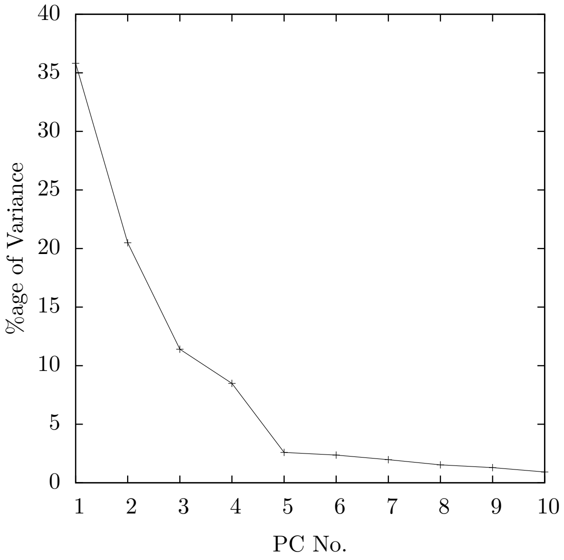


Figure 7.8: The percentage of the variation observed over the concatenated trajectory of all open conformation systems (OPN, DNA, RMD, EFZ and NVP) which is captured by each principal component (PC No, horizontal axis).

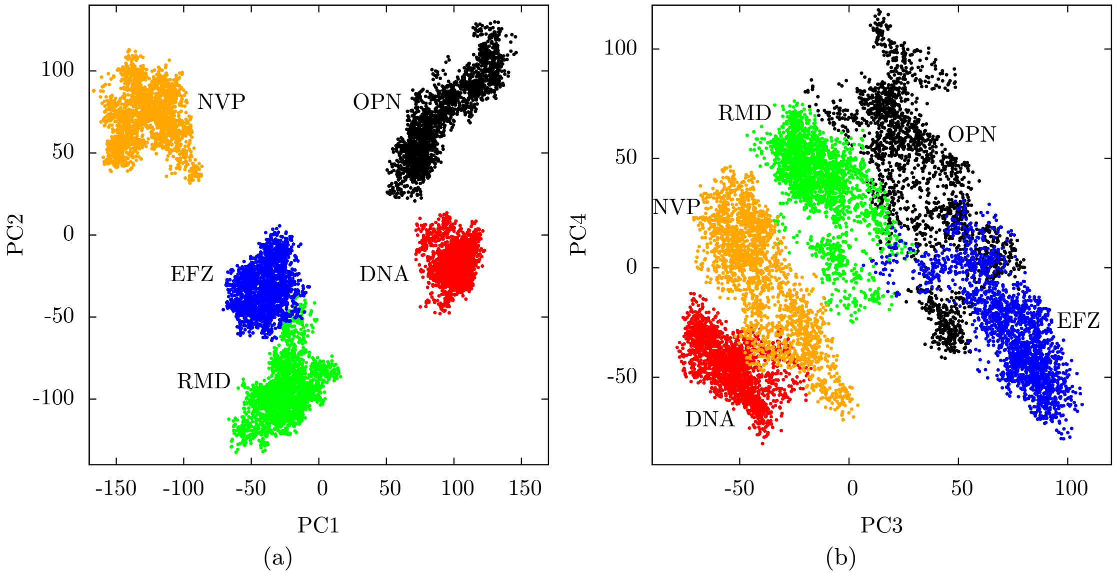


Figure 7.9: Projection of each snapshot of the concatenated trajectory for all open conformation systems (OPN, DNA, RMD, EFZ and NVP) simulated along the first four principal components. The first two PCs are shown in (a), the third and fourth in (b).

to identify metrics which help elucidate conformational and dynamic changes associated with NNRTI binding.

Figure 7.9 shows the projection of each snapshot of the concatenated trajectory along the first four principal components. Projections along PC1 distinguish between the three systems in which the NNRTI binding pocket is present (EFZ, RMD and NVP) and those without as the former have negative values and those without positive ones.

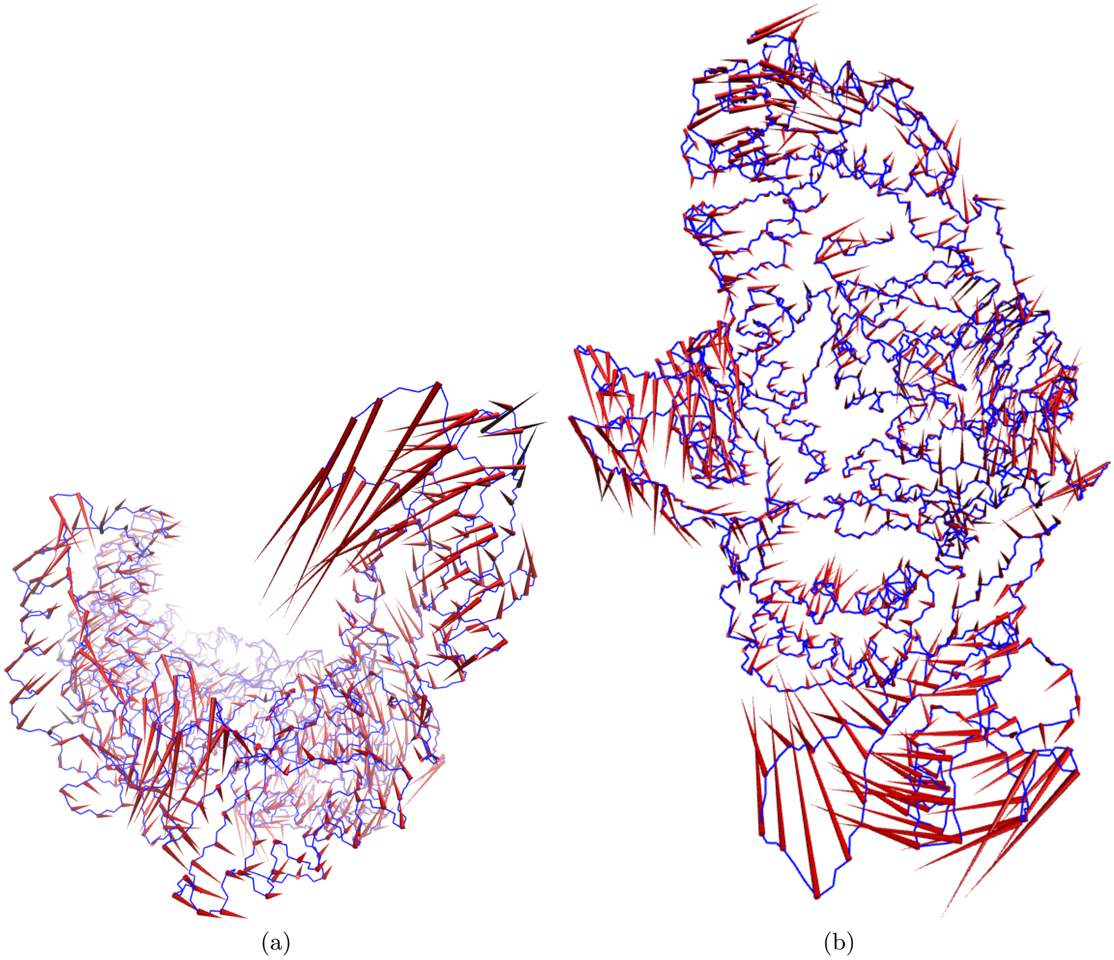


Figure 7.10: Porcupine plot showing the RT structural variation described by PC1. The structure in blue represents the most negative projection along the PC obtained from the concatenated trajectory of all open conformation systems (OPN, DNA, RMD, EFZ and NVP) simulated. The red cones show the direction of motion of the C_α atoms of the structure along the PC with the point representing the position in the most positive projection observed. The structure is shown (a) looking along the binding cleft from the polymerase active site to the RNaseH and (b) looking down from a point beyond the p66 fingers.

The projections along PC2 are similar for the EFZ and DNA systems which both have slightly negative values, the NVP and OPN are more positive and the RMD is more negative. In all systems except OPN the projections of both PC1 and PC2 are fairly tightly clustered. The OPN system moves along both vectors over the duration of the production phase (in both cases in the negative direction).

The conformational changes described by PC1 are illustrated in Figure 7.10. The key changes that can be seen are that the region at the base of the p66 thumb is deformed in the most negative projection, as is expected upon the formation of the NNRTI binding pocket, explaining the separation of the systems observed in Figure 7.9a. The more negative projections also have the thumb closer to the RNaseH and the loop containing

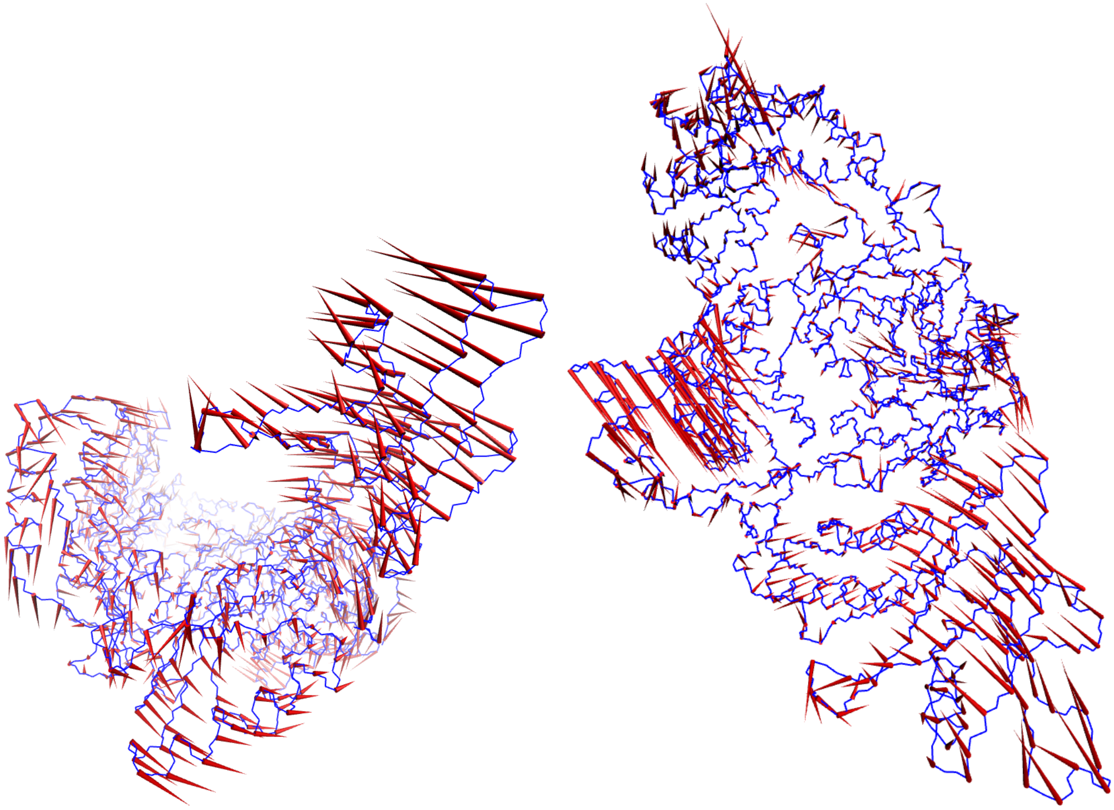
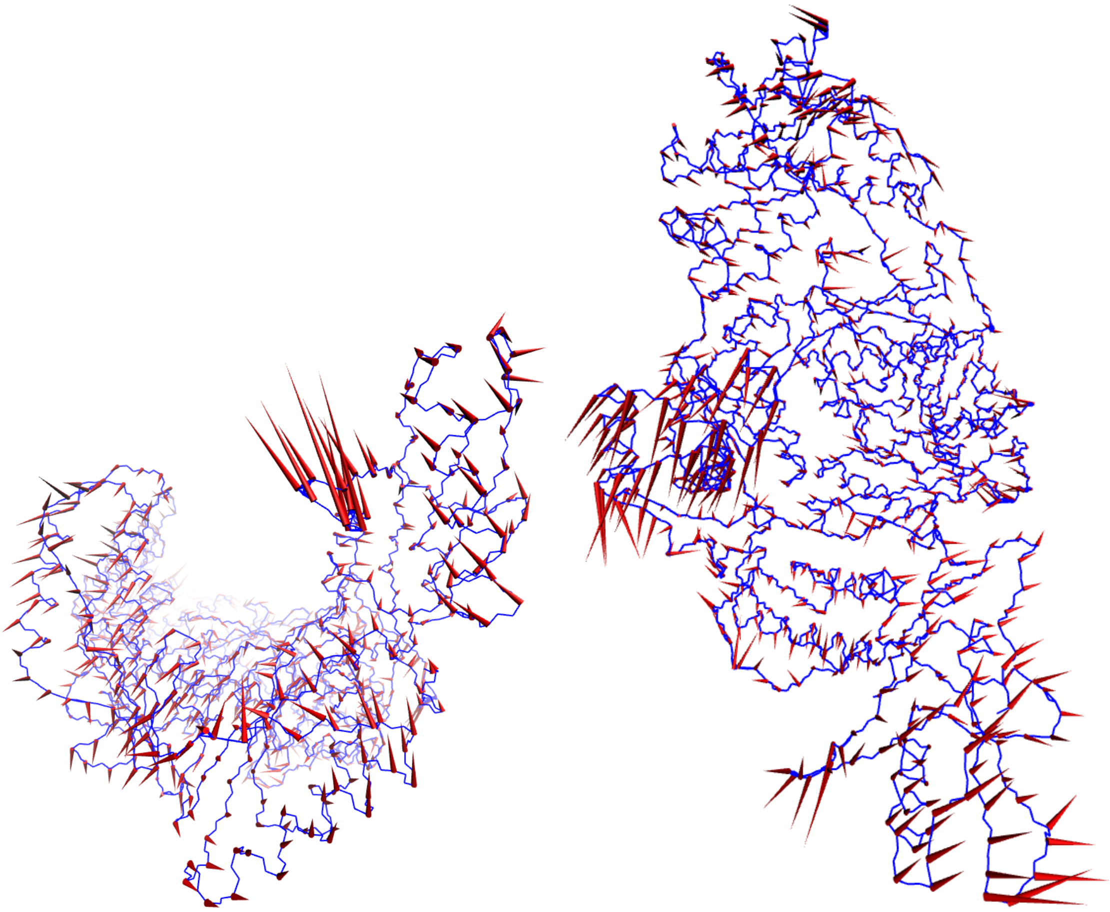


Figure 7.11: Porcupine plot showing the RT structural variation described by PC2. The structure in blue represents the most negative projection along the PC obtained from the concatenated trajectory of all open conformation systems (OPN, DNA, RMD, EFZ and NVP) simulated. The red cones show the direction of motion of the C_α atoms of the structure along the PC with the point representing the position in the most positive projection observed. The structure is shown (a) looking along the binding cleft from the polymerase active site to the RNaseH and (b) looking down from a point beyond the p66 fingers.

resides 65 and 72 in the fingers located further away from the polymerase active site. The RNaseH subdomain is also shifted to the side of the enzyme containing the p66 thumb in the more negative projections. The deformations creating the NNRTI binding pocket being correlated with a forward rotation of the thumb are expected from crystal structure evidence (as detailed in Section 4.7). However, changes in the p66 fingers are not commonly noted as effects of NNRTI binding.

PC2 is dominated by the opening of the DNA binding cleft with the p66 thumb and fingers moving in an anti-correlated fashion to close the cleft as the projection becomes more positive, as shown in Figure 7.11. Interestingly projections along this component do not separate the drug bound and apo or substrate bound systems. The projections of the OPN, RMD, EFZ and NVP systems show a spread along this component (see Figure 7.9a) indicating that they are undergoing motions in which the motions of p66



(a) (b)

Figure 7.12: Porcupine plot showing the RT structural variation described by PC3. The structure in blue represents the most negative projection along the PC obtained from the concatenated trajectory of all open conformation systems (OPN, DNA, RMD, EFZ and NVP) simulated. The red cones show the direction of motion of the C_α atoms of the structure along the PC with the point representing the position in the most positive projection observed. The structure is shown (a) looking along the binding cleft from the polymerase active site to the RNaseH and (b) looking down from a point beyond the p66 fingers.

finger and thumb domains are anti-correlated. This observation is counter to the established picture in which NNRTI binding alters the correlated motions between these subdomains.

The regions undergoing the most significant shifts in PC3 and PC4 are similar to those varying in the first two principal components. The changes represented by PC3 are seen in Figure 7.12a where the loop containing resides 65 and 72 in the fingers moves 'up' and away from the polymerase active site as the projections become more positive and Figure 7.12b that shows the p66 thumb moves away from the RNaseH. Similar motions in the fingers are described by PC4 as seen in Figure 7.13 but along with this as the projections become more positive the tip of the thumb bends towards the DNA binding

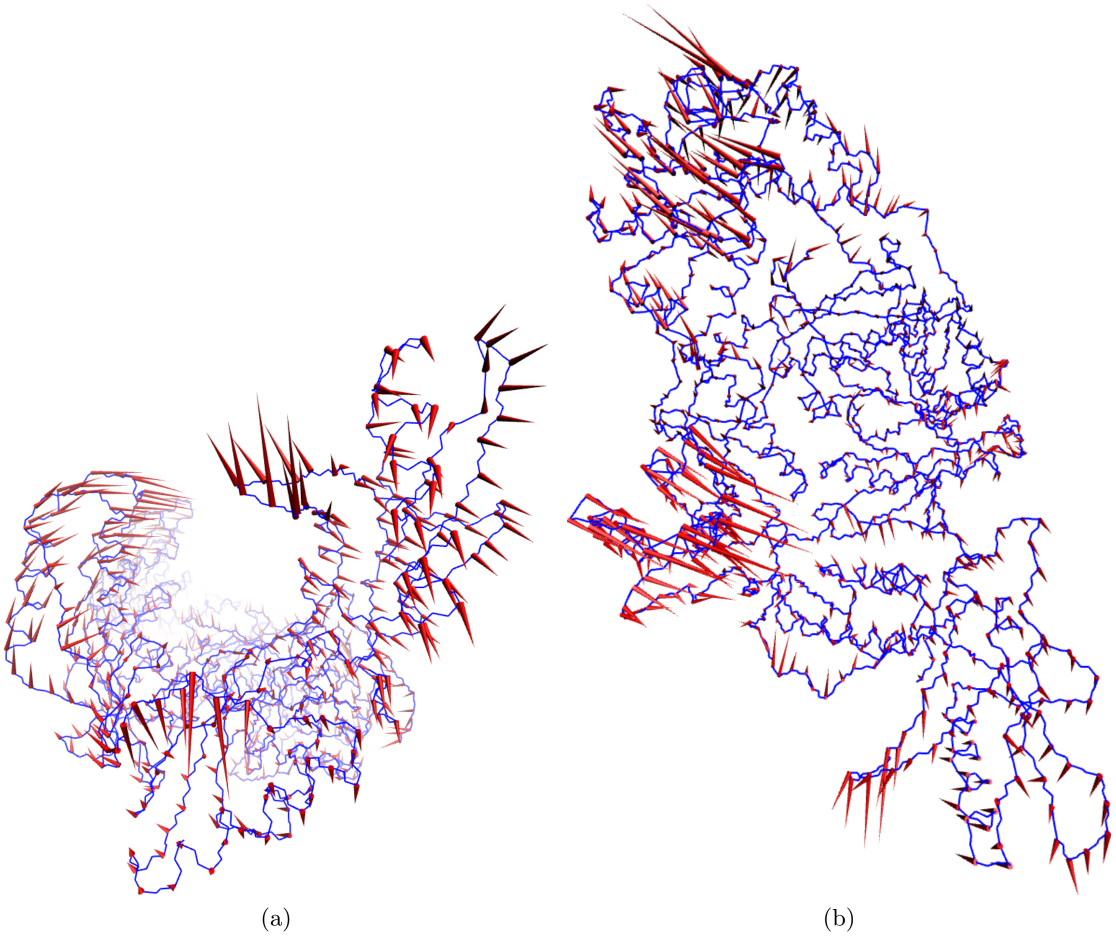


Figure 7.13: Porcupine plot showing the RT structural variation described by PC4. The structure in blue represents the most negative projection along the PC obtained from the concatenated trajectory of all open conformation systems (OPN, DNA, RMD, EFZ and NVP) simulated. The red cones show the direction of motion of the C_α atoms of the structure along the PC with the point representing the position in the most positive projection observed. The structure is shown (a) looking along the binding cleft from the polymerase active site to the RNaseH and b) looking down from a point beyond the p66 fingers.

cleft.

The fact that so many of the PCs impact upon the same areas make it hard to establish whether the less significant modes represent useful information about protein conformational changes or are simply artifacts of the PCA method. Consequently it is more instructive to concentrate on metrics suggested by the broad patterns observed in the motions they highlight.

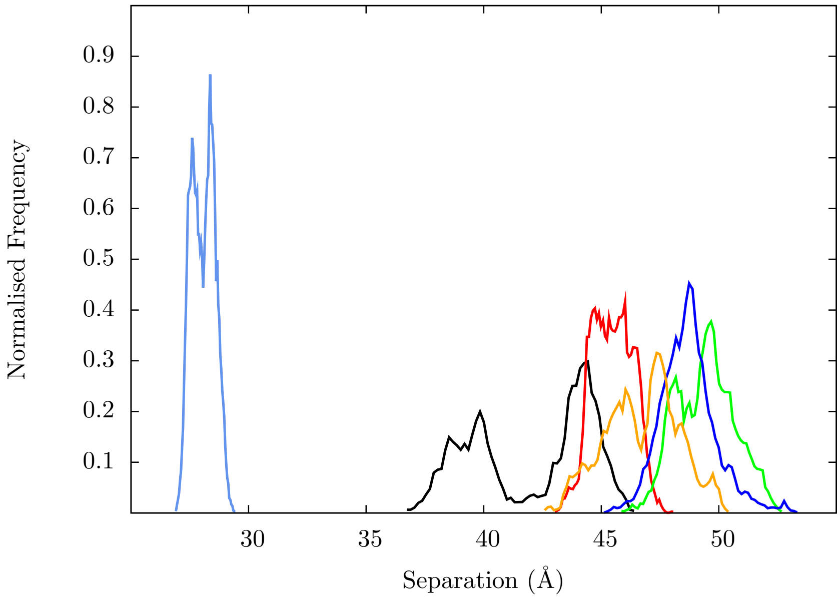


Figure 7.14: Normalised histogram of the frequency of observation of measurements of the separation of the centre of mass of the p66 fingers and thumb subdomains. Values for the CSD system are shown in light blue, OPN in black, DNA in red, RMD in green, EFZ in dark blue and NVP in orange.

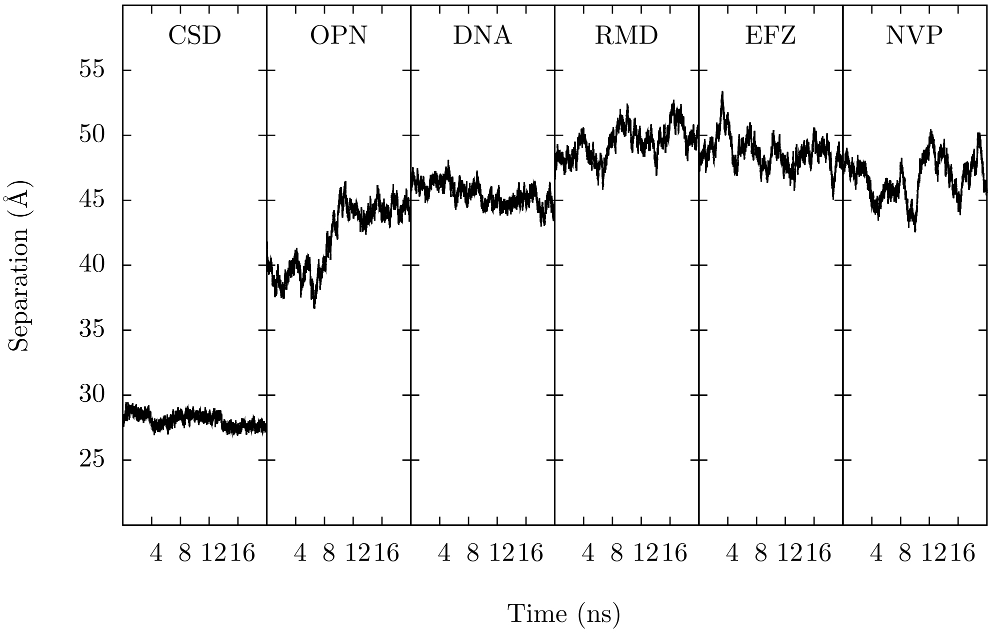


Figure 7.15: Distance between the centre of mass of the p66 fingers and thumb subdomains of all systems simulated in this study tracked over the duration of the production phase of simulation.

7.2.4.3 Characterisation of Structural Changes

Principal component analysis of the open conformation systems under study indicates that, as expected, the separation of the p66 thumb and finger domains plays a major role in differentiating the different ligation states of the HIV-1 RT. In order to assess the motions undergone by each system (including CSD) throughout the simulations the distance between the centre of mass of the p66 fingers and thumb subdomains was measured. Figure 7.14 shows the distribution of distances seen in each system. The large difference between the closed and open conformations is maintained throughout the simulations and, interestingly, in both the OPN and RMD systems (which contain neither drug nor substrate) the separation increases over the duration of the simulation (see Figure 7.15). The OPN and DNA bound systems originate in the same crystal structure but whilst, the DNA structure is constrained to a similar separation (of around 45 Å) over the simulation, the OPN system has drifted to a more closed conformation (with a separation of approximately 38 Å) by the start of the production phase of simulation and then shows significant flexibility by returning to a cleft width only slightly narrower than that exhibited with bound substrate. The changes observed in the OPN system are much greater than that of any other system and are indicative that the open conformation of the HIV-1 RT system is highly flexible when not bound to either natural substrate or an NNRTI. The lower peak and greater spread of the distance distribution of the NVP bound system compared to that containing EFZ can be interpreted as indicating that the former drug induces a less pronounced change in the flexibility of the protein.

The largest separation is exhibited in the RMD and EFZ systems with distribution peaks at 49.5 Å and 48.5 Å respectively. This is in line with the differences of the original crystal structure and suggests that the deformation of the structure to form the binding pocket is retained and continues to influence the conformation and dynamics of the rest of the system on the timescale of the simulations presented here, even after removal of the drug which caused them.

All of the first four principal components indicate that an important difference between the different systems resides in the conformation of the loop containing the residues K65 and R72, which form part of the dNTP binding pocket and are implicated in the binding of incoming nucleotides and hence the polymerase catalytic activity of HIV-1 RT. Experimental evidence suggest it is the polymerisation step (in which incoming dNTPs are incorporated into the nascent DNA chain) that is affected by NNRTI binding[331, 410] and deformations of the loop induced by the drugs would provide a plausible explanation of these results. The distance between the two key residues of this loop and residue 185 in the polymerase active site were measured to assess any changes caused by NNRTI

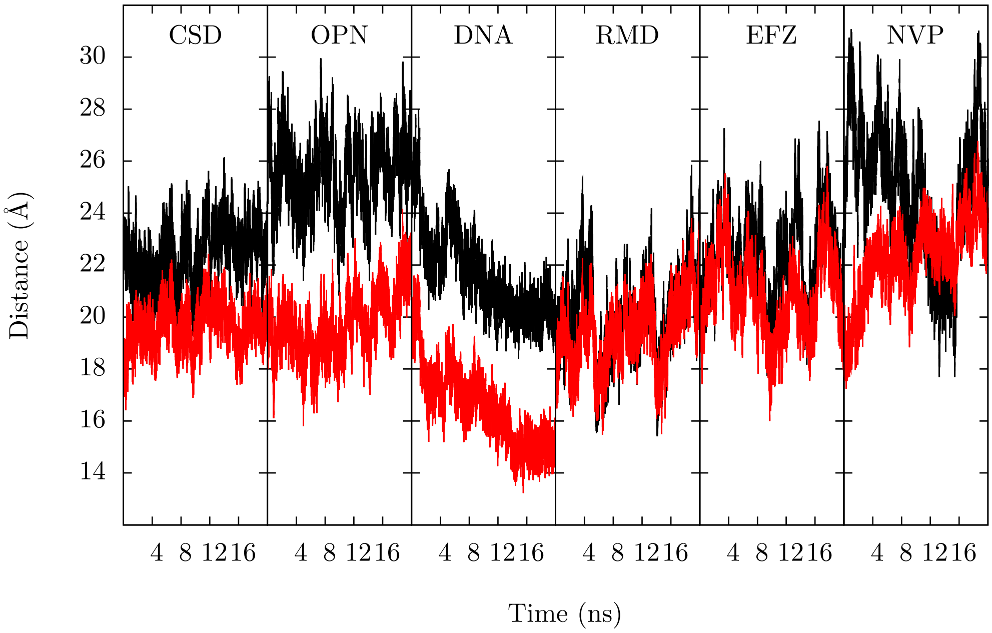


Figure 7.16: The variation of the distances between K65 (black) and R72 (red), implicated in dNTP binding and the polymerase active site residue 185 for all systems studied.

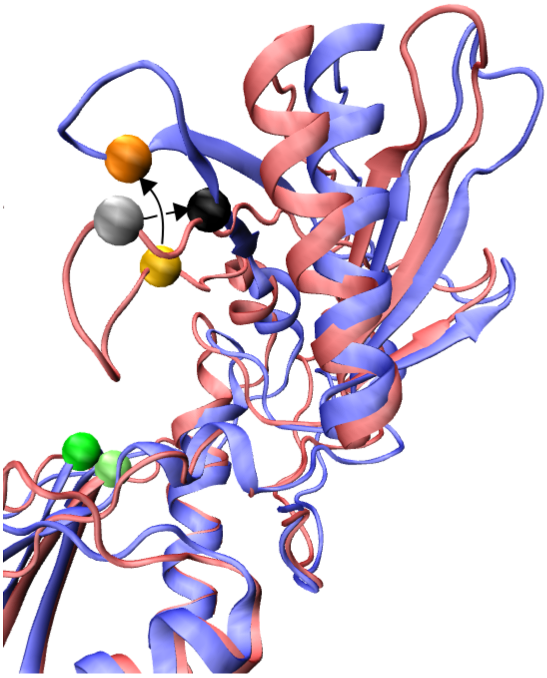


Figure 7.17: The difference in conformation of the p66 fingers in the average structures of the DNA and EFZ bound systems (shown in red and blue respectively) taken over the production phase simulation. Residues K65 (gray balls) and R72 (yellow balls) are shown in the dNTP binding loop and residue 185 is also highlighted (green balls) with a lighter shade used for the residues in the DNA bound system in all cases. The loop in the EFZ system is twisted over itself relative to the conformation seen in DNA.

binding. Figure 7.16 shows that in both apo structures and that bound to DNA the loop maintains a conformation in which residue 65 is further from the polymerase active site (by approximately 6 Å in the case of both open structures and 2 Å for the CSD system) in contrast both the EFZ and RMD systems have the two at an equal distance. This change represents a twisting of the loop as illustrated in Figure 7.17. The NVP system has a greater variability of behaviour, but in the second half of the production phase the separation is similar to that observed in the other systems containing the NNRTI binding pocket. These results seem to indicate that the formation of the NNRTI binding pocket has an impact on the conformations explored by the p66 fingers. A relationship between the two regions is also suggested by the coincidence of entry of water into the binding pocket of the NVP system and conformational rearrangements during equilibration.

In a study by Ivetac & McCammon [311] a similar set of simulations were run for a NVP bound system, a system with NVP removed, and the closed apo structure. Their work ran each system with 4 copies each producing 30 ns of simulation. They found that half of the copies of the closed system open to a similar binding cleft width as that in DNA bound crystal structures and that one of the four simulations of the system with NVP removed explored a similar conformation after 10 ns of simulation. No details of the fingers conformation are provided in their paper. They conclude that NNRTIs function by operating as a wedge that prevents the motion of a hinge between the the p66 thumb and surrounding palm residues and the fingers and surrounding palm residues, located with the pivot adjacent to the NNRTI binding pocket.

The largest stable changes in p66 thumb to finger distances observed in both the simulations by Ivetac & McCammon [311] and the simulations presented here are of the order of 5 Å and occur rarely (seen in the OPN system alone here and in 3 out of 8 non NNRTI bound systems in the ensembles of Ivetac & McCammon [311]). This is smaller than the variability of approximately 8 Å seen in the NVP bound system here. This would suggest that to confirm any hinge hypothesis considerably longer simulation lengths are needed. This is especially true as the direct structural rearrangement of the polymerase active site induced by NNRTI binding (see Section 4.7 and Figure 7.10) is hard to deconvolve from dynamic effects which may also impact upon polymerase function. Equally, confirmation of the impact of the conformational changes undergone in the p66 finger domains would require greater sampling.

7.2.5 Binding Affinity Comparison of NVP and EFZ

Along with investigating the conformational impact of NNRTI binding, this study aims to determine whether such changes impact upon the binding free energies of the drugs.

Table 7.3: Experimental differences in binding affinity values for the NNRTIs NVP and EFZ to the HXB2 wildtype HIV-1 RT sequence. In each case the results were originally presented as K_i values (or IC_{50} values which can be used as an approximation) and converted into binding free energy differences using the relation $\Delta G = RT\ln(K_i)$.

Experiment	$\mathbf{\Delta\Delta G_{expt}^{nvp-efz}}$
Butini *et al.* [94]	1.53
Högberg *et al.* [95]	2.23
Lindberg *et al.* [93]	2.50
Monforte *et al.* [96]	2.26

Energies are in kcal mol^{-1}.

As described in Section 7.2.2.3, our aim is to evaluate the possibility of adapting the MMPBSA protocols we have developed for HIV-1 protease (applications of which are described in Chapter 5 and Chapter 6) to be reliably applied to NNRTI binding to HIV-1 RT. Without calculating an estimate of the entropic component of the binding affinity it is not possible to reproduce experimental values for the absolute binding free energy change, ΔG. It is, however, possible to compare the relative binding affinities of the EFZ and NVP drugs, $\Delta\Delta G$. A selection of estimates of $\Delta\Delta G^{nvp-efz}$ from a variety of experiments are shown in Table 7.3, the average value is 2.13 kcal mol^{-1}. The impact of the water molecule observed entering the NNRTI binding pocket can also be investigated using this method, by including it as part of the receptor in the MMPBSA calculation. Over the course of the simulation, the water molecule in this position exchanges with those in the free solvent on a number of occasions and for each snapshot analysed only the closest water molecule to the NVP molecule was included in the computation.

The binding affinity computed for the 20 ns of production simulation of both EFZ and NVP is shown in Table 7.4. The EFZ drug is correctly observed to bind more tightly than NVP. However, the difference of binding affinity between the two inhibitors is much higher than the average experimental value of 2.13 kcal mol^{-1}. Including the water molecule reduces the $\Delta\Delta G_{MMPBSA}$ between the two NNRTIs from 8.28 to 6.14 kcal mol^{-1}. Accounting for the water molecule within the calculation as expected increases the binding affinity. The change of -2.14 kcal mol^{-1} is considerably less than the -5.45 kcal mol^{-1} obtained by Treesuwan & Hannongbua [402] in a study which calculated the difference from MD simulations based on a truncated system containing only the protein and water molecules within 10 Å of the NVP drug. In addition to this truncation the simulations performed by Treesuwan & Hannongbua [402] were 'equilibrated' (using the entire enzyme) for only 1 ns and produced just 3 ns of production dynamics, even the total duration of which would come within the equilibration phase of the simulations presented here. The analysis presented in Section 7.2.3.1 indicates that these simulations are likely to still be relaxing from the system setup and consequently it is possible that

Table 7.4: Computed free energy differences of binding (ΔG_{MMPBSA}) form MMPBSA analysis of NVP and EFZ bound to wildtype HIV-1 RT using 20 ns single simulations trajectories. Values for NVP are shown with a bridging water molecule included as part of the receptor (NVP WAT) and without.

Ligand	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{\Delta\Delta G^{nvp-efz}_{MMPBSA}}$
EFZ	-35.00 (0.06)	-
NVP	-26.72 (0.07)	8.28 (0.13)
NVP WAT	-28.86 (0.07)	6.14 (0.13)

Mean energies are in kcal mol^{-1}.
Standard errors are shown in parentheses.

the discrepancy in results with the simulations presented here is caused by insufficient equilibration in the earlier study.

A plausible explanation for the exaggerated difference between the binding affinities of NVP and EFZ obtained from these simulations is that the EFZ encounters a higher entropic barrier to binding. Evidence that this is a reasonable contention is provided by the observation, in Section 7.2.4.3, that the NVP system shows greater changes in the p66 fingers to thumb distance which is posited as the motion most impacted by NNRTI binding.

Table 7.5 shows the thermodynamic decomposition of the binding affinity. The binding of both NNRTIs is driven by strong attractive van der Waals interactions, as would be expected for drugs binding into a largely hydrophobic pocket. The difference between the NVP results including and excluding the bridging water molecule are largely due to a 6.97 kcal mol^{-1} more attractive electrostatic component, which more than compensates for the 0.79kcal mol^{-1} loss of attraction in the van der Waals component and 3.88 kcal mol^{-1} increase in the polar solvation penalty. The difference between NVP and EFZ is due to a reduced polar solvation penalty (4.97 kcal mol^{-1} compared to the NVP value computed including bridging water) and more attractive coulomb attraction (by 1.31 kcal mol^{-1}).

The binding affinities, ΔG_{MMPBSA}, of both systems remain stable throughout the 20 ns of production simulation as shown in Figure 7.18. The lack of changes in the binding affinities indicate that sampling of the free energy landscape using MMPBSA is largely unaffected by the domain motions observed in Section 7.2.4.3. In all cases fluctuations are present in both the polar solvation and electrostatic terms but, in general, they compensate one another, resulting in minimal changes in ΔG_{MMPBSA}. Significant deviations in these terms are seen for the NVP system when the bridging water is included in the calculation (Figure 7.18c) after 7 ns and 15 ns. These changes are associated with the exchange of water molecules between the solvent and NNRTI binding pocket. Even the

Table 7.5: Decomposed contributions to the free energy of binding for NVP and EFZ bound to wildtype HIV-1 RT using 20 ns single simulations trajectories based on the MMPBSA estimation method. Values for NVP are shown both with a bridging water molecule included as part of the receptor (NVP WAT) and without.

Sequence	$\mathbf{\Delta G^{MM}_{vdw}}$	$\mathbf{\Delta G^{MM}_{ele}}$	$\mathbf{\Delta G^{sol}_{pol}}$	$\mathbf{\Delta G^{sol}_{nonpol}}$	$\mathbf{\Delta G^{tot}_{ele}}$	$\mathbf{\Delta G_{MMPBSA}}$
EFZ	-41.21 (0.04)	-12.69 (0.08)	23.64 (0.07)	-4.74 (0.00)	10.95 (0.06)	-35.00 (0.06)
NVP	-42.26 (0.04)	-4.51 (0.06)	24.73 (0.07)	-4.69 (0.00)	20.22 (0.07)	-26.72 (0.07)
NVP WAT	-41.47 (0.05)	-11.38 (0.07)	28.61 (0.07)	-4.62 (0.00)	17.23 (0.07)	-28.86 (0.07)

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

swapping of water molecules included in the calculation does not result in any notable change in ΔG_{MMPBSA}.

7.3 Free Energy Calculations For Drug Resistant HIV-1 RT Mutants

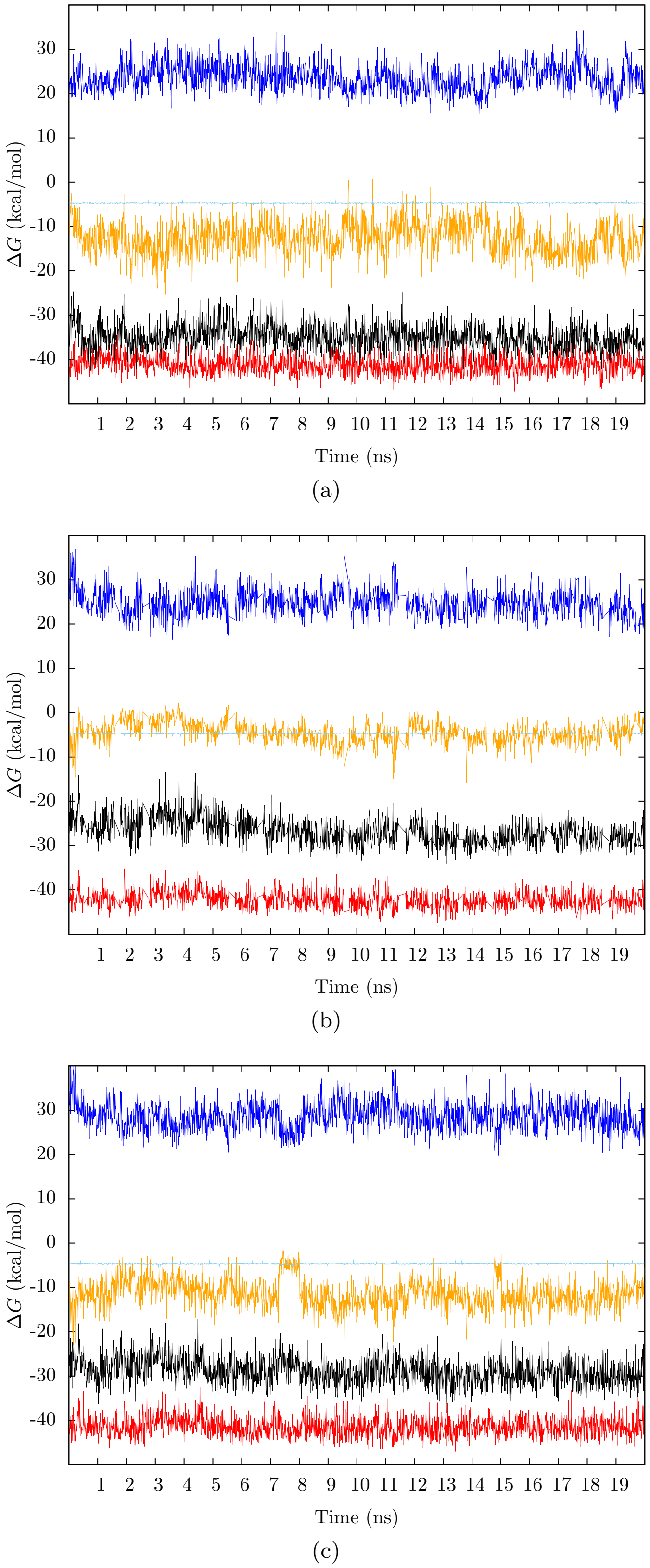
In order to assess the ability of MMPBSA to discriminate between more subtle changes than those between the two drugs, NVP and EFZ, binding to the HIV-1 RT, a series of resistant mutants have been simulated. The EFZ inhibitor was chosen to avoid the complication of the bridging water molecule which enters the NNRTI binding pocket alongside NVP. The most common mutation pathway for EFZ is that including K103N and L100I [200, 411] and these are the mutations which are studied here. Commonly the K103N mutation occurs first with L100I occurring as a secondary mutation [412]. Estimates of the binding affinities, obtained from a variety of experiments is shown in Table 7.6 indicating that K103N and L100I make similar levels of difference to the binding affinity and that the double mutant impact is more than the addition of the two. It should be noted that these results were estimated using IC_{50} enzyme efficacy measurements as a proxy for K_i and should consequently be viewed as approximations to the real binding affinities. The impact of K103N is sometimes thought to be caused by stabilising the closed structure of the NNRTI binding pocket [342–344]. This means that it may not be possible to detect the changes induced by the K103N single mutant via the MMPBSA methodology without using separate trajectories for the apo enzyme. In RT structures containing other inhibitors, L100I is seen to directly influence the shape of the binding pocket (and hence, presumably, the binding affinity) [413].

Table 7.6: Experimental differences in binding affinity values for the L100I, K103N and L100I-K103N mutants compared to the wildtype HIV-1 RT. The experimental values come from 1 Bacheler *et al.* [414], 2 Soriano & de Mendoza [415] and 3 Silvestri & Maga [416]. In each case the results were originally presented as K_i values (or IC_{50} values which can be used as an approximation) and converted into binding free energy differences using the relation $\Delta G = RT\ln(K_i)$.

Sequence	**$\Delta\Delta G_1$**	**$\Delta\Delta G_2$**	**$\Delta\Delta G_3$**
L100I	1.88	2.39	1.80
K103N	2.12	2.28	1.95
L100I-K103N	4.61	4.91	4.33

Energies are in kcal mol^{-1}.

This study focuses on assessing the use of a single trajectory to represent complex, inhibitor and apo receptor. The intention is to investigate both whether the changes binding affinity produced by the mutations L100I and K103N can be detected by MMPBSA



(a)

(b)

(c)

Figure 7.18: Binding free energy, ΔG_{MMPBSA}, and component values tracked for each snapshot throughout the 20 ns of production trajectory for (a) EFZ, (b) NVP and (c) NVP with bridging water molecule. The total, ΔG_{MMPBSA}, is shown in black with the components; ΔG^{MM}_{vdw} in red, ΔG^{MM}_{ele} in orange, ΔG^{sol}_{pol} in blue and ΔG^{sol}_{nonpol} ΔG^{tot}_{ele} in light blue.

and evaluating the performance of single trajectory and ensemble approaches to sample relevant conformations of the HIV-1 RT. Four systems are simulated the HXB2 wildtype (labelled WT) and the L100I and K103N single mutants and the L100I-K103N double mutant.

7.3.1 Methods and Analysis

All systems described in this section were based on the 1IKW crystal structure [93]. The sequence of this structure is that of the HXB2 wildtype and all residue substitutions described in this section are made with reference to this baseline. As in the HIV-1 protease case (see Chapter 5), mutations were inserted into the model using the VMD [315] visualisation package via the BAC [355] automation scripts. The simulation, equilibration and analysis protocols used for these simulations are identical to those described earlier for the simulations described in the Section 7.2. The single trajectory strategy is represented here by runs with production phases of 20 ns (labelled 1 × 20 ns) and an equivalent length of trajectory was generate using an ensemble of 5 replicas with 4 ns generating 4 ns of production trajectory (labelled 5 × 4 ns).

Assessment of convergence was performed using two primary methods, the assessment of the extent to which the data sets could be described as having a Gaussian distribution and the root mean squared difference between forward and reverse cumulative means, σ_{MMPBSA}, as described in Section 5.2.4.

7.3.2 Comparison of Binding Affinities From Single Trajectory and Ensemble Simulation Strategies

Table 7.7 shows the calculated binding affinities, ΔG_{MMPBSA}, for each of the four sequences under investigation using both the single trajectory, 1 × 20 ns, and ensemble, 5 × 4 ns, strategies along with the thermodynamic decomposition. The results from neither strategy reproduces the experimental trend.In the single trajectory results both the L100I and K103N single mutants are counted as resistant with relative binding affinities to wildtype, $\Delta\Delta G_{MMPBSA}$, values of 3.18 and 1.08 kcal mol^{-1}respectively. The double mutants system is however assessed as binding more tightly than the WT by 0.91 kcal mol^{-1}. The ΔG_{MMPBSA} value of the WT system is altered by only 0.36 kcal mol^{-1}between the the two simulation strategies but the overall ranking of mutants is very different. In the 5 × 4 ns data set the double mutant is ranked as highly resistant with a $\Delta\Delta G_{MMPBSA}$ of 2.40 kcal mol^{-1} but the K103N mutant is found to bind more tightly than wildtype and the L100I is observed to be only marginally resistant.

The thermodynamic decomposition shown in Table 7.7 does not present any single component as being particularly unreliable, with all measurements having similar levels of error associated with them. However, the systems showing the highest level of resistance in each strategy (L100I for 1 × 20 ns and L100I-L103N for 5 × 4 ns) both show substantially less attractive ΔG_{ele}^{MM}. The trajectories, however, show no unambiguous structural change that correlates with this. Furthermore, the only hydrogen bond (defined as a potential donor and receptor atom pair within 3.5Å of one another with a donor-hydrogen-receptor angle of less than 120°) present in more than 5% of simulation snapshots is between the backbone oxygen of residue 101 and the EFZ nitrogen and is found in 99% of snapshots in all of the simulations of all of the sequences. The loss of electrostatic attraction and the lowering of the polar solvation penalty compared to WT are found to be common features of all the mutant systems, using both simulation strategies. The balance of these two effects is predominantly responsible for the $\Delta\Delta G_{MMPBSA}$ obtained for each system.

7.3.2.1 Evaluation of the Free Energy Sampling and Convergence

The ΔG_{MMPBSA}values calculated here have limited success in reproducing the expected experimental trends. There are a number of possible explanations for this, including the failure to account for the energy cost of binding pocket formation, the failure of the MMPBSA methodology in this system and lack of sufficient sampling. The large variations seen between the replicas, in the double mutant system in particular, suggest that the last of these is at least plausible. The distribution of calculated ΔG_{MMPBSA}values in a well sampled system should be Gaussian. The real distribution for the values sampled in both the 1 × 20 ns and 5 × 4 ns simulations are shown in Figure 7.19. Whilst the results for the ensemble simulations are closer to replicating the correct distribution than those from the single trajectories significant deviations exist for all systems in both approaches. This observation would tend to suggest that we have not, as yet, obtained sufficient sampling to gain correctly converged results.

The RMS difference in the forward and reverse cumulative means, σ_{MMPBSA}, provides a metric indicating the level of variance in the average which is encountered as more snapshots are taken into account. Figure 7.20 shows a comparison of the difference in forward and reverse cumulative means, σ_{MMPBSA} (as defined in Section 5.2.4.3), for all four sequences under study. The only sequence using either ensemble or single trajectory strategies in which σ_{MMPBSA} fails to converges to below 2 kcal mol^{-1}is L100I where the 5 × 4 ns values plateaus after approximately 2 ns of sampling to around 2.5 kcal mol^{-1}. The convergence of both strategies in the case of the L100I-K103N double mutant is surprising considering the large discrepancy between the ΔG_{MMPBSA} values obtained by

Table 7.7: Decomposed contributions to the free energy of binding for for four HIV-1 RT sequences under investigation from both 1 × 20 ns single-trajectory and 5 × 4 ns ensemble strategies. The relative free energy difference between each mutant and the WT system, $\Delta\Delta G_{MMPBSA}$, is also shown.

Sequence	$\mathbf{\Delta G^{MM}_{vdw}}$	$\mathbf{\Delta G^{MM}_{ele}}$	$\mathbf{\Delta G^{sol}_{pol}}$	$\mathbf{\Delta G^{sol}_{nonpol}}$	$\mathbf{\Delta G^{tot}_{ele}}$	$\mathbf{\Delta G_{MMPBSA}}$	$\mathbf{\Delta\Delta G_{MMPBSA}}$
Single Trajectory (1 × 20 ns)							
WT	-41.21 (0.04)	-12.69 (0.08)	23.64 (0.07)	-4.74 (0.00)	10.95 (0.06)	-35.00 (0.06)	-
L100I	-40.68 (0.04)	-9.96 (0.08)	23.56 (0.07)	-4.74 (0.00)	13.60 (0.06)	-31.82 (0.07)	3.18 (0.13)
K103N	-41.04 (0.05)	-11.31 (0.07)	23.11 (0.07)	-4.69 (0.00)	11.80 (0.07)	-33.92 (0.08)	1.08 (0.14)
L100I-K103N	-41.20 (0.05)	-11.63 (0.08)	21.58 (0.07)	-4.65 (0.00)	9.94 (0.06)	-35.91 (0.07)	-0.91 (0.13)
Ensemble (5 × 4 ns)							
WT	-40.99 (0.05)	-12.32 (0.09)	23.38 (0.07)	-4.71 (0.00)	11.06 (0.07)	-34.64 (0.07)	-
L100I	-40.75 (0.04)	-10.57 (0.08)	21.68 (0.06)	-4.69 (0.00)	11.10 (0.06)	-34.34 (0.07)	0.30 (0.14)
K103N	-41.82 (0.05)	-10.37 (0.07)	21.04 (0.07)	-4.65 (0.00)	10.67 (0.07)	-35.80 (0.07)	-1.16 (0.14)
L100I-K103N	-40.38 (0.04)	-7.73 (0.06)	20.61 (0.07)	-4.74 (0.00)	12.88 (0.06)	-32.24 (0.07)	2.40 (0.14)

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

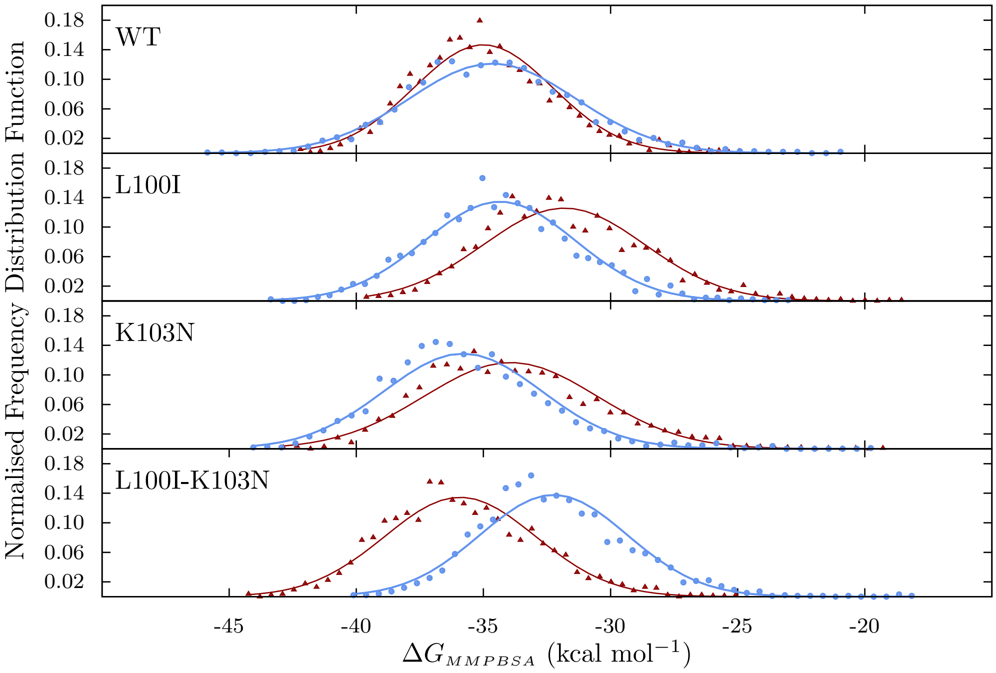


Figure 7.19: Normalised frequency distribution analysis of the MMPBSA derived binding free energy, ΔG_{MMPBSA}, for the 1 × 20 ns (red triangles) and 50 × 1 ns trajectories (blue circles) for each of the WT, L100I, K103N and L100I-K103N reverse transcriptase sequences. The expected normal distribution given the same mean and standard deviation for each data set is shown by the red and blue lines, respectively.

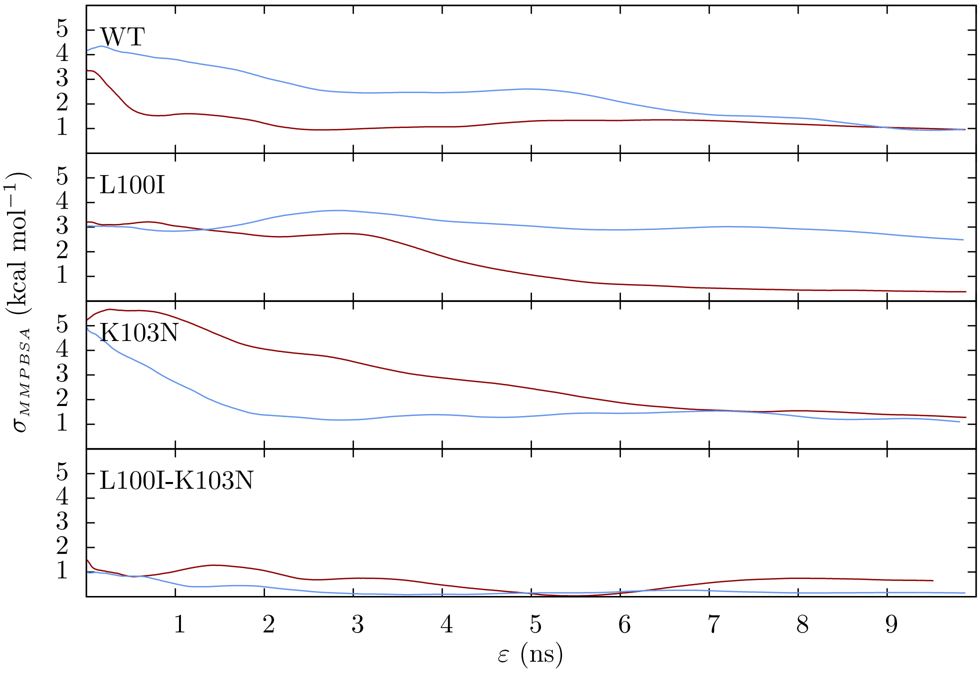


Figure 7.20: Comparison of the difference in forward and reverse cumulative means, σ_{MMPBSA}, as a function of snapshot, ε, for each of the four HIV-1 RT sequences under investigation. Values for the 1 × 20 ns, single trajectory computations are shown in red, those for the 5 × 4 ns ensemble in blue.

Table 7.8: Free energy differences of binding calculated using MMPBSA ,ΔG_{MMPBSA}, from each of the replicas in the 5 × 4 ns ensemble strategy for four HIV-1 RT sequences under investigation, showing the high variability between the replicas.

Sequence	$\mathbf{\Delta G^{rep1}_{MMPBSA}}$	$\mathbf{\Delta G^{rep2}_{MMPBSA}}$	$\mathbf{\Delta G^{rep3}_{MMPBSA}}$	$\mathbf{\Delta G^{rep4}_{MMPBSA}}$	$\mathbf{\Delta G^{rep5}_{MMPBSA}}$
WT	-35.14 (0.14)	-35.14 (0.14)	-35.52 (0.16)	-34.77 (0.16)	-32.63 (0.14)
L100I	-33.23 (0.13)	-32.47 (0.13)	-34.17 (0.13)	-35.13 (0.12)	-36.68 (0.12)
K103N	-34.97 (0.15)	-34.40 (0.14)	-35.11 (0.18)	-37.26 (0.13)	-36.27 (0.12)
L100I-K103N	-31.11 (0.13)	-33.02 (0.12)	-33.60 (0.11)	-33.43 (0.18)	-31.03 (0.16)

Mean energies are in kcal mol^{-1}. Standard errors are shown in parentheses.

the different strategies. The small value of σ_{MMPBSA} in the 1 × 20 ns case is indicative of the fact that the measured value does not drift over the course of the simulation and hence that long timescale motions are unlikely to be responsible for the discrepancy between the results of the two strategies. A possible explanation for these observations would lie in the idea that the energy landscape is such that in most cases, on the timescales accessed by these simulations, only a single minimum is explored and that the minimum sampled by the single trajectory for L100I-K103N happens to be very different to those explored by the replicas in the ensemble. If this is the case then significant variation should be seen between the averages of each replica within each ensemble. The mean values of ΔG_{MMPBSA} for each replica in the ensemble for each sequences is shown in Table 7.8. For each sequence there are replicas which vary by at least 2 kcal mol^{-1} which is less than the 3.67 kcal mol^{-1} difference between the single trajectory and ensemble results for the double mutant system but suggests that a variety of energy wells are indeed being sampled by each ensemble.

The differences between replicas and comparative stability of the averages obtained from single simulations suggest that, just as we found in the protease case, reverse transcriptase simulations with a length scale of the order of tens of nanoseconds are unlikely to be reproducible. The sampling presented here is far less extensive than that obtained for the HIV-1 PR case in Chapter 5. The results here suggest that we are seeing similar levels of variability between replicas to that in the protease system and consequently that approximately 50 replicas would be required in order to obtain reproducible results. Results obtained within our group have also shown that similar levels of sampling are required in the case of the cancer drug target epidermal growth factor receptor (EGFR) [417].

The σ_{MMPBSA} cannot be invoked to give any indication of the level of extra sampling that might be required to access areas of phase space not seen within the simulations it is used to analyse. It does, however, suggest a period of time beyond which sampling in a single trajectory is unlikely to provide extra information unless considerably extended time periods can be simulated, thereby increasing the chances of rare events that shift the system to a new area of phase space. In the case of the results presented here, the plateauing of σ_{MMPBSA}, seen for all sequences from between 4 and 6 ns in Figure 7.20, would indicate that this is a sensible length to adopt for replica simulations within an ensemble.

7.4 Conclusions

In the first part of this study, different HIV-1 RT systems representing the unliganded, NNRTI bound and natural double stranded DNA substrate bound enzyme were simulated. Principal component analysis of the conformers explored by the equilibrated systems indicate that major differences in the systems under study here reside in the conformation of the p66 thumb and fingers. Both of these regions are seen to be flexible in all these systems, although different conformations are available in different ligation states. In particular, the loop containing residues K65 and R72, which forms part of the dNTP binding pocket implicated in the binding of incoming nucleotides, is seen to exhibit different behaviour between the NNRTI bound and non-inhibited systems. In the drug bound systems the loop is seen to twist over itself relative to the configuration seen in the apo and DNA bound systems. This observation is consistent with experimental findings, according to which it is the DNA polymerisation step that is affected by NNRTI binding [331, 410]. The comparative importance of this indirect effect and other explanations for NNRTI efficacy, such as the distortion of the polymerase active site and surrounding residues, caused by the formation of the NNRTI binding pocket and seen in the crystal structures [93, 404], remains unclear.

During the equilibration phase of the NVP bound system, an entrance pathway via which water molecules enter the NNRTI binding pocket is observed. The water enters through a channel which opens between residues V106, P225, F227, H235 and P236, a locus which has been suggested as the entrance taken by NNRTIs to the binding pocket [336]. The ingress of this water molecule to the binding pocket of the NVP bound reverse transcriptase is coincident with several conformational changes, although caution must be applied when ascribing a causal relationship as water entry initially occurs as restraints on the system are being released.

Binding affinities were calculated for both the NNRTIs (EFZ and NVP) over the 20 ns of production simulation using the MMPBSA methodology. In the case of NVP the influence of the water which entered the binding pocket was also assessed. EFZ is correctly observed to bind more tightly than NVP whether or not the water molecule is included in the calculation. The difference of binding affinity between the two inhibitors is, however, higher than the average experimental value of 2.13 kcal mol^{-1}. The inclusion of the water molecule reduces the $\Delta\Delta G_{MMPBSA}$ between the two NNRTIs from 8.28 to 6.14 kcal mol^{-1}. A plausible explanation for the exaggeration of this difference is suggested by the extra mobility seen in the thumb and fingers in the EFZ system suggesting that it may face a lower entropic barrier to binding. The binding affinities calculated for both NNRTIs remain stable throughout the 20 ns of trajectory they are calculated over. The lack of changes in the binding affinities indicate that sampling of the free energy landscape is largely unaffected by the domain motions observed in the protein. This confirms the idea that meaningful free energy values can be computed from trajectories on the time scales explored here.

The second part of this study investigated the plausibility of using the MMPBSA protocol, automated within BAC, to assess the resistance levels of HIV-1 RT sequences by calculating the impact of the mutants L100I, K103N and L100I-K103N upon the calculated binding affinity, ΔG_{MMPBSA}. These mutations are all seen experimentally to lead to resistance, with the double mutant showing a greater than additive loss in binding efficacy compared to the two single mutants [414–416]. The simulation results presented here are unable to correctly reproduce this ranking. One potential explanation of this is that the sampling available within these simulations is insufficient to obtain correctly converged results. Convergence analysis of the ΔG_{MMPBSA} values obtained in this study suggest that neither simulation strategy obtains reliably converged results from 20 ns of sampling. The variation seen between the replicas in the ensemble approach and the stable values observed in most of the single trajectory mutant systems suggest that an extension of the ensemble strategy is likely to be the most effective way of accessing the necessary regions of phase space in order to obtain the required sampling. Analysis of the cumulative means for the single trajectory strategy suggests that replica lengths of between 4 and 6 ns sufficiently sample the local minima. Another possibility, suggested by other studies [342–344], is that an important component of K103N related resistance comes from the stabilisation of the apo RT structure without the NNRTI binding pocket. Were this to be true, the MMPBSA methodology used in this study would have to be amended to use separate trajectories for the unbound receptor, drug bound complex and NNRTI under investigation. A third possibility is that conformational entropy, neglected in the calculations presented here, has a significant effect on the relative binding free energies of the different systems. Currently, the ability to replicate fully the MMPBSA and normal mode free energy analysis used is limited by the fact that none of the available software solutions for calculating normal modes can accommodate the memory requirements made by a system as large as the HIV-1 RT. Results, presented elsewhere, in which a truncated HIV-1 RT structure have been analysed in this way have failed to alter the ranking of systems from that given by MMPBSA and are likely to have excluded important modes that contribute to the entropic barrier to drug binding [401].

Overall, our work in this chapter shows that differences in binding affinity produced by gross changes such as altering the inhibitor considered can be identified using MMPBSA in the case of NNRTIs binding to HIV-1 RT. The question of whether the same can be done for subtle shifts such as those introduced by point mutations remains open. In order to address this question, future studies will have to investigate extensions to the free energy methodology used, in particular finding ways in which to assess the contributions of binding pocket formation and entropy. The results presented here also suggest that ensembles of simulations might be the most efficient way to enhance the sampling of the available phase space compared to simply extending the time for which simulations are run. This is in line with the findings presented in Chapter 5 and work on EGFR [417], which both indicate that tens of replicas are required in order to produce converged, reproducible, free energy values using MMPBSA.

Chapter 8

Conclusions

In this thesis, fully atomistic molecular dynamics (MD) simulations have been used to investigate the binding of anti-retroviral drugs to target enzymes in HIV. In particular, changes to both structure and thermodynamic properties in the drug target enzymes protease and reverse transcriptase caused by mutations associated with drug resistance have been elucidated. The development of such mutations in response to therapy is well known, and represents the main obstacle to ultimate treatment success. The primary cause of drug resistance is the lowering of the binding affinity between drug and target protein. Whilst experimental techniques exist that measure this quantity, they cannot provide detailed molecular insight into the causes of resistance.

The sampling required to obtain accurate, converged binding affinities for a series of multi-drug resistant (MDR) protease mutants bound to the inhibitor lopinavir was investigated. An 'approximate' free energy method was applied to calculate the absolute and relative binding free energies of these systems and a comparison of the sampling achieved by ensembles of short simulations and longer single trajectories was evaluated. Only the ensemble method was shown to achieve correctly distributed and converged sampling of conformational microstates, implying that they explore conformational space more effectively than single long time-scale simulations. Using the ensemble methodology (with 50 replica simulations producing 4 nanoseconds of production trajectory each) a completely correct ranking for six HIV-1 protease variants was obtained, with a correlation coefficient of 0.89 and a mean relative deviation from experiment of 0.9 kcal mol^{-1}. The issue of reproducibility, crucial for the validity of any free energy calculation, is tied to the convergence of such calculations. The results presented in this thesis suggest that individual MD simulations are unlikely to be reproducible and caution must be exercised when considering properties extracted from single trajectory calculations. Further weight is given to these conclusions by preliminary calculations performed on reverse transcriptase bound to the inhibitor efavirenz, where again multiple copy simulations are seen to sample varying areas of phase space and improve the convergence of free energy calculations.

In clinical settings, the interpretation of the level of drug resistance is generally performed using mutation lists and rules-based algorithms embedded in so-called clinical decision support systems (CDSS)[1]. The EU funded 6th Framework Project (FP6) ViroLab[2] sought to enhance such systems by incorporating a wealth of tools for investigating the relationship between HIV genomic sequence and the level of resistance to anti retroviral drugs[382]. In order to demonstrate the potential of integrating diverse systems such as traditional drug ranking systems, literature mining, patient data and MD into a single interface the Virtual Patient Experiment (VPE) was designed. The VPE identified a patient sequence for which a range of CDSS gave differing resistance assessments, identifying three mutations within the sequence as being associated with resistance (L10I, A71I/V and L90M). Applying the simulation and free energy protocol, validated by the study on MDR protease systems, predictions of the resistance level of the identified mutations singly, in combination, and as part of the full patient sequence were generated. Only in the context of the full patient sequence was any reduction in binding affinity observed. The variants containing resistant associated mutants here adopt substantially different conformations to those of the MDR proteases previously investigated. In all resistant protease mutants studied in this thesis similar perturbation of the active site hydrogen bond network was observed and, additionally, water molecules were seen to occupy the catalytic cavity more frequently than in the wildtype. Uniquely, in the VPE mutants the overall enzyme conformation is expanded compared to wildtype with the distances between residues 35 and 45 on both monomers increased but with residue 79 on both monomers moving towards the active site.

Investigations of the application of the same simulation protocol to reverse transcriptase are still in the preliminary stages. The binding affinity of two common NNRTIs (efavirenz and nevirapine) have been successfully distinguished, but sufficient sampling to reliably evaluate resistance causing mutations has not yet been achieved. A comparison of simulations of the apo reverse transcriptase and the same enzyme bound to NNRTIs or natural DNA substrate have suggested that conformational changes, in regions of the p66 fingers domain key to catalytic function, are associated with drug binding. The importance of this change in conformation in the mechanism of inhibition of reverse transcriptase by NNRTIs compared to well established distortions of the active site and changes in enzyme flexibility is, as yet, unclear. It is hoped that the simulations

[1] Examples of CDSS include the Stanford HIVdb: `hivdb.stanford.edu`, ANRS: `www.anrs.fr` and RegaDB: `www.rega.kuleuven.be/cev/regadb/`.

[2] Virolab: `http://www.virolab.org`

presented in this thesis will provide the foundations of future work that could see us providing greater insight into drug binding in reverse transcriptase, comparable to that obtained in the protease case.

In order to envisage MD simulations becoming integrated into CDSS it is necessary that the results can be produced within a short and well defined turn around time. Given the vast number of cores on petascale supercomputers, all replicas within an ensemble simulation can, in principle, easily be run concurrently and be completed within a single day. Even accounting for the requirement of another 24 hours to perform post-processing analysis, the methodology applied in this thesis allows simulations to be turned around on clinically relevant timescales (2–3 days), opening the way for its potential use in a clinical setting to match proposed drug treatment to individual patients' genetic profiles. Such rapid turn around times are only achievable if the process of simulation creation, deployment, execution and analysis can be automated. In response to this requirement the Binding Affinity Calculator (BAC), which fully automates the free energy calculation workflow (see Appendix A), has been developed.

The approach adopted in this thesis is theoretically applicable to any system in which drugs bind to proteins and could be used to investigate the impact of mutations in a wide range of systems. Indeed, a recent study by Wan & Coveney [417] used BAC to compute binding affinities of anti-cancer drugs to genetic variants of the epidermal growth factor receptor (EGFR). This work was conducted as part of the EU ContraCancrum project[3] which aims to create a data warehouse collating data from both experimental and *in silico* sources in the hope that it may be used to inform future CDSS. Unlike the HIV case, genotypic testing is not currently standard for patients presenting with cancer, and consequently a much more limited range of data is available. Encouragingly, however, the U.K. National Health Service (NHS) has recently announced plans to deploy broad genetic testing for people with various forms of cancer, including lung carcinoma, and to implement personalised medicine based on individuals' genetic information [418]. The program will enroll up to 12,000 patients in its first phase, many more than any other current clinical trials for cancer treatment. It is to be hoped that such developments may well presage an era in which the use of genetic data to tailor treatment to individual patients, providing more reliable healthcare, becomes de rigeur. In such a future, methods which aid the interpretation of genetic data, such as MD, would become increasingly important in clinical practice.

[3] ContraCancrum: http://www.contracancrum.eu

Appendix A

Binding Affinity Calculator

Here we present a description of a tool, known as the 'Binding Affinity Calculator' (BAC), developed to automate the work flow involved in molecular dynamics based binding affinity calculations. It was originally designed for the simulations of HIV-1 protease (PR) bound to a variety of ligands [355] but has been extended for use with both the HIV-1 reverse transcriptase (RT) and human epidermal growth factor receptor (EGFR). We discuss the motivations that drove the development of the tool along with the architecture and methodology adopted within the BAC.

A.1 Motivation

The introduction of grid technology and the increasing availability of high performance computing (HPC) resources offers the opportunity to perform large numbers of CPU intensive simulations. One area in which this is particularly attractive is the study of biomolecular systems. The implementation of any such physically realistic molecular simulation, however, is a complicated and multistage process, often requiring the scientist to overcome a large manual overhead in the construction, preparation, and execution protocols needed to complete a set of simulations, not to mention any analysis protocols for determining desired properties post production. Within this context, provided a robust simulation protocol exists for the target biomolecular system, automated tools can relieve the user of the repetitive and time consuming steps involved in preparing and running simulations, freeing them to concentrate on the scientific aspects of their study. BAC was designed to be such an automation tool. Based around the simulation protocol established for the HIV-1 protease [286, 288] and later extended to the HIV-1 RT (see Chapter 7) and human epidermal growth factor receptor [419] BAC automates the various model construction, MD simulation and post-production analysis protocols, whilst requiring the specification of only a few biological input parameters.

BAC allows the execution of complete simulations with only the target protein, ligand and any mutations (relative to a designated wildtype) to be inserted, being specified by the user. The user is, however, limited to a selected range of initial crystal structures (PDBs) and pre-parameterised drugs. In the case of protease it is also possible for the user to specify a protonation state for the catalytic dyad.

A.2 BAC Workflow and Architecture

The workflow implemented by BAC incorporates model construction, simulation, and post production analyses, which are implemented by the BAC-Builder, Sim-Chain and FE-Calc applications, respectively. In this description we will assume the existence of a starting crystal structure of the complex to be simulated and that forcefield and charge parameters which describe both protein and ligand are available. In general, the preparation of the crystal structure (including the incorporation of mutations, etc.) and post production will be performed on a local cluster, whereas simulations execution takes place on a remote, perhaps Grid based, resource. In order to manage the data transfer this requires BAC is built upon the Application Hosting Environment (AHE) [385] middleware.

BAC is designed as set of modular applications, implemented in the Perl language. The overall architecture of the workflow is shown in Figure A.1. The overarching control script encountered by a command line user is called the Unit-Executor. This provides the information that the AHE requires to execute the BAC-Builder, Sim-Chain and FE-Calc applications on the designated resources and transfer the data between them at appropriate points of the work flow.

The BAC-Builder program, typically on a local resource that provides access to AMBER 9 [70] and VMD [315], builds all the pre-simulation and configuration files necessary for all equilibration and productions simulations, prior to execution of any simulation runs. An instance of the Sim-Chain application is then spawned. AHE stages all of the required flies onto an appropriate resource for simulations execution. A compiled copy of either the NAMD2 [50] or GROMACS 4[1] [72] molecular dynamics software, used by Sim-Chain, must be available on the target resource. Upon successful completion of each stage of simulation (either equilibration or production) data is staged back to a storage resource. A check is performed by the Unit-Executor to ensure each step has

[1] Topology conversion for use in GROMACS 4 is performed via the program acpypi (http://www.ohloh.net/p/acpypi).

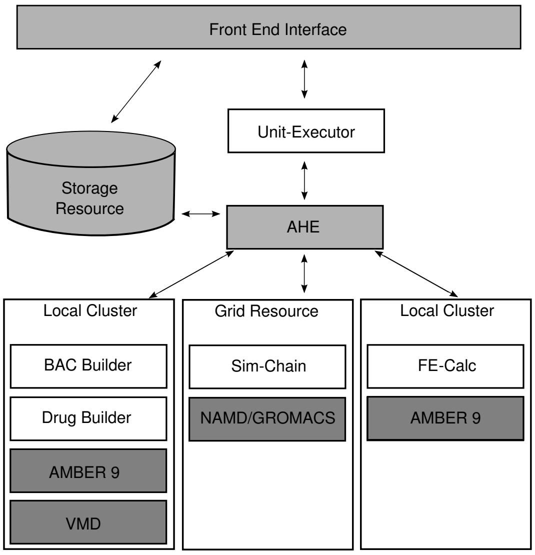


Figure A.1: Architecture of the BAC. Simulation workflow is managed by Unit-Executor, a perl script designed to utilise the application hosting environment (AHE) middleware. The components of the work flow, namely model construction, simulation, and post production analyses, are implemented by the BAC-Builder, Sim-Chain and FE-Calc applications, respectively. AHE automates the full workflow including the execution of each component and marshalling data transfer to, between and from distributed HPC resources.

successfully completed. If this check is passed the Sim-Chain program is then run again for the next stage of the simulation. Once all stages in the simulation are complete (and the data generated staged back to the storage resource) the post processing analysis can be performed. This is performed by the FE-Calc program, again typically on a local resource. Using parameters passed to it by the Unit-Executor the FE-Calc program generates input files and execution scripts for the AMBER 9 MMPBSA and normal mode analysis modules, which are used to calculate the total change in binding affinity. The generated scripts are submitted for calculation. Upon completion of the binding free energy computations AHE stages the output to the storage resource. Finally, a script is used to retrieve summary results.

In many situations, it is either necessary or desirable that parts of the workflow are executed in isolation. The module design of BAC allows all components, such as the BAC-Builder, Sim-Chain and FE-Calc applications to be used independently in scenarios when full automation is not required.

A.3 The BAC-Builder and Sim-Chain Applications

Simulation ready models of the target proteins are generated by the BAC-Builder application. BAC-Builder consists of a Perl script which requires the user to specify the

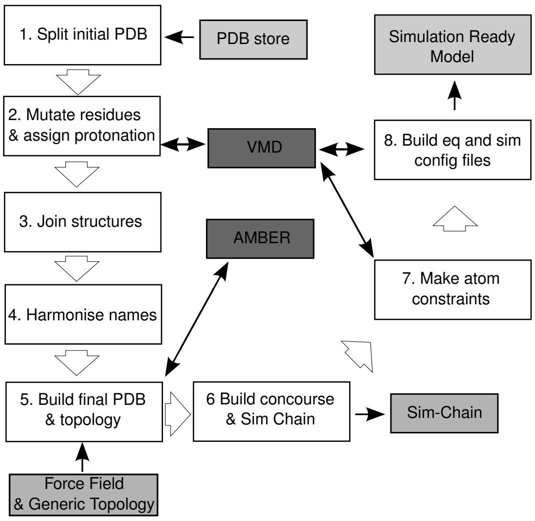

1. Split initial PDB
PDB store
Simulation Ready Model
2. Mutate residues & assign protonation
VMD
8. Build eq and sim config files
3. Join structures
AMBER
7. Make atom constraints
4. Harmonise names
5. Build final PDB & topology
6 Build concourse & Sim Chain
Sim-Chain
Force Field & Generic Topology

Figure A.2: Schematic representation of the BAC-Builder applications. The steps of the workflow (labelled 1 to 8) use a library of pre-modified PDB structures, together with standard forcefield and topology files and interface with the VMD and AMBER applications to construct the input files necessary for subsequent simulations.

forcefield, the initial pdb crystal structure, the complexed status of the protein (either ligand bound or apo) and only in the case of protease the protonation state of the catalytic dyad. Additional, optional parameters with default values may be specified, such as any desired mutations relative to the crystal structure chosen and the size of the solvation box. The input information is used to generate 'tcl' scripts which are run in VMD and input scripts executed using the 'tleap' module of AMBER 9. The locations of AMBER 9, VMD and the choice of target MD code and details of the equilibration protocol to be used are specified in a input file that is read when the script is run.

BAC-Builder contains a library of available, pre-prepared, PDB derived, crystal structures which includes 200 PR structures, 3 for RT and 2 for EGFR. In all structures the atomic nomenclature has been edited to conform to the AMBER format and chain in the PR and RT dimer chains are designated A and B sequentially. Atomic coordinates have been left unaltered. Parameterisations are available for all 9 FDA approved protease inhibitors, 3 NNRTI inhibitors and two drugs targeted at EGFR.

A schematic representation of the steps automated by the BAC-Builder is shown in Figure A.2. Initially, proteins and any solvent molecules captured crystalographically are separated into different files (protein dimers are split into two files, one for each monomeric chain). In ligand bound simulations the ligand is also placed in a separate file. If required, any mutations are inserted using VMD (in protease simulations the catalytic dyad protonation state is also set at this stage in the same manner). The separate coordinate files are merged and atomic nomenclature, which is reassigned by VMD, is reverted to conform to AMBER conventions. An input file is then generated

Table A.1: The steps involved in the BAC equilibration protocol. ^{a}M-region consists of all heavy ligand or protein atoms within a 5Å centred on each mutated residue (ligands are treated as a single residue). ^{b}NM-region consists of all heavy ligand or protein atoms outside the M-region.

			Force constraint (kcal/(mol Å^2))	
Stage	Process	Duration (ps)	Ligand	Protease
eq 0	Minimisation	2000 steps	4	4
eq 1	Annealing	50	4	4
eq 2	NPT solvation	200	4	4
	Mutation Relaxation		M-region^a	NM-region^b
eq (2 + 1)	M1-region relaxation	50	0	4
eq (2 + 2)	M2-region relaxation	50	0	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
eq (2 + n)	Mn-region relaxation	50	0	4
			Ligand	Protease
eq (2 + n + 1)	Constraint removal (NPT)	50	3	4
eq (2 + n + 2)		50	2	4
eq (2 + n + 3)		50	1	4
eq (2 + n + 4)		50	0	4
eq (2 + n + 5)		50	0	3
eq (2 + n + 6)		50	0	2
eq (2 + n + 7)		50	0	1
eq (2 + n + 8)	Unconstrained (NPT)	1400 - 50n	0	0

in order to solvate and neutralise the system in the 'tleap' module of AMBER 9. A directory is created into which all subsequent data corresponding to the system will be stored, this is termed the 'concourse'.

As part of this process the equilibration and simulation run files must be generated. The precise contents depends on the choice of simulation package and the contents of the equilibration protocol specified. In terms of the equilibration protocol, the user has the ability to set the number of simulation steps performed at each equilibration stage but the settings used during each one and the ordering of steps is pre-defined. The default equilibration protocol for PR simulations is shown in Table A.1 and detailed in Chapter 5 (changes appropriate for RT systems are provided in Chapter 7).

The simulation input files are transferred to a sub-directory within the concourse. A set of submission scripts for target compute resources is generated and copied to another concourse sub-directory. This collection of scripts is what we refer to as the Sim-Chain application. The submission scripts use relative paths, which means that once the concourse is transferred to a target resource, the Sim-Chain application can then be run by executing each individual submission script from within the appropriate concourse sub-directory. The submission scripts are designed to sequentially run a range of equilibration or simulation stages, executed by NAMD2 or GROMACS 4. By default all jobs are run on 64 processors. These submission scripts can either be run through the Unit-Executor, manually via AHE or the command line on the selected compute resource. Using the last two methods it is incumbent on the user to check for job completion before submitting the subsequent steps of the simulation. The Unit-Executor checks for completion of each stage via AHE and automatically submits any subsequent stages.

A.4 The FE-Calc Application

The FE-Calc application executes MMPBSA and normal mode analysis using the appropriate modules of the AMBER 9 software package. Again the application consists of a Perl script that generates all the input files necessary for a calculation and their subsequently submission to an appropriate compute resource.

Both the MMPBSA and normal mode calculations in AMBER 9 require separate topology files for the complex, ligand and receptor (stripped of solvent molecules) and input trajectories written in the AMBER .traj format. The pdb produced by BAC-Builder as input to the molecular dynamics simulations is split into 3 separate pdbs; for the complex, ligand and receptor. A 'tleap' source file is produced and then executed in order to produce the necessary topology files.

FE-Calc then generates and executes command files for the 'ptraj' module of AMBER 9 in order to convert the simulation trajectories into AMBER format. Whilst a common interface is provided in AMBER 9 for MMPBSA and normal mode analyses, the BAC protocol uses different parameters for each step, necessitating separate input files. These are generated from existing templates which are modified by FE-Calc. The atom numbers of the first and last atoms in each component are determined and the input scripts edited accordingly. The appropriate snapshot frequency and output filenames for the analysis program are also altered. Both calculations are then launched using generic job submission scripts on an appropriate compute resource.

A.5 Virolab Virtual Laboratory

As part of the EU ViroLab project[2] the BAC workflow was integrated into a virtual laboratory, consisting of a range of computational tools designed to investigate HIV resistance in a clinical setting [382–384], as discussed in Chapter 6. This implementation of BAC uses GridSpace Engine [386] to provide the User Interface component of the architecture shown in Figure A.1.

[2] ViroLab: www.virolab.org

Appendix B

Reverse Transcription in Detail

B.1 The Reverse Transcription Process

Reverse transcription of the viral genome proceeds via a number of steps, further details of each stage are provided in the following sections.

B.1.1 Initiation

Reverse transcription is initiated by the binding of a cellular tRNA$^{\mathrm{Lys}}$ primer to the primary binding site (PBS) of the viral genome. The PBS is a region approximately 200 nucleotides downstream from the 5′-end of the genome which is complementary to the 18 nucleotides at the 3′ end of human tRNA$^{\mathrm{Lys}}$ [420]. RT recognizes the tRNA/RNA complex and initiates the process of reverse transcription by extending the 3′-end of the annealed primer, using the RNA genome as a template for DNA synthesis.

The RT enzyme is a polymerase which is capable of using either RNA or DNA sequences as the template for DNA synthesis[301, 421–423]. It contains two active sites. The first catalyses the polymerisation reaction involved in extending DNA chains, while the other, known as the ribonuclease H (RNaseH), specifically degrades the RNA strand of RNA/DNA hybrids [301]. Further detail on the structure of RT is given in Section 4.7.

Once the RT has obtained a suitable template/primer complex it initiates the synthesis of the minus strand DNA, the first step in the reverse transcription process shown in Figure B.1 and described below.

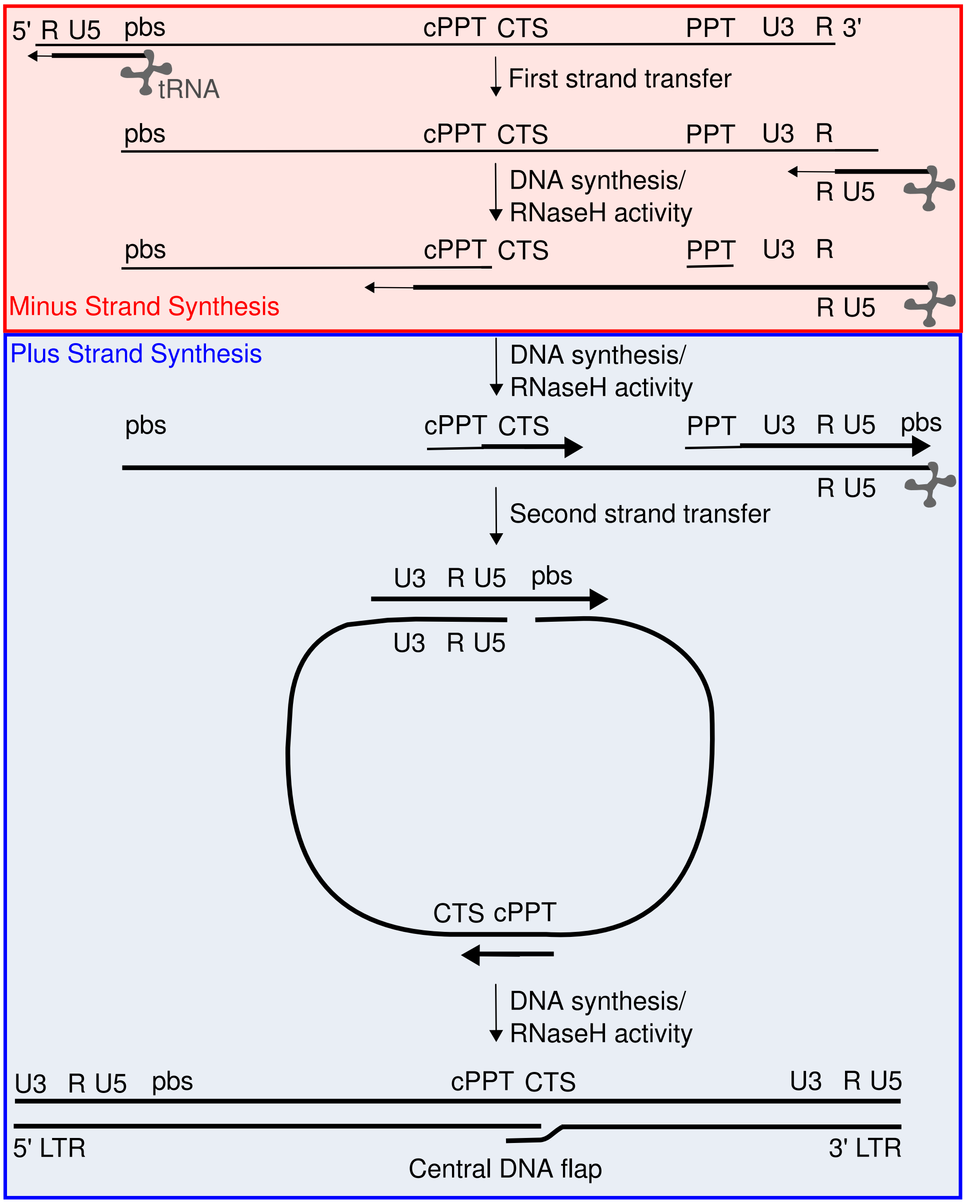


Figure B.1: The steps involved in the reverse transcription process. RNA is shown as a thin line, DNA as a thick one. The synthesis of DNA is punctuated by two transfers which change the position at which the nascent chain is being extended. Adapted from [161]

B.1.2 Minus Strand Synthesis

Following the initiation of reverse transcription, synthesis of minus strand DNA proceeds towards the 5′-end of the RNA template [424, 425]. Concurrent with, but 18 nucleotides further along the RNA template, the polymerisation reaction, the RNA component of the newly created DNA/RNA hybrid is degraded by the RNaseH domain of RT [426, 427].

Experimental evidence has shown that RT pauses several times during the synthesis of minus strand DNA [428]. In particular processivity is seen to stall early in reverse transcription 1 to 5 nucleotides after the primer terminus and at homopolymeric regions of the template during minus strand elongation [429]. These observations have been used as evidence to suggest that *in vivo* other factors, cellular or viral, are required to complete viral DNA synthesis.

The first discrete product of this process is a strand of DNA known as minus strand strong stop DNA (-ssDNA), created when the RNA genome is copied from the PBS to the 5′ terminus of the viral genome [430]. Once the 5′-end of the RNA genome has been reached continued minus strand synthesis requires a strand transfer reaction, with -ssDNA being transfered to the 3′-end of the genomic RNA, allowing this to become the template for continued synthesis. The transfer of the -ssDNA from one end of the RNA genome to the other is facilitated by sections of the 97 nucleotide R region of the viral genome (part of the LTR shown in Figure 4.2), which are present in DNA form in the -ssDNA and complementary RNA form at the 3′-end of the RNA template. It is also thought that NC plays some role in correct strand transfer [431]. This “jump” may occur as an intramolecular or intermolecular event due to the presence of two identical RNA genomes in the virion [432]. Once strand transfer has occurred synthesis of the minus strand DNA proceeds to the end of the template (now the 5′ end of the PBS as the R and U5 regions have been degraded by the RNaseH).

B.1.3 Plus Strand Synthesis

In order to create completed proviral DNA it is necessary to copy the minus strand DNA to create a DNA:DNA duplex. Again RNA primers are needed to initiate the copying of the template. Two sections of the RNA genome are resistant to degradation by RNaseH; these are the polypurine tract (PPT) and the central polypurine tract (cPPT) and it is these remaining fragments of the RNA genome which act as primers for the synthesis of the plus strand DNA. The primary site is the PPT, a purine rich sequence which is common to all retroviruses. The cPPT is unique to HIV and provides a second efficient priming site[161]. Recent single molecule flourescence resonance energy transfer (FRET) experiments by Abbondanzieri *et al.* [332] have shown that when arbitrary short RNA segments are bound to DNA the enzyme binds almost exclusively in a position to perform RNaseH activity. However, the when PPT primers are used a larger proportion bind in a polymerase competent direction. This proportion is significantly raised by the presence of dNTPs. Furthermore, in the case where a DNA primer is used RT binds almost exclusively in an orientation to extend the nascent DNA chain. The enzyme was also observed to flip orientation without dissociating from its nucleic acid substrate to allow synthesis to begin.

The PPT primed plus strand DNA is elongated until the 5′ end of the minus strand template is reached. This copies the U3, R and U5 sections of the minus strand. Synthesis continues using the tRNA minus strand primer as a template until a stop signal is reached, this reproduces the PBS sequence [433]. This DNA fragment is known as the plus strand strong stop DNA (+sssDNA).

At this stage the 3′ end of the +sssDNA is forming part of an RNA:DNA hybrid with the complimentary section of the tRNA primer. This hybrid can be degraded by the RNaseH leaving the newly synthesized PBS exposed [434]. Once this occurs the plus strand PBS sequence is free to base pair with the PBS sequence of the minus strand. The transfer of the +sssDNA to the 5′ end of the minus strand template facilitated by these complementary sequences is called the second strand transfer and creates the full LTRs at both end of the proviral genome.

Following the strand transfer plus strand synthesis continues until the central termination signal (CTS) is reached. As the CTS is at the 3′ end of the cPPT (which has acted as a second primer location with synthesis initiated at this point as well as at the PPT) approximately 100 nucleotides of plus strand DNA is displaced which results in the creation of a DNA "flap" (see Figure B.1) [435]. There is some evidence to suggest that this flap plays an important role in the transport of the PIC into the nucleus [436].

The minus strand is also completed, using the plus strand segment that originated as the +sssDNA as a template to form the full double stranded LTRs [437, 438]. The PPTs are removed during strand displacement or by RNaseH activity.

B.2 Template/Primer Binding and Positioning

In order allow to allow DNA synthesis RT must undergo significant changes in conformation. When RT is ligated to a template/primer duplex areas of the fingers, thumb and palm all contribute to keeping the ligand in the correct position to allow catalytic activity to proceed [439]. Tyr183 and Met184, which are part of the conserved YMDD motif[440] interact with the 3′ terminal nucleotide of the primer strand. The β12-β13 hairpin is known as the "primer grip" and interacts with the 3′ terminal phosphates of the primer strand, helping to align it correctly relative to the active site [439, 440]. Mutational studies have shown that residues 229 to 232 which form this hairpin loop affect both RNaseH and polymerase activity [441, 442]. Residues Asp76, Glu89, Glu151, Gly152, Lys154 and Pro157 constitute the "template grip" and as the name implies have close contact with and are responsible for maintaining the position of the template strand. Further positional stability is added by the $^{259}KLVGKL(X)_{16}KLLR^{284}$ motif which is found in the αH-turn-αI region, which is homologous to similar structures in a number of other DNA polymerases [443]. The αH section is partially inserted into the minor groove of the template/primer duplex with αI adjacent to the template strand.

Consistent with the multi-functional nature of RT in both RNA/DNA and DNA/DNA duplexes the majority of protein interactions are with the sugar-phosphate backbone of the template/primer [439].

B.3 Polymerase Function

The polymerase active site consists of residues Asp110, Asp185, Asp186 of the p66 subunit, all of which mutational studies have found to be essential for the enzyme to exhibit polymerase activity [302]. Both modelling and kinetic studies suggest that the polymerisation reaction proceeds with the three aspartate residues participating in the initial binding of the nucleotide through chelation of two Mg^{2+} ions. It is believed that Asp110 and Asp186 then stabilise the transition state of the polymerisation reaction. The model proposed in [444] indicates that the first Mg^{2+} ion is bound to the β and γ phosphates of the incoming deoxynucleotide triphosphate (dNTP) and to Asp110 and Asp186. The second ion creates an α-phosphate - Mg^{2+} - Asp185 complex. This complex facilitates the nucleophilic attack of the oxygen atom of the 3′-OH of the primer terminus. Evidence for this hypothesis comes from crystal structures which show two Mg^{2+} ions (designated metal A and metal B) coordinated with residues in the active site [306, 327, 439, 445]. In addition to this a recent structure of ATP crystallised with RT [327] (in the absence of a any template or primer) has shown metal B coordinated with the carboxylate oxygen atom of Asp185 and Asp110 and the N7 ATP nitrogen (which in this structure is in a position equivalent to that the 3′-OH adopts when a DNA primer is present [327] and a water molecule. This has been claimed as a model of a transition state in the polymerisation (or the reverse excision) reaction. The model suggests that when presented with a nucleotide Asp186 plays a role in the positioning of the α phosphate and orientating the scissile P-O bond for catalysis, whereas Asp110 and Asp185 are primarily responsible for positioning the triphosphate.

Appendix C

Reverse Transcriptase Crystal Structures

A list of all RT sequences used in the structural comparisons in Chapter 4 with information about any ligand present and any mutations within the sequence. The table was compiled in May 2007 when there were 92 structures of the complete HIV-1 RT duplex in the PDB, while more structures have been added since none containing substantial conformational differences have been reported. All of the structures have very similar sequences. The sequence shown in Table C.1 is that of the HXB2 wild type, and is referred to as sequence A in the following table. However, more than half of the sequences contain the following list of mutations; K172R, S280C, K416R, P468T, N471D, K512Q and I559V with respect to sequence A, the sequence with these present will be referred to as sequence B.

The mutation Q258C is commonly used to facilitate crosslinking of the template/primer duplex to RT and E478Q to eliminate RNaseH activity.

Table C.1: HXB2 wild type RT sequence.

1	PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPV
61	FAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPL
121	DEDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFKKQNPDIVI
181	YQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWGLTTPDKKHQKEPPFLWMGYELHPDKWT
241	VQPIVLPEKDSWTVNDIQKLVGKLNWASQIYPGIKVRQLSKLLRGTKALTEVIPLTEEAE
301	LELAENREILKEPVHGVYYDPSKDLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARMRGA
361	HTNDVKQLTEAVQKITTESIVIWGKTPKFKLPIQKETWETWWTEYWQATWIPEWEFVNTP
421	PLVKLWYQLEKEPIVGAETFYVDGAANRETKLGKAGYVTNKGRQKVVPLTNTTNQKTELQ
481	AIYLALQDSGLEVNIVTDSQYALGIIQAQPDKSESELVNQIIEQLIKKEKVYLAWVPAHK
541	GIGGNEQVDKLVSAGIRKIL

Table C.2: Table of all the HIV-1 RT structures used in the structure comparisons in Chapter 4. Also shown is the resolution (Res), R value, crystal space group, number of amino acids of the p66 chain included (AA No.), whether the protein has sequence A or B (Seq.), the mutations present relative to sequence A or B and the reference from which each structure originates(Ref.)

PDB	Ligand	Res (Å)	R Value	Space Grp.	AA No.	Seq.	Mutations	Ref.
1BQM	HBY 097	3.1	0.26	C 1 2 1	556	A		[446]
1BQN	HBY 097	3.3	0.25	C 1 2 1	558	A	Y188L E248Q	[446]
1C0T	BM + 21.1326	2.7	0.21	P 21 21 21	560	B		[447]
1C0U	BM + 50.0934	2.52	0.23	P 21 21 21	560	B		[447]
1C1B	GCA-186	2.5	0.2	P 21 21 21	560	B		[448]
1C1C	TNK-6123	2.5	0.23	P 21 21 21	560	B		[448]
1DLO	None	2.7	0.25	C 1 2 1	556	A		[323]
1DTQ	PETT-1 (PETT131A94)	2.8	0.22	P 21 21 21	560	B		[449]
1DTT	PETT-2 (PETT130A94)	3	0.2	P 21 21 21	560	B		[449]
1EET	MSC204	2.73	0.21	C 2 2 21	557	A		[95]
1EP4	DMP-266 (Efavirenz)	2.5	0.25	P 21 21 21	560	B		[450]
1FK9	DMP-266 (Efavirenz)	2.5	0.22	P 21 21 21	543	B		[451]
1FKO	DMP-266 (Efavirenz)	2.9	0.21	P 21 21 21	543	B	K103N	[451]
1FKP	Nevirpine (Viramune)	2.9	0.22	P 21 21 21	543	B	K103N	[451]
1HMV	None	3.2	0.25	C 1 2 1	560	A		[325]
1HNI	2,6-Br2 -APA (R95845)	2.8	0.26	C 1 2 1	558	A		[452]
1HNV	8-Cl TIBO (R86183)	3	0.25	C 1 2 1	558	A		[453]
1HPZ	2,6-Cl2 -APA (R90385)	3	0.25	C 1 2 1	560	A	K103N	[343]
1HQE	None	2.7	0.25	C 1 2 1	560	A	K103N	[343]
1HQU	HBY 097	2.7	0.25	C 1 2 1	560	A	K103N	[343]
1HVU	RNA Pseudoknot	4.75	0.34	C 1 2 1	554	A		[454]
1HYS	RNA/DNA	3	0.27	P 32 1 2	553	A		[409]
1IKV	DMP-266 (Efavirenz)	3	0.23	C 2 2 21	560	A	K103N	[93]
1IKW	DMP-266 (Efavirenz)	3	0.22	C 2 2 21	560	A		[93]
1IKX	PNU142721	2.8	0.21	C 2 2 21	560	A	K103N	[93]
1IKY	MSC194	3	0.21	C 2 2 21	560	A	K103N	[93]
1J5O	DNA/FAB	3.5	0.26	P 32 1 2	558	A	M184I	[455]
1JKH	DMP-266 (Efavirenz)	2.5	0.24	P 21 21 21	560	B	Y181C	[314]
1JLA	TNK-651	2.5	0.2	P 21 21 21	560	B	Y181C	[314]
1JLB	Nevirpine (Viramune)	3	0.21	P 21 21 21	560	B	Y181C	[314]
1JLC	PETT-2 (PETT130A94)	3	0.23	P 21 21 21	560	B	Y181C	[314]
1JLE	None	2.8	0.26	P 21 21 21	560	B	Y188C	[314]
1JLF	Nevirpine (Viramune)	2.6	0.24	P 21 21 21	560	B	Y188C	[314]
1JLG	UC-781	2.6	0.22	P 21 21 21	560	B	Y188C	[314]
1JLQ	739W34	3	0.22	P 21 21 21	560	B		[456]
1KLM	BHAP U-90152 (Delaviridine)	2.65	0.24	P 21 21 21	560	B		[336]
1LW0	Nevirpine (Viramune)	2.8	0.22	P 21 21 21	560	B	T215Y	[457]
1LW2	1051U19	3	0.21	P 21 21 21	560	B	T215Y	[457]
1LWC	Nevirpine (Viramune)	2.62	0.22	P 21 21 21	560	B		[457]
1LWE	Nevirpine (Viramune)	2.81	0.21	P 21 21 21	560	B	M41L T215Y	[457]
1LWF	Nevirpine (Viramune)	2.8	0.23	P 21 21 21	560	B	M41L D67N K70R M184V T215Y	[457]
1N5Y	DNA(AZTMP terminated)/FAB	3.1	0.26	P 32 1 2	558	A	Q258C	[445]
1N6Q	DNA(AZTMP terminated)/FAB	3	0.25	P 32 1 2	558	A	Q258C	[445]
1QE1	None	2.85	0.26	C 1 2 1	558	A	M184I	[455]
1R0A	DNA/FAB	2.8	0.24	P 32 1 2	558	A	Q258C	[458]
1REV	9-Cl TIBO (R82913)	2.6	0.22	P 21 21 21	560	B		[459]
1RT1	MKC-422 (Emivirine)	2.55	0.2	P 21 21 21	560	B		[460]
1RT2	TNK-651	2.55	0.21	P 21 21 21	560	B		[460]
1RT3	1051U19	3	0.26	P 21 21 21	560	B	D67N K70R T215F K219Q	[461]
1RT4	UC-781	2.9	0.24	P 21 21 21	560	B		[462]
1RT5	UC-10	2.9	0.23	P 21 21 21	560	B		[462]
1RT6	UC-38	2.8	0.24	P 21 21 21	560	B		[462]
1RT7	UC-84	3	0.26	P 21 21 21	560	B		[462]
1RTD	DNA/dNTP	3.2	0.22	P 21 21 21	554	B	P1K Q258C	[306]
1RTH	1051U19	2.2	0.21	P 21 21 21	560	B		[408]
1RTI	HEPT	3	0.24	P 21 21 21	560	B		[408]
1RTJ	None	2.35	0.22	P 21 21 21	560	B		[313]
1S1T	UC-781	2.4	0.21	P 21 21 21	560	B	L100I	[413]
1S1U	Nevirpine (Viramune)	3	0.23	P 21 21 21	560	B	L100I	[413]
1S1V	TNK-651	2.6	0.23	P 21 21 21	560	B	L100I	[413]
1S1W	UC-781	2.7	0.21	P 21 21 21	560	B	V106A	[413]
1S1X	Nevirpine (Viramune)	2.8	0.24	P 21 21 21	560	B	V108I	[413]
1S6P	R100943	2.9	0.25	C 1 2 1	560	A		[334]

Continued on next page

PDB	Ligand	Res (Å)	R Value	Space Grp.	AA No.	Seq.	Mutations	Ref.
1S6Q	R147681	3	0.25	C 1 2 1	560	A		[334]
1S9E	R129385	2.6	0.25	C 1 2 1	560	A		[334]
1S9G	R120394	2.8	0.24	C 1 2 1	560	A		[334]
1SUQ	R185545	3	0.26	C 1 2 1	560	A		[334]
1SV5	R165335(Etravirine-TMC125)	2.9	0.26	C 1 2 1	560	A	K103N	[334]
1T03	DNA(Tenofovir terminated)/FAB	3.1	0.26	P 32 1 2	558	A	Q258C	[463]
1T05	DNA/Tenofovir	3	0.25	P 31 1 2	558	A	Q258C	[463]
1TKT	GW426318	2.6	0.21	P 21 21 21	560	B		[464]
1TKX	GW490745	2.85	0.22	P 21 21 21	560	B		[465]
1TKZ	GW429576	2.81	0.21	P 21 21 21	560	B		[464]
1TL1	GW451211	2.9	0.22	P 21 21 21	560	B		[464]
1TL3	GW450557	2.8	0.21	P 21 21 21	560	B		[464]
1TV6	CP-94,707	2.8	0.26	C 1 2 1	560	A		[328]
1TVR	9-Cl TIBO (R82913)	3	0.26	C 1 2 1	558	A		[466]
1UWB	8-Cl TIBO (R86183)	3.2	0.27	C 1 2 1	558	A	Y181C	[466]
1VRT	Neviripine (Viramune)	2.2	0.19	P 21 21 21	560	B		[408]
1VRU	2,6-Cl2 -APA (R90385)	2.4	0.19	P 21 21 21	560	B		[408]
2B5J	JANSSEN-R165481	2.9	0.25	C 1 2 1	560	A		[467]
2BAN	JANSSEN-R157208	2.95	0.24	C 1 2 1	560	A		[467]
2BE2	R221239	2.43	0.24	C 1 2 1	560	A		[467]
2HMI	DNA/FAB	2.8	0.27	P 32 1 2	558	A		[403]
2HND	Neviripine (Viramune)	2.5	0.2	P 21 21 21	534	B	K101E	[468]
2HNY	Neviripine (Viramune)	2.5	0.21	P 21 21 21	534	B	E138K	[468]
2HNZ	PETT-2 (PETT130A94)	3	0.23	P 21 21 21	534	B	E138K	[468]
2I5J	DHBNH	3.15	0.27	C 1 2 1	552	A		[469]
2IAJ	ATP	2.5	0.23	C 1 2 1	560	A	K103N Y181C	[327]
2IC3	HBY 097	3	0.26	C 1 2 1	560	A	K103N Y181C	[327]
3HVT	Neviripine (Viramune)	2.9	0.27	C 1 2 1	556	A		[404]

It is generally contended that despite changes in overall conformation depending on the bound ligand the individual subdomains of RT only undergo relatively minor rearrangement. The average root mean squared deviations (RMSD) of the subdomains, calculated by aligning the subdomains of each of the crystal structures to the unliganded average structure, are shown in Table C.3. Almost all of these values are below 1.5 Å which indicates that the internal conformation of the subdomains remains similar throughout the structures. The notable exception to this is the palm subdomain of the NNRTI bound structures, where the value rises to 2.23 Å due to the formation of the NNRTIBP. A very similar value is obtained for the two open unliganded structures which adds credence to the idea they are more representative of the NNRTI bound form than the true open conformation of the apo enzyme.

Table C.3: Average subdomain RMSD in Å for all HIV-1 RT crystal structures in the PDB, broken down by subdomain and according to the class of ligand present. Figures in brackets are the standard deviations. All deviations are small, indicating that the individual subdomain structures are stable and undergo minor rearrangement.

Ligand Type	**Fingers**	**Palm**	**Thumb**	**Connection**	**RnaseH**
Unliganded	0.65 (0.16)	0.46 (0.03)	0.83 (0.11)	0.40 (0.04)	0.60 (0.13)
Template/Primer	1.27 (0.29)	0.85 (0.11)	1.46 (0.17)	0.57 (0.1)	0.77 (0.29)
NNRTI	1.26 (0.28)	2.23 (0.22)	1.21 (0.24)	0.88 (0.24)	0.85 (0.29)
Unliganded (open)	1.21 (0.16)	2.24 (0.09)	1.00 (0.04)	1.03 (0.05)	0.93 (0.31)
Overall	**1.24 (0.31)**	**1.99 (0.59)**	**1.21 (0.26)**	**0.82 (0.26)**	**0.83 (0.28)**

Bibliography

[1] L. Brocchieri and S. Karlin. 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res*, **33**, 3390–3400. URL http://dx.doi.org/10.1093/nar/gki615. (doi:10.1093/nar/gki615)

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter. *Molecular Biology of the Cell*. Garland Science, Fourth edition, 2002.

[3] C. Branden and J. Tooze. *Introduction to protein structure*. Garland Publishing, New York, 1991.

[4] L. Stryer, J. M. Berg and J. L. Tymoczko. *Biochemistry*. W. H. Freeman and Co. Ltd., Fifth edition, 2002.

[5] T. Nakamoto. 2009. Evolution and the universality of the mechanism of initiation of protein synthesis. *Gene*, **432**, 1–6. URL http://dx.doi.org/10.1016/j.gene.2008.11.001. (doi:10.1016/j.gene.2008.11.001)

[6] Madison Area Technical College: http://matcmadison.edu/biotech/resources/proteins/.

[7] G. N. Ramachandran, C. Ramakrishnan and V. Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations. *J Mol Biol*, **7**, 95–99.

[8] L. Pauling, R. B. Corey and H. R. Branson. 1951. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, **37**, 205–211.

[9] A. Maton, J. Hopkins, C. W. McLaughlin, S. Johnson, M. Q. Warner, D. LaHart and J.D. Wright. *Human biology and health. Englewood Cliffs, N.J: Prentice Hall.* Prentice Hall, 1993.

[10] M. Paoli, R. Liddington, J. Tame, A. Wilkinson and G. Dodson. 1996. Crystal structure of T state haemoglobin with oxygen bound at all four haems. *J Mol Biol*, **256**, 775–792. URL http://dx.doi.org/10.1006/jmbi.1996.0124. (doi:10.1006/jmbi.1996.0124)

[11] M. A. Mitz. 1979. CO2 biodynamics : A new concept of cellular control. *Journal of Theoretical Biology*, **80**, 537 – 551. ISSN 0022-5193. URL http://www.sciencedirect.com/science/article/B6WMD-4F1SV94-22/2/9671abd2fa69589f31c114fff0b45431. (doi:10.1016/0022-5193(79)90092-4)

[12] C. B. Anfinsen. 1973. Principles that govern the folding of protein chains. *Science*, **181**, 223–230.

[13] C. Levinthal. 1968. Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique*, **65**, 44–45. URL http://www.biochem.wisc.edu/courses/biochem704/Reading/Levinthal1968.pdf.

[14] H. J. Dyson and P. E. Wright. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, **6**, 197–208. URL http://dx.doi.org/10.1038/nrm1589. (doi:10.1038/nrm1589)

[15] D. Blow. *Outline of crystallography for biologists.* Oxford University Press, 2002.

[16] A. L. Morris, M. W. MacArthur, E. G. Hutchinson and J. M. Thornton. 1992. Stereochemical quality of protein structure coordinates. *Proteins*, **12**, 345–364. URL http://dx.doi.org/10.1002/prot.340120407. (doi:10.1002/prot.340120407)

[17] K. V. R. Chary and G. Govil. *NMR in biological systems : from molecules to human.* Springer, 2008.

[18] R. Ishima and D. A. Torchia. 2000. Protein dynamics from NMR. *Nat Struct Biol*, **7**, 740–743. URL http://dx.doi.org/10.1038/78963. (doi:10.1038/78963)

[19] E. Krieger, S. B. Nabuurs and G. Vriend. 2003. Homology modeling. *Methods Biochem Anal*, **44**, 509–523.

[20] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol*, **215**, 403–410. URL http://dx.doi.org/10.1006/jmbi.1990.9999. (doi:10.1006/jmbi.1990.9999)

[21] W. R. Pearson and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**, 2444–2448.

[22] D. M. Mount. *Bioinformatics: Sequence and Genome Analysis.* Cold Spring Harbor Laboratory Press, 2nd edition, 2003.

[23] D. G. Higgins, J. D. Thompson and T. J. Gibson. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*, **266**, 383–402.

[24] T. Schwede, J. Kopp, N. Guex and M. C. Peitsch. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res*, **31**, 3381–3385.

[25] N. Guex and M. C. Peitsch. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723. URL http://dx.doi.org/10.1002/elps.1150181505. (doi:10.1002/elps.1150181505)

[26] R. L. Dunbrack and M. Karplus. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol*, **230**, 543–574. URL http://dx.doi.org/10.1006/jmbi.1993.1170. (doi:10.1006/jmbi.1993.1170)

[27] V. De Filippis, C. Sander and G. Vriend. 1994. Predicting local structural changes that result from point mutations. *Protein Eng*, **7**, 1203–1208.

[28] S. Y. Chung and S. Subbiah. 1996. A structural explanation for the twilight zone of protein sequence homology. *Structure*, **4**, 1123–1127.

[29] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez and J. A. Pople. Gaussian 03, Revision C.02, 2004. Gaussian, Inc., Wallingford, CT, 2004.

[30] G.H. Grant and W.G. Richards. *Computational Chemistry*. Oxford University Press, 1995.

[31] F. Jensen. *Introduction to Computational Chemistry*. John Wiley & Sons, 1999.

[32] A. R. Leach. *Molecular modelling: principles and applications*. Prentice Hall, second edition edition, 2001.

[33] D. Frenkel and B. Smit. *Understanding Molecular Simulations: from Algorithms to Applications*. Academic Press, San Diego, 2002.

[34] M. Müller, K. Katsov and M. Schick. 2003. Coarse-grained models and collective phenomena in membranes: Computer simulation of membrane fusion. *Journal of Polymer Science Part B: Polymer Physics*, **41**, 1441–1450.

[35] J. C. Shelley and M. Y. Shelley. 2000. Computer simulation of surfactant solutions. *Current Opinion in Colloid & Interface Science*, **5**, 101–110. ISSN 1359-0294. URL http://www.sciencedirect.com/science/article/B6VRY-414WV0R-J/2/143bf8fcd57de833930bdd805e36a8f4. (doi:10.1016/S1359-0294(00)00042-X)

[36] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.N. Teller and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

[37] J. Norberg and L. Nilsson. 2003. Advances in biomolecular simulations: methodology and recent applications. *Q Rev Biophys*, **36**, 257–306.

[38] M. J. Field, P. A. Bash and M. Karplus. 1990. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.*, **11**, 700–733.

[39] N. Reuter, A. Dejaegere, B. Maigret and M. Karplus. 2000. Frontier bonds in QM/MM methods: A comparison of different approaches. *J. Phys. Chem. A,*, **104**, 1720–1735.

[40] D. L. Ermak and J. A. McCammon. 1978. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.*, **69**, 1352–1360. (doi:10.1063/1.436761)

[41] R. R. Gabdoulline and R. C. Wade. 2002. Biomolecular diffusional association. *Current Opinion in Structural Biology*, **12**, 204 – 213. ISSN 0959-440X. URL http://www.sciencedirect.com/science/article/B6VS6-45JPC7V-D/2/f8cd91ecd96a703181556f6c10a02852. (doi:10.1016/S0959-440X(02)00311-1)

[42] E. G. Flekkøy and P. V. Coveney. 1999. From Molecular Dynamics to Dissipative Particle Dynamics. *Phys. Rev. Lett.*, **83**, 1775–1778. (doi:10.1103/PhysRevLett.83.1775)

[43] E. G. Flekkøy, P. V. Coveney and G. De Fabritiis. 2000. Foundations of dissipative particle dynamics. *Phys. Rev. E*, **62**, 2140–2157. (doi:10.1103/PhysRevE.62.2140)

[44] I. Bahar, A. R. Atilgan and B. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, **2**, 173–181.

[45] T. Haliloglu and I. Bahar. 1999. Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. *Proteins*, **37**, 654–667.

[46] C. Chennubhotla, A. J. Rader, L. Yang and I. Bahar. 2005. Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Phys Biol*, **2**, 173–180. URL http://dx.doi.org/10.1088/1478-3975/2/4/S12. (doi:10.1088/1478-3975/2/4/S12)

[47] N. A. Temiz and I. Bahar. 2002. Inhibitor binding alters the directions of domain motions in HIV-1 reverse transcriptase. *Proteins*, **49**, 61–70. URL http://dx.doi.org/10.1002/prot.10183. (doi:10.1002/prot.10183)

[48] D. Tobi and I. Bahar. 2005. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc Natl Acad Sci U S A*, **102**, 18908–18913. URL http://dx.doi.org/10.1073/pnas.0507603102. (doi:10.1073/pnas.0507603102)

[49] L. D. Landau and E. M. Lifshitz. *Mechanics, Volume 1 of Course of Theoretical Physics.* Butterworth-Heinemann Ltd, 1980.

[50] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J Comput Chem*, **26**, 1781–1802. URL http://dx.doi.org/10.1002/jcc.20289. (doi:10.1002/jcc.20289)

[51] A. D. Jr. MacKerell, B. Brooks, C.L. Brooks III, L. Nilsson, B. Roux, Y. Won and M. Karplus. CHARMM: The energy function and its parameterization with an overview of the program. In P.v.R. Schleyer, P.R. Schreiner, N.L. Allinger, T. Clark, J. Gasteiger, P. Kollman and H.F. SchaeferIII, editors, *The Encyclopedia of Computational Chemistry*, volume 1, pages 271–277. John Wiley & Sons., 1998.

[52] J. W. Ponder and D. A. Case. 2003. Force fields for protein simulations. *Adv Protein Chem*, **66**, 27–85.

[53] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case. 2004. Development and testing of a general Amber force field. *J Comput Chem*, **25**, 1157–1174. URL http://dx.doi.org/10.1002/jcc.20035. (doi:10.1002/jcc.20035)

[54] L. Verlet. 1967. Computer ‘experiments’ on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review*, **159**, 98–103.

[55] W. C. Swope, H. C. Andersen and K. R. Wilson P. H. Berens. 1982. A computer simulation method for the calculation of equilibrium constants for the formation

of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, **76**, 637–649.

[56] B. V. Chirikov. 1979. A universal instability of many-dimensional oscillator systems. *Phys Rep*, **52**, 263–378.

[57] F. Calvo. 1999. Largest Lyapunov exponent in molecular systems. II: Quaternion coordinates and application to methane clusters. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, **60**, 2771–2778.

[58] R. Bowley and M. Sanchez. *Introductory Statistical Mechanics*. Oxford U, 1999.

[59] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, **81**, 3684–3690. URL http://link.aip.org/link/?JCP/81/3684/1. (doi:10.1063/1.448118)

[60] S. A. Adelman and J. D. Doll. 1976. Generalized Langevin equation approach for atom/solid-surface scattering: General formulation for classical scattering off harmonic solids. *The Journal of Chemical Physics*, **64**, 2375–2388. URL http://link.aip.org/link/?JCP/64/2375/1. (doi:10.1063/1.432526)

[61] S. A. Adelman. 1979. Generalized Langevin theory for many-body problems in chemical dynamics: General formulation and the equivalent harmonic chain representation. *The Journal of Chemical Physics*, **71**, 4471–4486. URL http://link.aip.org/link/?JCP/71/4471/1. (doi:10.1063/1.438200)

[62] P. P. Ewald. 1921. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, **369**, 1521–3889. URL http://dx.doi.org/10.1002/andp.19213690304. (doi:10.1002/andp.19213690304)

[63] Tom Darden, Darrin York and Lee Pedersen. 1993. Particle mesh ewald: An n [center-dot] log(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, **98**, 10089–10092. URL http://link.aip.org/link/?JCP/98/10089/1. (doi:10.1063/1.464397)

[64] H. Grubmüller, H. Heller, A. Windemuth and K. Schulten. 1991. Generalized verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Mol. Sim.*, **6**, 121–142.

[65] M. Tuckerman, B. J. Berne and G. J. Martyna. 1992. Reversible multiple time scale molecular dynamics. *Journal of Chemical Physics*, **97**, 1990–2001. (doi:10.1063/1.463137)

[66] J. P. Ryckaert, G. Ciccotti and H. J. C. Berendsen. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.

[67] S. Miyamoto and P. A. Kollman. 1992. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.*, **13**, 952–962.

[68] D. Hamelberg, J. Mongan and J. A. McCammon. 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys*, **120**, 11919–11929. URL http://dx.doi.org/10.1063/1.1755656. (doi:10.1063/1.1755656)

[69] B. Isralewitz, M. Gao and K. Schulten. 2001. Steered molecular dynamics and mechanical functions of proteins. *Curr Opin Struct Biol*, **11**, 224–230.

[70] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods. 2005. The Amber biomolecular simulation programs. *J Comput Chem*, **26**, 1668–1688. URL http://dx.doi.org/10.1002/jcc.20290. (doi:10.1002/jcc.20290)

[71] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen. 2005. GROMACS: fast, flexible, and free. *J Comput Chem*, **26**, 1701–1718. URL http://dx.doi.org/10.1002/jcc.20291. (doi:10.1002/jcc.20291)

[72] B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl. 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, **4**, 435–447.

[73] S. Plimpton. 1995. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.*, **117**, 1–19. ISSN 0021-9991. URL http://portal.acm.org/citation.cfm?id=201627.201628. (doi:10.1006/jcph.1995.1039)

[74] S. J. Plimpton and B. A. Hendrickson. 1996. A new parallel method for molecular-dynamics simulation of macromolecular systems. *J Comp Chem*, **17**, 326–327.

[75] K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossváry, M. A. Moraes, F. D. Sacerdoti, Y. Shan J. K. Salmon and D. E. Shaw. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, Florida, November 11–17*, 2006.

[76] J. Hein, F. Reid, L. Smith, I. Bush, M. Guest and P. Sherwood. 2005. On the performance of molecular dynamics applications on current high-end systems.

Philos Transact A Math Phys Eng Sci, **363**, 1987–1998. URL http://dx.doi.org/10.1098/rsta.2005.1624. (doi:10.1098/rsta.2005.1624)

[77] M. Shirts and V. S. Pande. 2000. COMPUTING: Screen Savers of the World Unite! *Science*, **290**, 1903–1904. URL http://dx.doi.org/10.1126/science.290.5498.1903. (doi:10.1126/science.290.5498.1903)

[78] M. J. Harvey, G. Giupponi and G. de Fabritiis. 2009. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *Journal of Chemical Theory and Computation*, **5**, 1632–1639. URL http://pubs.acs.org/doi/abs/10.1021/ct9000685. (doi:10.1021/ct9000685)

[79] J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco and K. Schulten. 2007. Accelerating molecular modeling applications with graphics processors. *J Comput Chem*, **28**, 2618–2640. URL http://dx.doi.org/10.1002/jcc.20829. (doi:10.1002/jcc.20829)

[80] P. V. Coveney. 2005. Scientific Grid computing. *Philos Transact A Math Phys Eng Sci*, **363**, 1707–1713. URL http://dx.doi.org/10.1098/rsta.2005.1632. (doi:10.1098/rsta.2005.1632)

[81] P. H. Beckman. 2005. Building the TeraGrid. *Philos Transact A Math Phys Eng Sci*, **363**, 1715–1728. URL http://dx.doi.org/10.1098/rsta.2005.1602. (doi:10.1098/rsta.2005.1602)

[82] R. S. Saksena, B. Boghosian, L. Fazendeiro, O. A. Kenway, S. Manos, M. D. Mazzeo, S. K. Sadiq, J. L. Suter, D. Wright and P. V. Coveney. 2009. Real Science at the Petascale. *Phil Trans R Soc A*, **367**, 2557–2571.

[83] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson and G. De Fabritiis. 2010. High-throughput all-atom molecular dynamics simulations using distributed computing. *J Chem Inf Model*, **50**, 397–403. URL http://dx.doi.org/10.1021/ci900455r. (doi:10.1021/ci900455r)

[84] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan and Brian Towles. Millisecond-Scale Molecular Dynamics Simulations on Anton. In *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis (SC09), New York*, 2009.

[85] P. V. Coveney, R. S. Saksena, S. J. Zasada, M. McKeown and S. Pickles. 2007. The Application Hosting Environment: Lightweight Middleware for Grid-Based Computational Science. *Comp. Phys. Comm*, **176**, 406–418.

[86] C. J. Adkins. *Equilibrium Thermodynamics*. Cambridge University Press, 3rd edition, 1983.

[87] R. A. Alberty. 2004. A short history of the thermodynamics of enzyme-catalyzed reactions. *J Biol Chem*, **279**, 27831–27836. URL http://dx.doi.org/10.1074/jbc.X400003200. (doi:10.1074/jbc.X400003200)

[88] X. Chen, M. Liu and M. K. Gilson. 2001. BindingDB: a web-accessible molecular recognition database. *Comb Chem High Throughput Screen*, **4**, 719–725.

[89] X. Chen, Y. Lin and M. K. Gilson. 2001. The binding database: overview and user's guide. *Biopolymers*, **61**, 127–141. URL http://dx.doi.org/gt;3.0.CO;2-N. (doi:gt;3.0.CO;2-N)

[90] Y. Cheng and W. H. Prusoff. 1973. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem Pharmacol*, **22**, 3099–3108.

[91] R. A. Copeland. *Enzymes: a practical introduction to structure, mechanism, and data analysis*. John Wiley & Sons., 2000.

[92] H. Ohtaka, A. Schön and E. Freire. 2003. Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations. *Biochemistry*, **42**, 13659–13666. URL http://dx.doi.org/10.1021/bi0350405. (doi:10.1021/bi0350405)

[93] J. Lindberg, S. Sigurdsson, S. Löwgren, H. O. Andersson, C. Sahlberg, R. Noréen, K. Fridborg, H. Zhang and T. Unge. 2002. Structural basis for the inhibitory efficacy of efavirenz (DMP-266), MSC194 and PNU142721 towards the HIV-1 RT K103N mutant. *Eur J Biochem*, **269**, 1670–1677.

[94] S. Butini, M. Brindisi, S. Cosconati, L. Marinelli, G. Borrelli, S. S. Coccone, A. Ramunno, G. Campiani, E. Novellino, S. Zanoli, A. Samuele, G. Giorgi, A. Bergamini, M. Di Mattia, S. Lalli, B. Galletti, S. Gemma and G. Maga. 2009. Specific targeting of highly conserved residues in the HIV-1 reverse transcriptase primer grip region. 2. Stereoselective interaction to overcome the effects of drug resistant mutations. *J Med Chem*, **52**, 1224–1228. URL http://dx.doi.org/10.1021/jm801395v. (doi:10.1021/jm801395v)

[95] M. Högberg, C. Sahlberg, P. Engelhardt, R. Noréen, J. Kangasmetsä, N. G. Johansson, B. Oberg, L. Vrang, H. Zhang, B. L. Sahlberg, T. Unge, S. Lövgren, K. Fridborg and K. Bäckbro. 1999. Urea-PETT compounds as a new class of HIV-1 reverse transcriptase inhibitors. 3. synthesis and further structure-activity relationship studies of PETT analogues. *J Med Chem*, **42**, 4150–4160.

[96] A. Monforte, P. Logoteta, L. De Luca, N. Iraci, S. Ferro, G. Maga, E. De Clercq, C. Pannecouque and A. Chimirri. 2010. Novel 1,3-dihydro-benzimidazol-2-ones and their analogues as potent non-nucleoside HIV-1 reverse transcriptase inhibitors. *Bioorg Med Chem*, **18**, 1702–1710. URL http://dx.doi.org/10.1016/j.bmc.2009.12.059. (doi:10.1016/j.bmc.2009.12.059)

[97] C. Yun, T. J. Boggon, Y. Li, M. S. Woo, H. Greulich, M. Meyerson and M. J. Eck. 2007. Structures of lung cancer-derived egfr mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell*, **11**, 217–227. URL http://dx.doi.org/10.1016/j.ccr.2006.12.017. (doi:10.1016/j.ccr.2006.12.017)

[98] C. Yun, K. E. Mengwasser, A. V. Toms, M. S. Woo, H. Greulich, K. Wong, M. Meyerson and M. J. Eck. 2008. The t790m mutation in egfr kinase causes drug resistance by increasing the affinity for atp. *Proc Natl Acad Sci U S A*, **105**, 2070–2075. URL http://dx.doi.org/10.1073/pnas.0709662105. (doi:10.1073/pnas.0709662105)

[99] C. Chipot. *Free Energy Calculations.* Springer, 2007.

[100] M. R. Shirts, D. L. Mobley and J. D. Chodera. Chapter 4 alchemical free energy calculations: Ready for prime time? In D.C. Spellmeyer and R. Wheeler, editors, *Annual Reports in Computational Chemistry*, volume 3 of *Annual Reports in Computational Chemistry*, pages 41–59. Elsevier, 2007. URL http://www.sciencedirect.com/science/article/B7RNN-4PPMX91-6/2/9dd135f05a62e2a2c2fe8550da6632fb.

[101] Y. Deng and B. Roux. 2009. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *The Journal of Physical Chemistry B*, **113**, 2234–2246. URL http://pubs.acs.org/doi/abs/10.1021/jp807701h. (doi:10.1021/jp807701h)

[102] C. Chipot and D. A. Pearlman. 2002. Free Energy Calculations. The Long and Winding Gilded Road. *Molecular Simulation*, **28**, 1–12. (doi:10.1080/08927020211974)

[103] M. K. Gilson and H. Zhou. 2007. Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct*, **36**, 21–42. URL http://dx.doi.org/10.1146/annurev.biophys.36.040306.132550. (doi:10.1146/annurev.biophys.36.040306.132550)

[104] T. Steinbrecher and A. Labahn. 2010. Towards accurate free energy calculations in ligand protein-binding studies. *Curr Med Chem*.

[105] O. Trott and A. J. Olson. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, **31**, 455–461. URL http://dx.doi.org/10.1002/jcc.21334. (doi:10.1002/jcc.21334)

[106] R. Wang, L. and S. Wang. 2002. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des*, **16**, 11–26.

[107] H. F. G. Velec, H. Gohlke and G. Klebe. 2005. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem*, **48**, 6296–6303. URL http://dx.doi.org/10.1021/jm050436v. (doi:10.1021/jm050436v)

[108] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee. 1997. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design*, **11**, 425–445. ISSN 0920-654X. URL http://dx.doi.org/10.1023/A:1007996124545. 10.1023/A:1007996124545. (doi:10.1023/A:1007996124545)

[109] G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor. 1997. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, **267**, 727–748. URL http://dx.doi.org/10.1006/jmbi.1996.0897. (doi:10.1006/jmbi.1996.0897)

[110] M. Rarey, B. Kramer, T. Lengauer and G. Klebe. 1996. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, **261**, 470–489. URL http://dx.doi.org/10.1006/jmbi.1996.0477. (doi:10.1006/jmbi.1996.0477)

[111] A. Krammer, P. D. Kirchhoff, X. J., C.M. Venkatachalam and M. Waldman. 2005. LigScore: a novel scoring function for predicting binding affinities. *Journal of Molecular Graphics and Modelling*, **23**, 395–407. ISSN 1093-3263.

URL http://www.sciencedirect.com/science/article/B6TGP-4F3NY68-1/2/75d3aa2dff28900452c40414f7bcf207. (doi:10.1016/j.jmgm.2004.11.007)

[112] H. J. Böhm. 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Design*, **8**, 243–256.

[113] David L Mobley and Ken A Dill. 2009. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure*, **17**, 489–498. URL http://dx.doi.org/10.1016/j.str.2009.02.010. (doi:10.1016/j.str.2009.02.010)

[114] J. Aqvist, C. Medina and J. E. Samuelsson. 1994. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng*, **7**, 385–391.

[115] N. Foloppe and R. Hubbard. 2006. Towards predictive ligand design with free-energy based computational methods? *Curr Med Chem*, **13**, 3583–3608.

[116] T. Hansson and J. Aqvist. 1995. Estimation of binding free energies for HIV proteinase inhibitors by molecular dynamics simulations. *Protein Eng*, **8**, 1137–1144.

[117] J. Aqvist and J. Marelius. 2001. The linear interaction energy method for predicting ligand binding free energies. *Comb Chem High Throughput Screen*, **4**, 613–626.

[118] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case and T. E. Cheatham. 2000. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res*, **33**, 889–897.

[119] D. Sitkoff, K. Sharp and B. Honig. 1998. Accurate calculation of hydration free energies using macroscopic continuum models. *J. Phys. Chem.*, **98**, 1978–1983.

[120] J. Wang, P. Morin, W. Wang and P. A. Kollman. 2001. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J Am Chem Soc*, **123**, 5221–5230.

[121] C. Holm, P. Kkicheff and R. Podgornik, editors. *Electrostatic effects in soft matter and biophysics*. Springer, 2001.

[122] J. D. Jackson. *Classical Electrodynamics*. Wiley, 1998.

[123] F. Dong, B. Olsen and N. A. Baker. 2008. Computational methods for biomolecular electrostatics. *Methods Cell Biol*, **84**, 843–870. URL http://dx.doi.org/10.1016/S0091-679X(07)84026-X. (doi:10.1016/S0091-679X(07)84026-X)

[124] M. K. Gilson and B. H. Honig. 1986. The dielectric constant of a folded protein. *Biopolymers*, **25**, 2097–2119. URL http://dx.doi.org/10.1002/bip.360251106. (doi:10.1002/bip.360251106)

[125] B. Lee and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, **55**, 379–400.

[126] J. Warwicker and H. C. Watson. 1982. Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. *J Mol Biol*, **157**, 671–679.

[127] M. L. Connolly. 1985. Computation of molecular volume. *J AM Chem Soc*, **107**, 1118–1124.

[128] J. A. Grant, B. T. Pickup and A. Nicholls. 2001. A smooth permittivity function for Poisson-Boltzmann solvation methods. *Journal of Computational Chemistry*, **22**, 608–640. ISSN 1096-987X. URL http://dx.doi.org/10.1002/jcc.1032. (doi:10.1002/jcc.1032)

[129] P. Grochowski and J. Trylska. 2008. Continuum molecular electrostatics, salt effects, and counterion binding–a review of the Poisson-Boltzmann theory and its modifications. *Biopolymers*, **89**, 93–113. URL http://dx.doi.org/10.1002/bip.20877. (doi:10.1002/bip.20877)

[130] F. Fogolari, P. Zuccato, G. Esposito and P. Viglino. 1999. Biomolecular electrostatics with the linearized Poisson-Boltzmann equation. *Biophys J*, **76**, 1–16. URL http://dx.doi.org/10.1016/S0006-3495(99)77173-0. (doi:10.1016/S0006-3495(99)77173-0)

[131] Nathan A Baker. 2004. Poisson-boltzmann methods for biomolecular electrostatics. *Methods Enzymol*, **383**, 94–118. URL http://dx.doi.org/10.1016/S0076-6879(04)83005-2. (doi:10.1016/S0076-6879(04)83005-2)

[132] M. F. Sanner, A. J. Olson and J. C. Spehner. 1996. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, **38**, 305–320. URL http://dx.doi.org/gt;3.0.CO;2-Y. (doi:gt;3.0.CO;2-Y)

[133] C. Chipot. 2003. Rational determination of charge distributions for free energy calculations. *Journal of Computational Chemistry*, **24**, 409–415.

[134] D. L. Mobley, E. Dumont, J. D. Chodera and K. A. Dill. 2007. Comparison of charge models for fixed-charge force fields: small-molecule hydration free energies in explicit solvent. *J Phys Chem B*, **111**, 2242–2254. URL http://dx.doi.org/10.1021/jp0667442. (doi:10.1021/jp0667442)

[135] B. R. Brooks, D. Janezic and M. Karplus. 1995. Harmonic analysis of large systems. I. Methodology . *J. Comput. Chem.*, **16**, 1522–1542. URL http://dx.doi.org/10.1002/jcc.540161209. (doi:10.1002/jcc.540161209)

[136] J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman and D. A. Case. 1998. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *Journal of the American Chemical Society*, **120**, 9401–9409. URL http://pubs.acs.org/doi/abs/10.1021/ja981844%2B. (doi:10.1021/ja981844+)

[137] S. M. Schwarzl, T. B. Tschopp, J. C. Smith and S. Fischer. 2002. Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction? *J Comput Chem*, **23**, 1143–1149. URL http://dx.doi.org/10.1002/jcc.10112. (doi:10.1002/jcc.10112)

[138] I. Andricioaei and M. Karplus. 2001. On the calculation of entropy from covariance matrices of the atomic fluctuations. *The Journal of Chemical Physics*, **115**, 6289–6292. URL http://link.aip.org/link/?JCP/115/6289/1. (doi:10.1063/1.1401821)

[139] C. Chang, W. Wei and M. K. Gilson. 2005. Evaluating the Accuracy of the Quasiharmonic Approximation. *Journal of Chemical Theory and Computation*, **1**, 1017–1028. URL http://pubs.acs.org/doi/abs/10.1021/ct0500904. (doi:10.1021/ct0500904)

[140] UNAIDS. AIDS Epidemic Update 2009, 2009.

[141] B. M. Kuehn. 2006. UNAIDS report: AIDS epidemic slowing, but huge challenges remain. *JAMA*, **296**, 29–30. URL http://dx.doi.org/10.1001/jama.296.1.29. (doi:10.1001/jama.296.1.29)

[142] E. S. Daar and D. D. Richman. 2005. Confronting the emergence of drug-resistant HIV type 1: impact of antiretroviral therapy on individual and population resistance. *AIDS Res Hum Retroviruses*, **21**, 343–357. URL http://dx.doi.org/10.1089/aid.2005.21.343. (doi:10.1089/aid.2005.21.343)

[143] R. C. Hunt. Mircobiology and immunology online. University of South Carolina, School of Medicine (http://pathmicro.med.sc.edu/book/welcome.htm).

[144] A. R. Moss and P. Bacchetti. 1989. Natural history of HIV infection. *AIDS*, **3**, 55–61.

[145] G. F. Lemp, S. F. Payne, G. W. Rutherford, N. A. Hessol, W. Winkelstein, J. A. Wiley, A. R. Moss, R. E. Chaisson, R. T. Chen and D. W. Feigal. 1990. Projections of AIDS morbidity and mortality in San Francisco. *JAMA*, **263**, 1497–1501.

[146] J. O. Kahn and B. D. Walker. 1998. Acute human immunodeficiency virus type 1 infection. *N Engl J Med*, **339**, 33–39. URL http://dx.doi.org/10.1056/NEJM199807023390107. (doi:10.1056/NEJM199807023390107)

[147] B. Tindall and D. A. Cooper. 1991. Primary HIV infection: host responses and intervention strategies. *AIDS*, **5**, 1–14.

[148] J. M. Coffin. 1996. Hiv viral dynamics. *AIDS*, **10 Suppl 3**, S75–S84.

[149] S. G. Lim, A. Condez, C. A. Lee, M. A. Johnson, C. Elia and L. W. Poulter. 1993. Loss of mucosal CD4 lymphocytes is an early feature of HIV infection. *Clin Exp Immunol*, **92**, 448–454.

[150] J. Chinen and W. T. Shearer. 2002. Molecular virology and immunology of HIV infection. *J Allergy Clin Immunol*, **110**, 189–198.

[151] J. Stebbing, B. Gazzard and D. C. Douek. 2004. Where does HIV live? *N Engl J Med*, **350**, 1872–1880. URL http://dx.doi.org/10.1056/NEJMra032395. (doi:10.1056/NEJMra032395)

[152] H. Jiang and L. Chess. 2004. An integrated view of suppressor T cell subsets in immunoregulation. *J Clin Invest*, **114**, 1198–1208. URL http://dx.doi.org/10.1172/JCI23411. (doi:10.1172/JCI23411)

[153] G. S. Ogg, X. Jin, S. Bonhoeffer, P. R. Dunbar, M. A. Nowak, S. Monard, J. P. Segal, Y. Cao, S. L. Rowland-Jones, V. Cerundolo, A. Hurley, M. Markowitz, D. D. Ho, D. F. Nixon and A. J. McMichael. 1998. Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA. *Science*, **279**, 2103–2106.

[154] G. F. Burton, B. F. Keele, J. D. Estes, T. C. Thacker and S. Gartner. 2002. Follicular dendritic cell contributions to HIV pathogenesis. *Semin Immunol*, **14**, 275–284.

[155] R. Ahmed and D. Gray. 1996. Immunological memory and protective immunity: understanding their relation. *Science*, **272**, 54–60.

[156] R. W. Dutton, L. M. Bradley and S. L. Swain. 1998. T cell memory. *Annu Rev Immunol*, **16**, 201–223. URL http://dx.doi.org/10.1146/annurev.immunol.16.1.201. (doi:10.1146/annurev.immunol.16.1.201)

[157] J. J. Chen and M. W. Cloyd. 1999. The potential importance of HIV-induction of lymphocyte homing to lymph nodes. *Int Immunol*, **11**, 1591–1594.

[158] B. Ahr, V. Robert-Hebmann, C. Devaux and M. Biard-Piechaczyk. 2004. Apoptosis of uninfected cells induced by HIV envelope glycoproteins. *Retrovirology*, **1**, 12. URL http://dx.doi.org/10.1186/1742-4690-1-12. (doi:10.1186/1742-4690-1-12)

[159] G. H. Holm, C. Zhang, P. R. Gorry, K. Peden, D. Schols, E. De Clercq and D. Gabuzda. 2004. Apoptosis of bystander T cells induced by human immunodeficiency virus type 1 with increased envelope/receptor affinity and coreceptor binding site exposure. *J Virol*, **78**, 4541–4551.

[160] E. O. Freed and M. A. Martin. HIVs and their replication. in: , eds. , 4th ed. philadelphia: , williams, and wilkins, 2001:1971-2041. In D. M. Knipe, P. M. Howley, D. Griffin, R. A. Lamb, B. Roizman, M. A. Martin and S. E. Straus, editors, *Fields Virology*. Lippincott, 4th edition, 2001.

[161] E. O. Freed. 2001. HIV-1 replication. *Somat Cell Mol Genet*, **26**, 13–33.

[162] J. M. Coffin., S. H. Hughes and H. E. Varmus. *Retroviruses*. Cold Spring Harbor Laboratory Press, 1993.

[163] C. A. Janeway, P. Travers, Mark Walport and Mark J Shlomchik. *Immunobiology: The Immune System in Health and Disease*. Garland Publishing, New York, 5th edition, 2001.

[164] A. D. Frankel and J. A. Young. 1998. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem*, **67**, 1–25. URL http://dx.doi.org/10.1146/annurev.biochem.67.1.1. (doi:10.1146/annurev.biochem.67.1.1)

[165] BioAfrica: http://www.bioafrica.net/.

[166] CCBC: http://student.ccbcmd.edu/courses/bio141/.

[167] P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski and W. A. Hendrickson. 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, **393**, 648–659. URL http://dx.doi.org/10.1038/31405. (doi:10.1038/31405)

[168] D. M. Eckert and P. S. Kim. 2001. Mechanisms of viral membrane fusion and its inhibition. *Annu Rev Biochem*, **70**, 777–810. URL http://dx.doi.org/10.1146/annurev.biochem.70.1.777. (doi:10.1146/annurev.biochem.70.1.777)

[169] P. Poignard, E. O. Saphire, P. W. Parren and D. R. Burton. 2001. gp120: Biologic aspects of structural features. *Annu Rev Immunol*, **19**, 253–274. URL http://dx.doi.org/10.1146/annurev.immunol.19.1.253. (doi:10.1146/annurev.immunol.19.1.253)

[170] C. D. Rizzuto, R. Wyatt, N. Hernndez-Ramos, Y. Sun, P. D. Kwong, W. A. Hendrickson and J. Sodroski. 1998. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science*, **280**, 1949–1953.

[171] M. Dean, M. Carrington, C. Winkler, G. A. Huttley, M. W. Smith, R. Allikmets, J. J. Goedert, S. P. Buchbinder, E. Vittinghoff, E. Gomperts, S. Donfield, D. Vlahov, R. Kaslow, A. Saah, C. Rinaldo, R. Detels and S. J. O'Brien. 1996. Genetic Restriction of HIV-1 Infection and Progression to AIDS by a Deletion Allele of the CKR5 Structural Gene. *Science*, **273**, 1856–1862. URL http://www.sciencemag.org/content/273/5283/1856.abstract. (doi:10.1126/science.273.5283.1856)

[172] Y. Huang, W. A. Paxton, S. M. Wolinsky, A. U. Neumann, L. Zhang, T. He, S. Kang, D. Ceradini, Z. Jin, K. Yazdanbakhsh, K. Kunstman, D. Erickson, E. Dragon, N. R. Landau, J. Phair, D. D. Ho and R. A. Koup. 1996. The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nat Med*, **2**, 1240–1243.

[173] W. A. Paxton and R. A. Koup. 1997. Mechanisms of resistance to HIV infection. *Springer Semin Immunopathol*, **18**, 323–340.

[174] C. D. Weiss. 2003. HIV-1 gp41: mediator of fusion and target for inhibition. *AIDS Rev*, **5**, 214–221.

[175] M. P. Sherman and W. C. Greene. 2002. Slipping through the door: HIV entry into the nucleus. *Microbes Infect*, **4**, 67–73.

[176] H. M. Temin. 1993. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. *Proc Natl Acad Sci U S A*, **90**, 6900–6903.

[177] V.K. Pathak and W-S. Hu. 1997. ”might as well jump!” Template switching by retroviral reverse transcriptase, defective genome formation and recombination. *Semin Virol*, **8**, 141–150.

[178] L. M. Mansky and H. M. Temin. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol*, **69**, 5087–5094.

[179] B. D. Preston, B. J. Poiesz and L. A. Loeb. 1988. Fidelity of HIV-1 reverse transcriptase. *Science*, **242**, 1168–1171.

[180] J. D. Roberts, K. Bebenek and T. A. Kunkel. 1988. The accuracy of reverse transcriptase from HIV-1. *Science*, **242**, 1171–1173.

[181] D. D. Ho. 1997. Perspectives series: host/pathogen interactions. Dynamics of HIV-1 replication in vivo. *J Clin Invest*, **99**, 2565–2567. URL http://dx.doi.org/10.1172/JCI119443. (doi:10.1172/JCI119443)

[182] M. Alfano and G. Poli. 2004. The HIV life cycle: Multiple targets for antiretroviral agents. *Drug Design Reviews - Online*, **1**, 83–92.

[183] M. I. Bukrinsky, S. Haggerty, M. P. Dempsey, N. Sharova, A. Adzhubel, L. Spitz, P. Lewis, D. Goldfarb, M. Emerman and M. Stevenson. 1993. A nuclear localization signal within HIV-1 matrix protein that governs infection of non-dividing cells. *Nature*, **365**, 666–669. URL http://dx.doi.org/10.1038/365666a0. (doi:10.1038/365666a0)

[184] R. A. Fouchier, B. E. Meyer, J. H. Simon, U. Fischer and M. H. Malim. 1997. HIV-1 infection of non-dividing cells: evidence that the amino-terminal basic region of the viral matrix protein is important for Gag processing but not for post-entry nuclear import. *EMBO J*, **16**, 4531–4539. URL http://dx.doi.org/10.1093/emboj/16.15.4531. (doi:10.1093/emboj/16.15.4531)

[185] P. O. Brown. Integration. In J.M. Coffin, S.H. Hughes and H.E. Varmus, editors, *Retroviruses*, pages 161–203, 1997.

[186] D. F. Purcell and M. A. Martin. 1993. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J Virol*, **67**, 6365–6378.

[187] V. W. Pollard and M. H. Malim. 1998. The HIV-1 Rev protein. *Annu Rev Microbiol*, **52**, 491–532. URL http://dx.doi.org/10.1146/annurev.micro.52.1.491. (doi:10.1146/annurev.micro.52.1.491)

[188] M. Hill, G. Tachedjian and J. Mak. 2005. The packaging and maturation of the HIV-1 Pol proteins. *Curr HIV Res*, **3**, 73–85.

[189] W. Zhou, L. J. Parent, J. W. Wills and M. D. Resh. 1994. Identification of a membrane-binding domain within the amino-terminal region of human immunodeficiency virus type 1 Gag protein which interacts with acidic phospholipids. *J Virol*, **68**, 2556–2569.

[190] S. Oroszlan and R. B. Luftig. 1990. Retroviral proteinases. *Curr Top Microbiol Immunol*, **157**, 153–185.

[191] N. E. Kohl, E. A. Emini, W. A. Schleif, L. J. Davis, J. C. Heimbach, R. A. Dixon, E. M. Scolnick and I. S. Sigal. 1988. Active human immunodeficiency virus protease is required for viral infectivity. *Proc Natl Acad Sci U S A*, **85**, 4686–4690.

[192] D. Warnke, J. Barreto and Z. Temesgen. 2007. Antiretroviral drugs. *J Clin Pharmacol*, **47**, 1570–1579. URL http://dx.doi.org/10.1177/0091270007308034. (doi:10.1177/0091270007308034)

[193] C. J. Cohen. 2006. Successful HIV treatment: lessons learned. *J Manag Care Pharm*, **12**, S6–11.

[194] M. A. Thompson, J. A. Aberg, P. Cahn, J. S. G. Montaner, G. Rizzardini, A. Telenti, J. M. Gatell, H. F. G"unthard, S. M. Hammer, M. S. Hirsch, D. M. Jacobsen, P. Reiss, D. D. Richman, P. A. Volberding, P. Yeni, R. T. Schooley and International AIDS Society-USA. 2010. Antiretroviral treatment of adult HIV infection: 2010 recommendations of the International AIDS Society-USA panel. *JAMA*, **304**, 321–333.

[195] C. T. Fang, Y. Y. Chang, H. M. Hsu, S. J. Twu, K. T. Chen, C. C. Lin, L. Y L Huang, M. Y. Chen, J. S. Hwang, J. D. Wang and C. Y. Chuang. 2007. Life expectancy of patients with newly-diagnosed hiv infection in the era of highly active antiretroviral therapy. *QJM*, **100**, 97–105. URL http://dx.doi.org/10.1093/qjmed/hcl141. (doi:10.1093/qjmed/hcl141)

[196] E. De Clercq. 2009. The history of antiretrovirals: key discoveries over the past 25 years. *Rev Med Virol*, **19**, 287–299. URL http://dx.doi.org/10.1002/rmv.624. (doi:10.1002/rmv.624)

[197] N. Sluis-Cremer, N. A. Temiz and I. Bahar. 2004. Conformational changes in HIV-1 reverse transcriptase induced by nonnucleoside reverse transcriptase inhibitor binding. *Curr HIV Res*, **2**, 323–332.

[198] A. L. Parrill. 2003. HIV-1 integrase inhibition: binding sites, structure activity relationships and future perspectives. *Curr Med Chem*, **10**, 1811–1824.

[199] R. T. Steigbigel, D. A. Cooper, P. N. Kumar, J. E. Eron, M. Schechter, M. Markowitz, M. R. Loutfy, J. L. Lennox, J. M. Gatell, J. K. Rockstroh, C. Katlama, P. Yeni, A. Lazzarin, B. Clotet, J. Zhao, J. Chen, D. M. Ryan, R. R. Rhodes, J. A. Killar, L. R. Gilde, K. M. Strohmaier, A. R. Meibohm, M. D. Miller, D. J.

Hazuda, M. L. Nessly, M. J. DiNubile, R. D. Isaacs, B. Nguyen, H. Teppler and B. E. N. C. H. M. R. K. Study Teams. 2008. Raltegravir with optimized background therapy for resistant HIV-1 infection. *N Engl J Med*, **359**, 339–354. URL http://dx.doi.org/10.1056/NEJMoa0708975. (doi:10.1056/NEJMoa0708975)

[200] V. A. Johnson, F. Brun-Vezinet, B. Clotet, H. F. Gunthard, D. R. Kuritzkes, D. Pillay, J. M. Schapiro and D. D. Richman. 2009. Update of the drug resistance mutations in HIV-1: December 2009. *Top HIV Med*, **17**, 138–145.

[201] R. E. Myers and D. Pillay. 2008. Analysis of natural sequence variation and covariation in human immunodeficiency virus type 1 integrase. *J Virol*, **82**, 9228–9235. URL http://dx.doi.org/10.1128/JVI.01535-07. (doi:10.1128/JVI.01535-07)

[202] Z. G. Luo, J. J. Tan, Y. Zeng, C. X. Wang and L. M. Hu. 2010. Development of integrase inhibitors of quinolone acid derivatives for treatment of AIDS: an overview. *Mini Rev Med Chem*, **10**, 1046–1057.

[203] J. P. Lalezari, J. J. Eron, M. Carlson, C. Cohen, E. DeJesus, R. C. Arduino, J. E. Gallant, P. Volberding, R. L. Murphy, F. Valentine, E. L. Nelson, P. R. Sista, A. Dusek and J. M. Kilby. 2003. A phase II clinical study of the long-term safety and antiviral activity of enfuvirtide-based antiretroviral therapy. *AIDS*, **17**, 691–698. URL http://dx.doi.org/10.1097/01.aids.0000050825.06065.84. (doi:10.1097/01.aids.0000050825.06065.84)

[204] J. A. Levy. 2009. HIV pathogenesis: 25 years of progress and persistent challenges. *AIDS*, **23**, 147–160. URL http://dx.doi.org/10.1097/QAD.0b013e3283217f9f. (doi:10.1097/QAD.0b013e3283217f9f)

[205] P. Biswas, G. Tambussi and A. Lazzarin. 2007. Access denied? The status of co-receptor inhibition to counter HIV entry. *Expert Opin Pharmacother*, **8**, 923–933. URL http://dx.doi.org/10.1517/14656566.8.7.923. (doi:10.1517/14656566.8.7.923)

[206] D. D. Richman, S. J. Little, D. M. Smith, T. Wrin, C. Petropoulos and J. K. Wong. 2004. HIV evolution and escape. *Trans Am Clin Climatol Assoc*, **115**, 289–303.

[207] J. Durant, P. Clevenbergh, P. Halfon, P. Delgiudice, S. Porsin, P. Simonet, N. Montagne, C. A. Boucher, J. M. Schapiro and P. Dellamonica. 1999. Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. *Lancet*, **353**, 2195–2199.

[208] M. M. Kitahata, S. E. Van Rompaey and A. W. Shields. 2000. Physician experience in the care of HIV-infected persons is associated with earlier adoption of new antiretroviral therapy. *J Acquir Immune Defic Syndr*, **24**, 106–114.

[209] A. De Luca, A. Cingolani, S. Di Giambenedetto, M. P. Trotta, F. Baldini, M. G. Rizzo, A. Bertoli, G. Liuzzi, P. Narciso, R. Murri, A. Ammassari, C. F. Perno and A. Antinori. 2003. Variable prediction of antiretroviral treatment outcome by different systems for interpreting genotypic human immunodeficiency virus type 1 drug resistance. *J Infect Dis*, **187**, 1934–1943. URL http://dx.doi.org/10.1086/375355. (doi:10.1086/375355)

[210] M. C. F. Prosperi, A. Altmann, M. Rosen-Zvi, E. Aharoni, G. Borgulya, F. Bazso, A. Sönnerborg, E. Schülter, D. Struck, G. Ulivi, A. Vandamme, J. Vercauteren, M. Zazzi, EuResist and Virolab study groups. 2009. Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. *Antivir Ther*, **14**, 433–442.

[211] M. Zazzi, M. Prosperi, I. Vicenti, S. Di Giambenedetto, A. Callegaro, B. Bruzzone, F. Baldanti, A. Gonnelli, E. Boeri, E. Paolini, S. Rusconi, A. Giacometti, F. Maggiolo, S. Menzo, A. De Luca and A. R. C. A. Collaborative Group. 2009. Rules-based HIV-1 genotypic resistance interpretation systems predict 8 week and 24 week virological antiretroviral treatment outcome and benefit from drug potency weighting. *J Antimicrob Chemother*, **64**, 616–624. URL http://dx.doi.org/10.1093/jac/dkp252. (doi:10.1093/jac/dkp252)

[212] S. Rhee, W. J. Fessel, T. F. Liu, N. M. Marlowe, C. M. Rowland, R. A. Rode, A. Vandamme, K. van Laethem, F. Brun-Vezinet, V. Calvez, J. Taylor, L. Hurley, M. Horberg and R. W. Shafer. 2009. Predictive value of HIV-1 genotypic resistance test interpretation algorithms. *J Infect Dis*, **200**, 453–463. URL http://dx.doi.org/10.1086/600073. (doi:10.1086/600073)

[213] Z. V. Fox, A. M. Geretti, J. Kjaer, U. B. Dragsted, A. N. Phillips, J. Gerstoft, S. Staszewski, B. Clotet, V. von Wyl and J. D. Lundgren. 2007. The ability of four genotypic interpretation systems to predict virological response to ritonavir-boosted protease inhibitors. *AIDS*, **21**, 2033–2042. URL http://dx.doi.org/10.1097/QAD.0b013e32825a69e4. (doi:10.1097/QAD.0b013e32825a69e4)

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

[215] A. L. Perryman, J. Lin and J. A. McCammon. 2004. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs. *Protein Sci*, **13**, 1108–1123. URL http://dx.doi.org/110/ps.03468904. (doi:110/ps.03468904)

[216] A. Wlodawer and J. W. Erickson. 1993. Structure-based inhibitors of HIV-1 protease. *Annu Rev Biochem*, **62**, 543–585. URL http://dx.doi.org/10.1146/annurev.bi.62.070193.002551. (doi:10.1146/annurev.bi.62.070193.002551)

[217] B. M. Dunn, M. M. Goodenow, A. Gustchina and A. Wlodawer. 2002. Retroviral proteases. *Genome Biol*, **3**, REVIEWS3006.

[218] K. Strisovsky, U. Tessmer, J. Langner, J. Konvalinka and H. G. Kräusslich. 2000. Systematic mutational analysis of the active-site threonine of HIV-1 proteinase: rethinking the "fireman's grip" hypothesis. *Protein Sci*, **9**, 1631–1641.

[219] W. E. Harte, S. Swaminathan, M. M. Mansuri, J. C. Martin, I. E. Rosenberg and D. L. Beveridge. 1990. Domain communication in the dynamical structure of human immunodeficiency virus 1 protease. *Proc Natl Acad Sci U S A*, **87**, 8864–8868.

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

[221] A. Krohn, S. Redshaw, J. C. Ritchie, B. J. Graves and M. H. Hatada. 1991. Novel binding mode of highly potent HIV-proteinase inhibitors incorporating the (R)-hydroxyethylamine isostere. *J Med Chem*, **34**, 3340–3342.

[222] D. I. Freedberg, R. Ishima, J. Jacob, Y. Wang, I. Kustanovich, J. M. Louis and D. A. Torchia. 2002. Rapid structural fluctuations of the free HIV protease flaps in solution: Relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Sci*, **11**, 221–232. URL http://www.proteinscience.org/cgi/content/abstract/11/2/221. (doi:10.1110/ps.33202)

[223] L. K. Nicholson, T. Yamazaki, D. A. Torchia, S. Grzesiek, A. Bax, S. J. Stahl, J. D. Kaufman, P. T. Wingfield, P. Y. Lam and P. K. Jadhav. 1995. Flexibility and function in HIV-1 protease. *Nat Struct Biol*, **2**, 274–280.

[224] V. Zoete, O. Michielin and M. Karplus. 2002. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J Mol Biol*, **315**, 21–52. URL http://dx.doi.org/10.1006/jmbi.2001.5173. (doi:10.1006/jmbi.2001.5173)

[225] V. Hornak, A. Okur, R. C. Rizzo and C. Simmerling. 2006. HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. *J Am Chem Soc*, **128**, 2812–2813. URL http://dx.doi.org/10.1021/ja058211x. (doi:10.1021/ja058211x)

[226] R. Ishima and J. M. Louis. 2008. A diverse view of protein dynamics from NMR studies of HIV-1 protease flaps. *Proteins: Structure, Function, and Bioinformatics*, **70**, 1408–1415. URL http://dx.doi.org/10.1002/prot.21632. (doi:10.1002/prot.21632)

[227] A. L. Perryman, J. Lin and J. A. McCammon. 2006. Restrained molecular dynamics simulations of HIV-1 protease: the first step in validating a new target for drug design. *Biopolymers*, **82**, 272–284. URL http://dx.doi.org/10.1002/bip.20497. (doi:10.1002/bip.20497)

[228] M. Layten, V. Hornak and C. Simmerling. 2006. The Open Structure of a Multi-Drug-Resistant HIV-1 Protease is Stabilized by Crystal Packing Contacts. *Journal of the American Chemical Society*, **128**, 13360–13361. ISSN 0002-7863. URL http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ja065133k. (doi:10.1021/ja065133k)

[229] S. K. Sadiq and G. De Fabritiis. 2010. Explicit solvent dynamics and energetics of HIV-1 protease flap-opening and closing. *Proteins: Structure, Function and Bioinformatics*.

[230] V. Hornak, A.Okur, R. C. Rizzo and C. Simmerling. 2006. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc Natl Acad Sci U S A*, **103**, 915–920. URL http://dx.doi.org/10.1073/pnas.0508452103. (doi:10.1073/pnas.0508452103)

[231] C. Chang, T. Shen, J. Trylska, V. Tozzini and J. A. McCammon. 2006. Gated binding of ligands to HIV-1 protease: Brownian dynamics simulations in a coarse-grained model. *Biophys J*, **90**, 3880–3885. URL http://dx.doi.org/10.1529/biophysj.105.074575. (doi:10.1529/biophysj.105.074575)

[232] C. A. Chang, J. Trylska, V. Tozzini and J. A. McCammon. 2007. Binding pathways of ligands to HIV-1 protease: coarse-grained and atomistic simulations. *Chem Biol Drug Des*, **69**, 5–13. URL http://dx.doi.org/10.1111/j.1747-0285.2007.00464.x. (doi:10.1111/j.1747-0285.2007.00464.x)

[233] P. L. Darke, C. T. Leu, L. J. Davis, J. C. Heimbach, R. E. Diehl, W. S. Hill, R. A. Dixon and I. S. Sigal. 1989. Human immunodeficiency virus protease. Bacterial

expression and characterization of the purified aspartic protease. *J Biol Chem*, **264**, 2307–2312.

[234] S. C. Pettit, S. Gulnik, L. Everitt and A. H. Kaplan. 2003. The dimer interfaces of protease and extra-protease domains influence the activation of protease and the specificity of GagPol cleavage. *J Virol*, **77**, 366–374.

[235] S. C. Pettit, L. E. Everitt, S. Choudhury, B. M. Dunn and A. H. Kaplan. 2004. Initial cleavage of the human immunodeficiency virus type 1 GagPol precursor by its activated protease occurs by an intramolecular mechanism. *J Virol*, **78**, 8477–8485. URL http://dx.doi.org/10.1128/JVI.78.16.8477-8485.2004. (doi:10.1128/JVI.78.16.8477-8485.2004)

[236] L. J. Hyland, T. A. Tomaszek and T. D. Meek. 1991. Human immunodeficiency virus-1 protease. 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanism. *Biochemistry*, **30**, 8454–8463.

[237] L. J. Hyland, T. A. Tomaszek, G. D. Roberts, S. A. Carr, V. W. Magaard, H. L. Bryan, S. A. Fakhoury, M. L. Moore, M. D. Minnich and J. S. Culp. 1991. Human immunodeficiency virus-1 protease. 1. Initial velocity studies and kinetic characterization of reaction intermediates by 18O isotope exchange. *Biochemistry*, **30**, 8441–8453.

[238] D.C. Chatfield and B. R. Brooks. 1995. HIV-1 protease cleavage mechanism elucidated with molecular dynamics simulation. *J Am Chem Soc*, **117**, 5561–5572.

[239] S. Piana, P. Carloni and M. Parrinello. 2002. Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease. *J Mol Biol*, **319**, 567–583. URL http://dx.doi.org/10.1016/S0022-2836(02)00301-7. (doi:10.1016/S0022-2836(02)00301-7)

[240] J. Trylska, P. Grochowski and J. A. McCammon. 2004. The role of hydrogen bonding in the enzymatic reaction catalyzed by HIV-1 protease. *Protein Sci*, **13**, 513–528. URL http://dx.doi.org/10.1110/ps.03372304. (doi:10.1110/ps.03372304)

[241] S. Bihani, A. Das, V. Prashar, J-L. Ferrer and M. V. Hosur. 2009. X-ray structure of HIV-1 protease in situ product complex. *Proteins*, **74**, 594–602. URL http://dx.doi.org/10.1002/prot.22174. (doi:10.1002/prot.22174)

[242] A. Das, S. Mahale, V. Prashar, S. Bihani, J-L. Ferrer and M. V. Hosur. 2010. X-ray snapshot of hiv-1 protease in action: observation of tetrahedral intermediate and short ionic hydrogen bond sihb with catalytic aspartate. *J Am*

Chem Soc, **132**, 6366–6373. URL http://dx.doi.org/10.1021/ja100002b. (doi:10.1021/ja100002b)

[243] A. Brik and C. Wong. 2003. HIV-1 protease: mechanism and drug discovery. *Org Biomol Chem*, **1**, 5–14.

[244] T. Hofmann, R. S. Hodges and M. N. James. 1984. Effect of pH on the activities of penicillopepsin and Rhizopus pepsin and a proposal for the productive substrate binding mode in penicillopepsin. *Biochemistry*, **23**, 635–643.

[245] T. Yamazaki, L. K. Nicholson, P. Wingfield, S. J. Stahl, J. D. Kaufman, C. J. Eyermann, C. N. Hodge, P. Y. S. Lam and D. A. Torchia. 1994. NMR and X-ray Evidence That the HIV Protease Catalytic Aspartyl Groups Are Protonated in the Complex Formed by the Protease and a Non-Peptide Cyclic Urea-Based Inhibitor. . *J. Am. Chem. Soc.*, **116**, 10791–10792.

[246] Y. Wang, D.I. Freedberg, T. Yamazaki, P.T. Wingfield, S.J. Stahl, J.D. Kaufman, Y. Kiso and D.A. Torchia. 1996. Solution NMR Evidence That the HIV-1 Protease Catalytic Aspartyl Groups Have Different Ionization States in the Complex Formed with the Asymmetric Drug KNI-272. *Biochemistry*, **35**, 9945–9950. URL http://pubs.acs.org/doi/abs/10.1021/bi961268z. PMID: 8756455. (doi:10.1021/bi961268z)

[247] K.Y. Nam, B.H. Chang, C.K. Han, S.G. Ahn and K.T. No. 2003. Investigation of the Protonated State of HIV-1 Protease Active Site. *Bull Korean Chem Soc*, **24**, 817–823.

[248] S. Piana, D. Sebastiani, P. Carloni and M. Parrinello. 2001. Ab initio molecular dynamics-based assignment of the protonation state of pepstatin A/HIV-1 protease cleavage site. *J Am Chem Soc*, **123**, 8730–8737.

[249] R. Smith, I. M. Brereton, R. Y. Chai and S. B. Kent. 1996. Ionization states of the catalytic residues in HIV-1 protease. *Nat Struct Biol*, **3**, 946–950.

[250] W. Wang and P. A. Kollman. 2000. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. *J Mol Biol*, **303**, 567–582. URL http://dx.doi.org/10.1006/jmbi.2000.4057. (doi:10.1006/jmbi.2000.4057)

[251] P. L. Darke, R. F. Nutt, S. F. Brady, V. M. Garsky, T. M. Ciccarone, C. T. Leu, P. K. Lumma, R. M. Freidinger, D. F. Veber and I. S. Sigal. 1988. HIV-1 protease specificity of peptide cleavage is sufficient for processing of gag and pol polyproteins. *Biochem Biophys Res Commun*, **156**, 297–303.

[252] J. Tozser, A. Gustchina, I. T. Weber, I. Blaha, E. M. Wondrak and S. Oroszlan. 1991. Studies on the role of the S4 substrate binding site of HIV proteinases. *FEBS Lett*, **279**, 356–360.

[253] X. Li, H. Hu and L. Shu. 2010. Predicting human immunodeficiency virus protease cleavage sites in nonlinear projection space. *Mol Cell Biochem*, **339**, 127–133. URL http://dx.doi.org/10.1007/s11010-009-0376-y. (doi:10.1007/s11010-009-0376-y)

[254] B. Niu, L. Lu, L. Liu, T. Gu, K. Feng, W. Lu and Y. Cai. 2009. HIV-1 protease cleavage site prediction based on amino acid property. *J Comput Chem*, **30**, 33–39. URL http://dx.doi.org/10.1002/jcc.21024. (doi:10.1002/jcc.21024)

[255] Hasan Ogul. 2009. Variable context Markov chains for HIV protease cleavage site prediction. *Biosystems*, **96**, 246–250. URL http://dx.doi.org/10.1016/j.biosystems.2009.03.001. (doi:10.1016/j.biosystems.2009.03.001)

[256] M. Tritel and M. D. Resh. 2000. Kinetic analysis of human immunodeficiency virus type 1 assembly reveals the presence of sequential intermediates. *J Virol*, **74**, 5845–5855.

[257] K. Wiegers, G. Rutter, H. Kottler, U. Tessmer, H. Hohenberg and H. G. Kräusslich. 1998. Sequential steps in human immunodeficiency virus particle maturation revealed by alterations of individual Gag polyprotein cleavage sites. *J Virol*, **72**, 2846–2854.

[258] A. Wlodawer. 2002. Rational approach to AIDS drug design through structural biology. *Annu Rev Med*, **53**, 595–614. URL http://dx.doi.org/10.1146/annurev.med.53.052901.131947. (doi:10.1146/annurev.med.53.052901.131947)

[259] P. Y. Lam, P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner and Y. N. Wong. 1994. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science*, **263**, 380–384.

[260] H. L. Sham, D. J. Kempf, A. Molla, K. C. Marsh, G. N. Kumar, C. M. Chen, W. Kati, K. Stewart, R. Lal, A. Hsu, D. Betebenner, M. Korneyeva, S. Vasavanonda, E. McDonald, A. Saldivar, N. Wideburg, X. Chen, P. Niu, C. Park, V. Jayanti, B. Grabowski, G. R. Granneman, E. Sun, A. J. Japour, J. M. Leonard, J. J. Plattner and D. W. Norbeck. 1998. ABT-378, a highly potent inhibitor of the human immunodeficiency virus protease. *Antimicrob Agents Chemother*, **42**, 3218–3224.

[261] J. M. Llibre. 2009. First-line boosted protease inhibitor-based regimens in treatment-naive HIV-1-infected patients–making a good thing better. *AIDS Rev*, **11**, 215–222.

[262] C. E. Patterson, R. Seetharam, C. A. Kettner and Y. S. Cheng. 1992. Human immunodeficiency virus type 1 and type 2 protease monomers are functionally interchangeable in the dimeric enzymes. *J Virol*, **66**, 1228–1231.

[263] A. J. Barrett, N. D. Rawlings and J. F. Woessner. *Handbook of Proteolytic Enzymes*. London: Academic Press, 1998.

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

[265] R. F. Shinazi and B. A. Larder J. W. Mellors. 1993. Mutations in retroviral genes associated with drug resistance. *Intl Antiviral News*, **5**, 129–142.

[266] R. W. Shafer, S. Rhee, D. Pillay, V. Miller, P. Sandstrom, J. M. Schapiro, D. R. Kuritzkes and D. Bennett. 2007. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. *AIDS*, **21**, 215–223. URL http://dx.doi.org/10.1097/QAD.0b013e328011e691. (doi:10.1097/QAD.0b013e328011e691)

[267] B. C. Logsdon, J. F. Vickrey, P. Martin, G. Proteasa, J. I. Koepke, S. R. Terlecky, Z. Wawrzak, M. A. Winters, T. C. Merigan and L. C. Kovari. 2004. Crystal structures of a multidrug-resistant human immunodeficiency virus type 1 protease reveal an expanded active-site cavity. *J Virol*, **78**, 3123–3132.

[268] H. Tamamura and N. Fujii. 2004. Two orthogonal approaches to overcome multidrug resistant HIV-1s: development of protease inhibitors and entry inhibitors based on CXCR4 antagonists. *Curr Drug Targets Infect Disord*, **4**, 103–110.

[269] Omar Haq, Ronald M Levy, Alexandre V Morozov and Michael Andrec. 2009. Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease. *BMC Bioinformatics*, **10 Suppl 8**, S10. URL http://dx.doi.org/10.1186/1471-2105-10-S8-S10. (doi:10.1186/1471-2105-10-S8-S10)

[270] A. J. Brown, B. T. Korber and J. H. Condra. 1999. Associations between amino acids in the evolution of HIV type 1 protease sequences under indinavir therapy. *AIDS Res Hum Retroviruses*, **15**, 247–253. URL http://dx.doi.org/10.1089/088922299311420. (doi:10.1089/088922299311420)

[271] J. H. Condra, D. J. Holder, W. A. Schleif, O. M. Blahy, R. M. Danovich, L. J. Gabryelski, D. J. Graham, D. Laird, J. C. Quintero, A. Rhodes, H. L. Robbins, E. Roth, M. Shivaprakash, T. Yang, J. A. Chodakewitz, P. J. Deutsch, R. Y. Leavitt, F. E. Massari, J. W. Mellors, K. E. Squires, R. T. Steigbigel, H. Teppler and E. A. Eminimolla. 1996. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. *J Virol*, **70**, 8270–8276.

[272] A. Molla, M. Korneyeva, Q. Gao, S. Vasavanonda, P. J. Schipper, H. M. Mo, M. Markowitz, T. Chernyavskiy, P. Niu, N. Lyons, A. Hsu, G. R. Granneman, D. D. Ho, C. A. Boucher, J. M. Leonard, D. W. Norbeck and D. J. Kempf. 1996. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat Med*, **2**, 760–766.

[273] A. K. Patick, M. Duran, Y. Cao, D. Shugarts, M. R. Keller, E. Mazabel, M. Knowles, S. Chapman, D. R. Kuritzkes and M. Markowitz. 1998. Genotypic and phenotypic characterization of human immunodeficiency virus type 1 variants isolated from patients treated with the protease inhibitor nelfinavir. *Antimicrob Agents Chemother*, **42**, 2637–2644.

[274] M. Prabu-Jeyabalan, N. M. King, E. A. Nalivaika, G. Heilek-Snyder, N. Cammack and C. A. Schiffer. 2006. Substrate envelope and drug resistance: crystal structure of RO1 in complex with wild-type human immunodeficiency virus type 1 protease. *Antimicrob Agents Chemother*, **50**, 1518–1521. URL http://dx.doi.org/10.1128/AAC.50.4.1518-1521.2006. (doi:10.1128/AAC.50.4.1518-1521.2006)

[275] S. Chellappan, V. Kairys, M. X. Fernandes, C. A. Schiffer and M. K. Gilson. 2007. Evaluation of the substrate envelope hypothesis for inhibitors of HIV-1 protease. *Proteins*, **68**, 561–567. URL http://dx.doi.org/10.1002/prot.21431. (doi:10.1002/prot.21431)

[276] M. N. L. Nalam, A. Ali, M. D. Altman, G. S. K. K. Reddy, S. Chellappan, V. Kairys, A. Ozen, H. Cao, M. K. Gilson, B. Tidor, T. M. Rana and C. A. Schiffer. 2010. Evaluating the substrate-envelope hypothesis: structural analysis of novel hiv-1 protease inhibitors designed to be robust against drug resistance. *J Virol*, **84**, 5368–5378. URL http://dx.doi.org/10.1128/JVI.02531-09. (doi:10.1128/JVI.02531-09)

[277] E. Lefebvre and C. A. Schiffer. 2008. Resilience to resistance of HIV-1 protease inhibitors: profile of darunavir. *AIDS Rev*, **10**, 131–142.

[278] L. Hong, X. C. Zhang, J. A. Hartsuck and J. Tang. 2000. Crystal structure of an in vivo HIV-1 protease mutant in complex with saquinavir: insights into the

mechanisms of drug resistance. *Protein Sci*, **9**, 1898–1904. URL http://dx.doi.org/10.1110/ps.9.10.1898. (doi:10.1110/ps.9.10.1898)

[279] O. Aruksakunwong, P. Wolschann, S. Hannongbua and P. Sompornpisut. 2006. Molecular dynamic and free energy studies of primary resistance mutations in HIV-1 protease-ritonavir complexes. *J Chem Inf Model*, **46**, 2085–2092. URL http://dx.doi.org/10.1021/ci060090c. (doi:10.1021/ci060090c)

[280] J. Zhang, T. Hou, W. Wang and J. S. Liu. 2010. Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. *Proc Natl Acad Sci U S A*, **107**, 1321–1326. URL http://dx.doi.org/10.1073/pnas.0907304107. (doi:10.1073/pnas.0907304107)

[281] J. C. Clemente, R. Hemrajani, L. E. Blum, M. M. Goodenow and B. M. Dunn. 2003. Secondary mutations M36I and A71V in the human immunodeficiency virus type 1 protease can provide an advantage for the emergence of the primary mutation D30N. *Biochemistry*, **42**, 15029–15035. URL http://dx.doi.org/10.1021/bi035701y. (doi:10.1021/bi035701y)

[282] J. C. Clemente, R. E. Moose, R. Hemrajani, L. R. S. Whitford, L. Govindasamy, R. Reutzel, R. McKenna, M. Agbandje-McKenna, M. M. Goodenow and B. M. Dunn. 2004. Comparing the accumulation of active- and nonactive-site mutations in the HIV-1 protease. *Biochemistry*, **43**, 12141–12151. URL http://dx.doi.org/10.1021/bi049459m. (doi:10.1021/bi049459m)

[283] J. R. Collins, S. K. Burt and J. W. Erickson. 1995. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. *Nat Struct Biol*, **2**, 334–338.

[284] S. W. Rick, I. A. Topol, J. W. Erickson and S. K. Burt. 1998. Molecular mechanisms of resistance: free energy calculations of mutation effects on inhibitor binding to HIV-1 protease. *Protein Sci*, **7**, 1750–1756. URL http://dx.doi.org/10.1002/pro.5560070809. (doi:10.1002/pro.5560070809)

[285] J. E. Foulkes-Murzycki, W. R. P. Scott and C. A. Schiffer. 2007. Hydrophobic sliding: a possible mechanism for drug resistance in human immunodeficiency virus type 1 protease. *Structure*, **15**, 225–233. URL http://dx.doi.org/10.1016/j.str.2007.01.006. (doi:10.1016/j.str.2007.01.006)

[286] K. S. Sadiq, S. Wan and P. V. Coveney. 2007. Insights into a Mutation-Assisted Lateral Drug Escape Mechanism from the HIV-1 Protease Active Site. *Biochemistry*, **46**, 14865–14877. URL http://pubs.acs.org/doi/abs/10.1021/bi700864p. (doi:10.1021/bi700864p)

[287] M. A. McCarrick and P. Kollman. 1994. Use of molecular dynamics and free energy perturbation calculations in anti-human immunodeficiency virus drug design. *Methods Enzymol*, **241**, 370–384.

[288] I. Stoica, S.K. Sadiq and P.V. Coveney. 2008. Rapid and Accurate Prediction of Binding Free Energies for Saquinavir-Bound HIV-1 Proteases. *Journal of the American Chemical Society*, **130**, 2639–2648. ISSN 0002-7863. URL http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ja0779250. (doi:10.1021/ja0779250)

[289] G. Hu, T. Zhu, S. Zhang, D. Wang and Q. Zhang. 2010. Some insights into mechanism for binding and drug resistance of wild type and I50V V82A and I84V mutations in HIV-1 protease with GRL-98065 inhibitor from molecular dynamic simulations. *Eur J Med Chem*, **45**, 227–235. URL http://dx.doi.org/10.1016/j.ejmech.2009.09.048. (doi:10.1016/j.ejmech.2009.09.048)

[290] M. Parera, G. Fernàndez, B. Clotet and M. A. Martínez. 2007. HIV-1 protease catalytic efficiency effects caused by random single amino acid substitutions. *Mol Biol Evol*, **24**, 382–387. URL http://dx.doi.org/10.1093/molbev/msl168. (doi:10.1093/molbev/msl168)

[291] J. Martinez-Picado, A. V. Savara, L. Sutton and R. T. D'Aquila. 1999. Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. *J Virol*, **73**, 3744–3752.

[292] M. Nijhuis, R. Schuurman, D. de Jong, J. Erickson, E. Gustchina, J. Albert, P. Schipper, S. Gulnik and C. A. Boucher. 1999. Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. *AIDS*, **13**, 2349–2359.

[293] A. Fehér, I. T. Weber, P. Bagossi, P. Boross, B. Mahalingam, J. M. Louis, T. D. Copeland, I. Y. Torshin, R. W. Harrison and J. Tözsér. 2002. Effect of sequence polymorphism and drug resistance on two HIV-1 Gag processing sites. *Eur J Biochem*, **269**, 4114–4120.

[294] S. C. Pettit, G. J. Henderson, C. A. Schiffer and R. Swanstrom. 2002. Replacement of the P1 amino acid of human immunodeficiency virus type 1 Gag processing sites can inhibit or enhance the rate of cleavage by the viral protease. *J Virol*, **76**, 10226–10233.

[295] R. M. Kagan, M. D. Shenderovich, P. N. R. Heseltine and K. Ramnarayan. 2005. Structural analysis of an HIV-1 protease I47A mutant resistant to the protease

inhibitor lopinavir. *Protein Sci*, **14**, 1870–1878. URL http://dx.doi.org/10.1110/ps.051347405. (doi:10.1110/ps.051347405)

[296] J. Yanchunas, D. R. Langley, L. Tao, R. E. Rose, J. Friborg, R. J. Colonno and M. L. Doyle. 2005. Molecular basis for increased susceptibility of isolates with atazanavir resistance-conferring substitution I50L to other protease inhibitors. *Antimicrob Agents Chemother*, **49**, 3825–3832. URL http://dx.doi.org/10.1128/AAC.49.9.3825-3832.2005. (doi:10.1128/AAC.49.9.3825-3832.2005)

[297] I. Malet, B. Roquebert, C. Dalban, M. Wirden, B. Amellal, R. Agher, A. Simon, C. Katlama, D. Costagliola, V. Calvez and A. Marcelin. 2007. Association of Gag cleavage sites to protease mutations and to virological response in HIV-1 treated patients. *J Infect*, **54**, 367–374. URL http://dx.doi.org/10.1016/j.jinf.2006.06.012. (doi:10.1016/j.jinf.2006.06.012)

[298] E. Dam, R. Quercia, B. Glass, D. Descamps, O. Launay, X. Duval, H. Kräusslich, A. J. Hance, F. Clavel and A. N. R. S. 109 Study Group. 2009. Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. *PLoS Pathog*, **5**, e1000345.

[299] M. Prabu-Jeyabalan, E. A. Nalivaika, N. M. King and C. A. Schiffer. 2004. Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease. *J Virol*, **78**, 12446–12454. URL http://dx.doi.org/10.1128/JVI.78.22.12446-12454.2004. (doi:10.1128/JVI.78.22.12446-12454.2004)

[300] D. Chattopadhyay, D. B. Evans, M. R. Deibel, A. F. Vosters, F. M. Eckenrode, H. M. Einspahr, J. O. Hui, A. G. Tomasselli, H. A. Zurcher-Neely and R. L. Heinrikson. 1992. Purification and characterization of heterodimeric human immunodeficiency virus type 1 (HIV-1) reverse transcriptase produced by in vitro processing of p66 with recombinant HIV-1 protease. *J Biol Chem*, **267**, 14227–14232.

[301] A. Jacobo-Molina and E. Arnold. 1991. HIV reverse transcriptase structure-function relationships. *Biochemistry*, **30**, 6351–6356.

[302] B. A. Larder, D. J. Purifoy, K. L. Powell and G. Darby. 1987. Site-specific mutagenesis of AIDS virus reverse transcriptase. *Nature*, **327**, 716–717. URL http://dx.doi.org/10.1038/327716a0. (doi:10.1038/327716a0)

[303] S. F. Le Grice, T. Naas, B. Wohlgensinger and O. Schatz. 1991. Subunit-selective mutagenesis indicates minimal polymerase activity in heterodimer-associated p51 HIV-1 reverse transcriptase. *EMBO J*, **10**, 3905–3911.

[304] J. J. DeStefano, W. Wu, J. Seehra, J. McCoy, D. Laston, E. Albone, P. J. Fay and R. A. Bambara. 1994. Characterization of an RNase H deficient mutant of human immunodeficiency virus-1 reverse transcriptase having an aspartate to asparagine change at position 498. *Biochim Biophys Acta*, **1219**, 380–388.

[305] V. Mizrahi, R. L. Brooksbank and N. C. Nkabinde. 1994. Mutagenesis of the conserved aspartic acid 443, glutamic acid 478, asparagine 494, and aspartic acid 498 residues in the ribonuclease H domain of p66/p51 human immunodeficiency virus type i reverse transcriptase. Expression and biochemical analysis. *J Biol Chem*, **269**, 19245–19249.

[306] H. Huang, R. Chopra, G. L. Verdine and S. C. Harrison. 1998. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science*, **282**, 1669–1675.

[307] V. N. Pandey, N. Kaushik, N. Rege, S. G. Sarafianos, P. N. Yadav and M. J. Modak. 1996. Role of methionine 184 of human immunodeficiency virus type-1 reverse transcriptase in the polymerase function and fidelity of DNA synthesis. *Biochemistry*, **35**, 2168–2179. URL http://dx.doi.org/10.1021/bi9516642. (doi:10.1021/bi9516642)

[308] W. C. Drosopoulos and V. R. Prasad. 1998. Increased misincorporation fidelity observed for nucleoside analog resistance mutations M184V and E89G in human immunodeficiency virus type 1 reverse transcriptase does not correlate with the overall error rate measured in vitro. *J Virol*, **72**, 4224–4230.

[309] L. F. Rezende, K. Curr, T. Ueno, H. Mitsuya and V. R. Prasad. 1998. The impact of multidideoxynucleoside resistance-conferring mutations in human immunodeficiency virus type 1 reverse transcriptase on polymerase fidelity and error specificity. *J Virol*, **72**, 2890–2895.

[310] H. Jonckheere, E. De Clercq and J. Anné. 2000. Fidelity analysis of HIV-1 reverse transcriptase mutants with an altered amino-acid sequence at residues Leu74, Glu89, Tyr115, Tyr183 and Met184. *Eur J Biochem*, **267**, 2658–2665.

[311] A. Ivetac and J. A. McCammon. 2009. Elucidating the inhibition mechanism of HIV-1 non-nucleoside reverse transcriptase inhibitors through multicopy molecular dynamics simulations. *J Mol Biol*, **388**, 644–658. URL http://dx.doi.org/10.1016/j.jmb.2009.03.037. (doi:10.1016/j.jmb.2009.03.037)

[312] M. Madrid, A. Jacobo-Molina, J. Ding and E. Arnold. 1999. Major subdomain rearrangement in HIV-1 reverse transcriptase simulated by molecular dynamics. *Proteins*, **35**, 332–337.

[313] R. Esnouf, J. Ren, C. Ross, Y. Jones, D. Stammers and D. Stuart. 1995. Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. *Nat Struct Biol*, **2**, 303–308.

[314] J. Ren, C. Nichols, L. Bird, P. Chamberlain, K. Weaver, S. Short, D. I. Stuart and D. K. Stammers. 2001. Structural mechanisms of drug resistance for mutations at codons 181 and 188 in HIV-1 reverse transcriptase and the improved resilience of second generation non-nucleoside inhibitors. *J Mol Biol*, **312**, 795–805. URL http://dx.doi.org/10.1006/jmbi.2001.4988. (doi:10.1006/jmbi.2001.4988)

[315] W. Humphrey, A. Dalke and K. Schulten. 1996. VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, **14**, 33–38.

[316] P. A. Keller, S. P. Leach, T. T. Luu, S. J. Titmuss and R. Griffith. 2000. Development of computational and graphical tools for analysis of movement and flexibility in large molecules. *J Mol Graph Model*, **18**, 235–41, 299.

[317] O. Kensch, T. Restle, B. M. Wöhrl, R. S. Goody and H. J. Steinhoff. 2000. Temperature-dependent equilibrium between the open and closed conformation of the p66 subunit of HIV-1 reverse transcriptase revealed by site-directed spin labelling. *J Mol Biol*, **301**, 1029–1039. URL http://dx.doi.org/10.1006/jmbi.2000.3998. (doi:10.1006/jmbi.2000.3998)

[318] V. Vivet-Boudou, J. Didierjean, C. Isel and R. Marquet. 2006. Nucleoside and nucleotide inhibitors of HIV-1 replication. *Cell Mol Life Sci*, **63**, 163–186. URL http://dx.doi.org/10.1007/s00018-005-5367-x. (doi:10.1007/s00018-005-5367-x)

[319] J. Balzarini, P. Herdewijn and E. De Clercq. 1989. Differential patterns of intracellular metabolism of 2',3'-didehydro-2',3'-dideoxythymidine and 3'-azido-2',3'-dideoxythymidine, two potent anti-human immunodeficiency virus compounds. *J Biol Chem*, **264**, 6127–6133.

[320] P. A. Furman, J. A. Fyfe, M. H. St Clair, K. Weinhold, J. L. Rideout, G. A. Freeman, S. N. Lehrman, D. P. Bolognesi, S. Broder and H. Mitsuya. 1986. Phosphorylation of 3'-azido-3'-deoxythymidine and selective interaction of the 5'-triphosphate with human immunodeficiency virus reverse transcriptase. *Proc Natl Acad Sci U S A*, **83**, 8333–8337.

[321] E. J. Eisenberg, G. X. He and W. A. Lee. 2001. Metabolism of GS-7340, a novel phenyl monophosphoramidate intracellular prodrug of PMPA, in blood. *Nucleosides Nucleotides Nucleic Acids*, **20**, 1091–1098.

[322] Avert: http://www.avert.org/.

[323] Y. Hsiou, J. Ding, K. Das, A. D. Clark, S. H. Hughes and E. Arnold. 1996. Structure of unliganded HIV-1 reverse transcriptase at 2.7 å resolution: implications of conformational changes for polymerization and inhibition mechanisms. *Structure*, **4**, 853–860.

[324] L. A. Kohlstaedt, J. Wang, J. M. Friedman, P. A. Rice and T. A. Steitz. 1992. Crystal structure at 3.5 å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science*, **256**, 1783–1790.

[325] D. W. Rodgers, S. J. Gamblin, B. A. Harris, S. Ray, J. S. Culp, B. Hellmig, D. J. Woolf, C. Debouck and S. C. Harrison. 1995. The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1. *Proc Natl Acad Sci U S A*, **92**, 1222–1226.

[326] L. Lawtrakul, A. Beyer, S. Hannongbua and P. Wolschann. 2004. Quantitative Structural Rearrangement of HIV-1 Reverse Transcriptase on Binding to Non-Nucleoside Inhibitors. *Monatshefte fr Chemie*, **135**, 1033–1046.

[327] K. Das, S. G. Sarafianos, A. D. Clark, P. L. Boyer, S. H. Hughes and E. Arnold. 2007. Crystal structures of clinically relevant Lys103Asn/Tyr181Cys double mutant HIV-1 reverse transcriptase in complexes with ATP and non-nucleoside inhibitor HBY 097. *J Mol Biol*, **365**, 77–89. URL http://dx.doi.org/10.1016/j.jmb.2006.08.097. (doi:10.1016/j.jmb.2006.08.097)

[328] J. D. Pata, W. G. Stirtan, S. W. Goldstein and T. A. Steitz. 2004. Structure of HIV-1 reverse transcriptase bound to an inhibitor active against mutant reverse transcriptases resistant to other nonnucleoside inhibitors. *Proc Natl Acad Sci U S A*, **101**, 10548–10553. URL http://dx.doi.org/10.1073/pnas.0404151101. (doi:10.1073/pnas.0404151101)

[329] C. Tantillo, J. Ding, A. Jacobo-Molina, R. G. Nanni, P. L. Boyer, S. H. Hughes, R. Pauwels, K. Andries, P. A. Janssen and E. Arnold. 1994. Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. implications for mechanisms of drug inhibition and resistance. *J Mol Biol*, **243**, 369–387. URL http://dx.doi.org/10.1006/jmbi.1994.1665. (doi:10.1006/jmbi.1994.1665)

[330] T. Restle, B. Müller and R. S. Goody. 1990. Dimerization of human immunodeficiency virus type 1 reverse transcriptase. A target for chemotherapeutic intervention. *J Biol Chem*, **265**, 8986–8988.

[331] R. A. Spence, W. M. Kati, K. S. Anderson and K. A. Johnson. 1995. Mechanism of inhibition of HIV-1 reverse transcriptase by nonnucleoside inhibitors. *Science*, **267**, 988–993.

[332] E. A. Abbondanzieri, G. Bokinsky, J. W. Rausch, J. X. Zhang, S. F. J. Le Grice and X. Zhuang. 2008. Dynamic binding orientations direct activity of HIV reverse transcriptase. *Nature*, **453**, 184–189. URL http://dx.doi.org/10.1038/nature06941. (doi:10.1038/nature06941)

[333] L. Shen, J. Shen, X. Luo, F. Cheng, Y. Xu, K. Chen, E. Arnold, J. Ding and H. Jiang. 2003. Steered molecular dynamics simulation on the binding of NNRTI to HIV-1 RT. *Biophys J*, **84**, 3547–3563.

[334] K. Das, A. D. Clark, P. J. Lewi, J. Heeres, M. R. De Jonge, L. M. H. Koymans, H. M. Vinkers, F. Daeyaert, D. W. Ludovici, M. J. Kukla, B. De Corte, R. W. Kavash, C. Y. Ho, H. Ye, M. A. Lichtenstein, K. Andries, R. Pauwels, M. De Bthune, P. L. Boyer, P. Clark, S. H. Hughes, P. A. J Janssen and E. Arnold. 2004. Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 variants. *J Med Chem*, **47**, 2550–2560. URL http://dx.doi.org/10.1021/jm030558s. (doi:10.1021/jm030558s)

[335] J. S. James. 2005. TMC125: new results, large phase III trial begins. *AIDS Treat News*, pages 2–3.

[336] R. M. Esnouf, J. Ren, A. L. Hopkins, C. K. Ross, E. Y. Jones, D. K. Stammers and D. I. Stuart. 1997. Unique features in the structure of the complex between HIV-1 reverse transcriptase and the bis(heteroaryl)piperazine (BHAP) U-90152 explain resistance mutations for this nonnucleoside inhibitor. *Proc Natl Acad Sci U S A*, **94**, 3984–3989.

[337] W. C. Drosopoulos and V. R. Prasad. 1996. Increased polymerase fidelity of E89G, a nucleoside analog-resistant variant of human immunodeficiency virus type 1 reverse transcriptase. *J Virol*, **70**, 4834–4838.

[338] J. L. Martin, J. E. Wilson, R. L. Haynes and P. A. Furman. 1993. Mechanism of resistance of human immunodeficiency virus type 1 to 2',3'-dideoxyinosine. *Proc Natl Acad Sci U S A*, **90**, 6135–6139.

[339] M. A. Wainberg, W. C. Drosopoulos, H. Salomon, M. Hsu, G. Borkow, M. Parniak, Z. Gu, Q. Song, J. Manne, S. Islam, G. Castriota and V. R. Prasad. 1996. Enhanced fidelity of 3TC-selected mutant HIV-1 reverse transcriptase. *Science*, **271**, 1282–1285.

[340] P. L. Boyer, S. G. Sarafianos, E. Arnold and S. H. Hughes. 2001. Selective excision of AZTMP by drug-resistant human immunodeficiency virus reverse transcriptase. *J Virol*, **75**, 4832–4842. URL http://dx.doi.org/10.1128/JVI.75.10.4832-4842.2001. (doi:10.1128/JVI.75.10.4832-4842.2001)

[341] S. G Sarafianos, A. D. Clark, S. Tuske, C. J. Squire, K. Das, D. Sheng, P. Ilankumaran, A. R. Ramesha, H. Kroth, J. M. Sayer, D. M. Jerina, P. L. Boyer, S. H. Hughes and E. Arnold. 2003. Trapping HIV-1 reverse transcriptase before and after translocation on DNA. *J Biol Chem*, **278**, 16280–16288. URL http://dx.doi.org/10.1074/jbc.M212911200. (doi:10.1074/jbc.M212911200)

[342] F. Rodríguez-Barrios and F. Gago. 2004. Understanding the basis of resistance in the irksome Lys103Asn HIV-1 reverse transcriptase mutant through targeted molecular dynamics simulations. *J Am Chem Soc*, **126**, 15386–15387. URL http://dx.doi.org/10.1021/ja045409t. (doi:10.1021/ja045409t)

[343] Y. Hsiou, J. Ding, K. Das, A. D. Clark, P. L. Boyer, P. Lewi, P. A. Janssen, J. P. Kleim, M. Rösner, S. H. Hughes and E. Arnold. 2001. The Lys103Asn mutation of HIV-1 RT: a novel mechanism of drug resistance. *J Mol Biol*, **309**, 437–445. URL http://dx.doi.org/10.1006/jmbi.2001.4648. (doi:10.1006/jmbi.2001.4648)

[344] F. Rodríguez-Barrios, J. Balzarini and F. Gago. 2005. The molecular basis of resilience to the effect of the Lys103Asn mutation in non-nucleoside HIV-1 reverse transcriptase inhibitors studied by targeted molecular dynamics simulations. *J Am Chem Soc*, **127**, 7570–7578. URL http://dx.doi.org/10.1021/ja042289g. (doi:10.1021/ja042289g)

[345] V. A. Johnson, F. Brun-Vezinet, B. Clotet, D. R. Kuritzkes, D. Pillay, J. M. Schapiro and D. D. Richman. 2006. Update of the drug resistance mutations in HIV-1: Fall 2006. *Top HIV Med*, **14**, 125–130.

[346] P. A. Cane, H. Green, E. Fearnhill, D. Dunn and on behalf of the UK collaborative group on HIV Drug Resistance. 2007. Identification of accessory mutations associated with high-level resistance in HIV-1 reverse transcriptase. *AIDS*, **21**, 447–455.

[347] M. M. Schuckmann, B. Marchand, A. Hachiya, E. N. Kodama, K. A. Kirby, K. Singh and S. G. Sarafianos. 2010. The N348I mutation at the connection

subdomain of HIV-1 reverse transcriptase decreases binding to nevirapine. *J Biol Chem*, **285**, 38700–38709. URL http://dx.doi.org/10.1074/jbc.M110.153783. (doi:10.1074/jbc.M110.153783)

[348] G. N Nikolenko, K. A. Delviks-Frankenberry and V. K. Pathak. 2010. A novel molecular mechanism of dual resistance to nucleoside and nonnucleoside reverse transcriptase inhibitors. *J Virol*, **84**, 5238–5249. URL http://dx.doi.org/10.1128/JVI.01545-09. (doi:10.1128/JVI.01545-09)

[349] S. A. Clark, N. S. Shulman, R. J. Bosch and J. W. Mellors. 2006. Reverse transcriptase mutations 118I, 208Y, and 215Y cause HIV-1 hypersusceptibility to non-nucleoside reverse transcriptase inhibitors. *AIDS*, **20**, 981–984. URL http://dx.doi.org/10.1097/01.aids.0000222069.14878.44. (doi:10.1097/01.aids.0000222069.14878.44)

[350] J. M. Whitcomb, W. Huang, K. Limoli, E. Paxinos, T. Wrin, G. Skowron, S. G. Deeks, M. Bates, N. S. Hellmann and C. J. Petropoulos. 2002. Hypersusceptibility to non-nucleoside reverse transcriptase inhibitors in HIV-1: clinical, phenotypic and genotypic correlates. *AIDS*, **16**, F41–F47.

[351] R. H. Haubrich, C. A. Kemper, N. S. Hellmann, P. H. Keiser, M. D. Witt, D. N. Forthal, J. Leedom, M. Leibowitz, J. M. Whitcomb, D. Richman, J. A. McCutchan and California Collaborative Treatment Group. 2002. The clinical relevance of non-nucleoside reverse transcriptase inhibitor hypersusceptibility. A prospective cohort analysis. *AIDS*, **16**, F33–F40.

[352] J. Garner, J. Deadman, D. Rhodes, R. Griffith and P. A. Keller. 2006. A new methodology for the simulation of flexible protein-ligand interactions. *J Mol Graph Model*, **26**, 187–197. URL http://dx.doi.org/10.1016/j.jmgm.2006.11.004. (doi:10.1016/j.jmgm.2006.11.004)

[353] S. K. Sadiq, D. W. Wright, O. A. Kenway and P. V. Coveney. 2010. Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases. *J Chem Inf Model*, **50**, 890–905. URL http://dx.doi.org/10.1021/ci100007w. (doi:10.1021/ci100007w)

[354] R. D'Aquila, International AIDS Society-USA, J., F. Brun-Vézinet, B. Clotet, B. Conway, L. Demeter, R. Grant, V. Johnson, D. Kuritzkes, C. Loveday, R. Shafer and D. Richman. 2002. Drug Resistance Mutations in HIV-1. *Top HIV Med*, **10**, 21–25.

[355] S. K. Sadiq, D. Wright, S. J. Watson, S. J. Zasada, I. Stoica and P.V. Coveney. 2008. Automated Molecular Simulation Based Binding Affinity Calculator for

Ligand-Bound HIV-1 Proteases. *Journal of Chemical Information and Modeling*, **48**, 1909–1919. ISSN 1549-9596. URL http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ci8000937. (doi:10.1021/ci8000937)

[356] V. Stoll, W. Qin, K. D. Stewart, C. Jakob, C. Park, K. Walter, R. L. Simmer, R. Helfrich, D. Bussiere, J. Kao, D. Kempf, H. L. Sham and D. W. Norbeck. 2002. X-ray crystallographic structure of ABT-378 (lopinavir) bound to HIV-1 protease. *Bioorg Med Chem*, **10**, 2803–2806.

[357] A. W. Schüttelkopf and D. M. F. van Aalten. 2004. *PRODRG*: a tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallographica Section D*, **60**, 1355–1363. URL http://dx.doi.org/10.1107/S0907444904011679. (doi:10.1107/S0907444904011679)

[358] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowsk, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, Al M. A. Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, Head M. Gordon, E. S. Replogle and J. A. Pople. Gaussian 98. Gaussian, Inc., 1998.

[359] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, **79**, 926–935. URL http://link.aip.org/link/?JCP/79/926/1. (doi:10.1063/1.445869)

[360] Y. Duan, C. Wu., S. Chowdhury., M. C. Lee, G. Xiong, W. Zhang., R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang and P. Kollman. 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem*, **24**, 1999–2012. ISSN 0192-8651. URL http://dx.doi.org/10.1002/jcc.10349. (doi:10.1002/jcc.10349)

[361] Kristin L Meagher and Heather A Carlson. 2005. Solvation influences flap collapse in HIV-1 protease. *Proteins*, **58**, 119–125. URL http://dx.doi.org/10.1002/prot.20274. (doi:10.1002/prot.20274)

[362] K. Wittayanarakul, S. Hannongbua and M. Feig. 2008. Accurate prediction of protonation state as a prerequisite for reliable MM-PB(GB)SA binding free energy calculations of HIV-1 protease inhibitors. *J Comput Chem*, **29**, 673–685. URL http://dx.doi.org/10.1002/jcc.20821. (doi:10.1002/jcc.20821)

[363] O. Aruksakunwong, K. Wittayanarakul, P. Sompornpisut, V. Sanghiran, V. Parasuk and S. Hannongbua. 2006. Structural and dynamical properties of different protonated states of mutant HIV-1 protease complexed with the saquinavir inhibitor studied by molecular dynamics simulations. *J Mol Graph Model*, **25**, 324–332. URL http://dx.doi.org/10.1016/j.jmgm.2006.01.004. (doi:10.1016/j.jmgm.2006.01.004)

[364] Y. Won. 2000. Binding Free Energy Simulations of HIV-1 Protease and Hydroxyethylene isostere Inhibitors. *Bull Korean Chem Soc*, **21**, 1207–1212.

[365] H. Ode, S. Neya, M. Hata, W. Sugiura and T. Hoshino. 2006. Computational simulations of HIV-1 proteases–multi-drug resistance due to nonactive site mutation L90M. *J Am Chem Soc*, **128**, 7887–7895. URL http://dx.doi.org/10.1021/ja060682b. (doi:10.1021/ja060682b)

[366] F. Pietrucci, F. Marinelli, P. Carloni and A. Laio. 2009. Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *J Am Chem Soc*, **131**, 11811–11818. URL http://dx.doi.org/10.1021/ja903045y. (doi:10.1021/ja903045y)

[367] S. W. Rick, J. W. Erickson and S. K. Burt. 1998. Reaction path and free energy calculations of the transition between alternate conformations of HIV-1 protease. *Proteins*, **32**, 7–16.

[368] D. Kovalskyy, V. Dubyna, A. E. Mark and A. Kornelyuk. 2005. A molecular dynamics study of the structural stability of HIV-1 protease under physiological conditions: the role of Na^+ ions in stabilizing the active site. *Proteins*, **58**, 450–458. URL http://dx.doi.org/10.1002/prot.20304. (doi:10.1002/prot.20304)

[369] D. Hamelberg and J. A. McCammon. 2004. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. *J Am Chem Soc*, **126**, 7683–7689. URL http://dx.doi.org/10.1021/ja0377908. (doi:10.1021/ja0377908)

[370] Y. Lu, C. Yang and S. Wang. 2006. Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes. *J Am Chem Soc*, **128**, 11830–11839. URL http://dx.doi.org/10.1021/ja058042g. (doi:10.1021/ja058042g)

[371] M. Fornabaio, F. Spyrakis, A. Mozzarelli, P. Cozzini, D. J. Abraham and G. E. Kellogg. 2004. Simple, intuitive calculations of free energy of binding for protein-ligand complexes. 3. The free energy contribution of structural water molecules in HIV-1 protease complexes. *J Med Chem*, **47**, 4507–4516. URL http://dx.doi.org/10.1021/jm030596b. (doi:10.1021/jm030596b)

[372] T. Hou and R. Yu. 2007. Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: mechanism for binding and drug resistance. *J Med Chem*, **50**, 1177–1188. URL http://dx.doi.org/10.1021/jm0609162. (doi:10.1021/jm0609162)

[373] P. Lorenzi, M. Opravil, B. Hirschel, J. P. Chave, H. J. Furrer, H. Sax, T. V. Perneger, L. Perrin, L. Kaiser and S. Yerly. 1999. Impact of drug resistance mutations on virologic response to salvage therapy. Swiss HIV Cohort Study. *AIDS*, **13**, F17–F21.

[374] A. Cingolani, A. Antinori, M. G. Rizzo, R. Murri, A. Ammassari, F. Baldini, S. Di Giambenedetto, R. Cauda and A. De Luca. 2002. Usefulness of monitoring HIV drug resistance and adherence in individuals failing highly active antiretroviral therapy: a randomized study (ARGENTA). *AIDS*, **16**, 369–379.

[375] J. Delgado, K. V. Heath, B. Yip, S. Marion, V. Alfonso, J. S. G. Montaner, M. V. O'Shaughnessy and R. S. Hogg. 2003. Highly active antiretroviral therapy: physician experience and enhanced adherence to prescription refill. *Antivir Ther*, **8**, 471–478.

[376] C. Tural, L.Ruiz, C. Holtzer, J. Schapiro, P. Viciana, J. González, P. Domingo, C. Boucher, C. Rey-Joly, B. Clotet and Havana Study Group. 2002. Clinical utility of HIV-1 genotyping and expert advice: the Havana trial. *AIDS*, **16**, 209–218.

[377] D. Frentz, C. A. B. Boucher, M. Assel, A. De Luca, M. Fabbiani, F. Incardona, P. Libin, N. Manca, V. Müller, B. O Nualláin, R. Paredes, M. Prosperi, E. Quiros-Roldan, L. Ruiz, P. M. A. Sloot, C. Torti, A. Vandamme, K. Van Laethem, M. Zazzi and D. A. M. C. van de Vijver. 2010. Comparison of HIV-1 genotypic resistance test interpretation systems in predicting virological outcomes over time. *PLoS One*, **5**, e11505. URL http://dx.doi.org/10.1371/journal.pone.0011505. (doi:10.1371/journal.pone.0011505)

[378] C. Reid, R. Bassett, S. Day, B. Larder, V. DeGruttola and D Winslow. A dynamic rules-based interpretation system derived by an expert panel is predictive of virological failure. In *Antivir Ther.*, volume 7 of *S91*, 2002.

[379] J. Ravela, B. J. Betts, F. Brun-Vézinet, A. Vandamme, D. Descamps, K. van Laethem, K. Smith, J. M. Schapiro, D. L. Winslow, C. Reid and R. W. Shafer. 2003. HIV-1 protease and reverse transcriptase mutation patterns responsible for discordances between genotypic drug resistance interpretation algorithms. *J Acquir Immune Defic Syndr*, **33**, 8–14.

[380] M. Parera, N. Perez-Alvarez, B. Clotet and M. A. Martínez. 2009. Epistasis among deleterious mutations in the HIV-1 protease. *J Mol Biol*, **392**, 243–250. URL http://dx.doi.org/10.1016/j.jmb.2009.07.015. (doi:10.1016/j.jmb.2009.07.015)

[381] T. Lengauer and T. Sing. 2006. Bioinformatics-assisted anti-HIV therapy. *Nat Rev Microbiol*, **4**, 790–797. URL http://dx.doi.org/10.1038/nrmicro1477. (doi:10.1038/nrmicro1477)

[382] M. Bubak, T. Gubala, M. Malawski, B. Balis, W. Funika, T. Bartynski, E. Ciepiela, D. Harezlak, M. Kasztelnik, J. Kocot, D. Krol, P. Nowakowski, M. Pelczar, J. Wach, M. Assel and A. Tirado-Ramos. 2008. Virtual laboratory for development and execution of biomedical collaborative applications. *Computer-Based Medical Systems, IEEE Symposium on*, pages 373–378. ISSN 1063-7125. (doi:10.1109/CBMS.2008.47)

[383] M. Assel, D. van de Vijver, P. Libin, K. Theys, D. Harezlak, B. O Nualláin, P. Nowakowski, M. Bubak, A. Vandamme, S. Imbrechts, R. Sangeda, T. Jiang, D. Frentz and P. Sloot. 2009. A collaborative environment allowing clinical investigations on integrated biomedical databases. *Studies in health technology and informatics*, **147**, 51–61.

[384] P. M. A. Sloot, P. V. Coveney, G. Ertaylan, V. Müller, C. A. Boucher and M. Bubak. 2009. HIV decision support: from molecule to man. *Phil. Trans. R. Soc. A*, **367**, 2691–2703.

[385] S.J. Zasada and P.V. Coveney. 2009. Virtualizing access to scientific applications with the Application Hosting Environment. *Computer Physics Communications*, **180**, 2513–2525.

[386] M. Bubak, T. Gubala, M. Malawski, B. Balis, W. Funika, T. Bartynski, E. Ciepiela, D. Harezlak, M. Kasztelnik, J. Kocot, D. Krol, P. Nowakowski, M. Pelczar, J. Wach, M. Assel and A. Tirado-Ramos. Virtual Laboratory for Development and Execution of Biomedical Collaborative Applications. In *21st IEEE Int. Symp. on Computer-Based Medical Systems (CBMS 2008)*, pages 373–378. IEEE Computer Society: Washington, DC, 2008.

[387] C. Garriga, M. J. Pérez-Elías, R. Delgado, L. Ruiz, R. Nájera, T. Pumarola, M. del Mar Alonso-Socas, S. García-Bujalance and L. Menéndez-Arias. Nov 2007. Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after virological failure to nelfinavir- and lopinavir/ritonavir-based treatments. *J Med Virol*, **79**, 1617–1628.

[388] L. M. F. Gonzalez, A. F. Santos, A. B. Abecasis, K. Van Laethem, E. A. Soares, K. Deforche, A. Tanuri, R. Camacho, A. Vandamme and M. A. Soares. 2008. Impact of HIV-1 protease mutations A71V/T and T74S on M89I/V-mediated protease inhibitor resistance in subtype G isolates. *J Antimicrob Chemother*, **61**, 1201–1204. URL `http://dx.doi.org/10.1093/jac/dkn099`. (doi:10.1093/jac/dkn099)

[389] J. Martinez-Picado, T. Wrin, S. D. W. Frost, B. Clotet, L. Ruiz, A. J. Brown, C. J. Petropoulos and N. T. Parkin. 2005. Phenotypic hypersusceptibility to multiple protease inhibitors and low replicative capacity in patients who are chronically infected with human immunodeficiency virus type 1. *J Virol*, **79**, 5907–5913. URL `http://dx.doi.org/10.1128/JVI.79.10.5907-5913.2005`. (doi:10.1128/JVI.79.10.5907-5913.2005)

[390] V. Tozzini, J. Trylska, C. A. Chang and J. A. McCammon. 2007. Flap opening dynamics in HIV-1 protease explored with a coarse-grained model. *J Struct Biol*, **157**, 606–615. URL `http://dx.doi.org/10.1016/j.jsb.2006.08.005`. (doi:10.1016/j.jsb.2006.08.005)

[391] J. Trylska, V. Tozzini, C. A. Chang and J. A. McCammon. 2007. HIV-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics. *Biophys J*, **92**, 4179–4187. URL `http://dx.doi.org/10.1529/biophysj.106.100560`. (doi:10.1529/biophysj.106.100560)

[392] N. Kurt, W. R. P. Scott, C. A. Schiffer and T. Haliloglu. 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*, **51**, 409–422. URL `http://dx.doi.org/10.1002/prot.10350`. (doi:10.1002/prot.10350)

[393] T. Skalova, J. Dohnalek, J. Duskova, H. Petrokova, M. Hradilek, M. Soucek, J. Konvalinka and J. Hasek. 2006. HIV-1 protease mutations and inhibitor modifications monitored on a series of complexes. Structural basis for the effect of the A71V mutation on the active site. *J Med Chem*, **49**, 5777–5784. URL `http://dx.doi.org/10.1021/jm0605583`. (doi:10.1021/jm0605583)

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

[395] A. Kontijevskis, P. Prusis, R. Petrovska, S. Yahorava, F. Mutulis, I. Mutule, J. Komorowski and J. E. S. Wikberg. 2007. A look inside HIV resistance through retroviral protease interaction maps. *PLoS Comput Biol*, **3**, e48. URL http://dx.doi.org/10.1371/journal.pcbi.0030048. (doi:10.1371/journal.pcbi.0030048)

[396] A. T. P. Carvalho, P. A. Fernandes and M. J. Ramos. 2006. Molecular dynamics model of unliganded HIV-1 reverse transcriptase. *Med Chem*, **2**, 491–498.

[397] M. Madrid, J. A. Lukin, J. D. Madura, J. Ding and E. Arnold. 2001. Molecular dynamics of HIV-1 reverse transcriptase indicates increased flexibility upon DNA binding. *Proteins*, **45**, 176–182.

[398] I. Bahar, B. Erman, R. L. Jernigan, A. R. Atilgan and D. G. Covell. 1999. Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J Mol Biol*, **285**, 1023–1037. URL http://dx.doi.org/10.1006/jmbi.1998.2371. (doi:10.1006/jmbi.1998.2371)

[399] Z. Zhou, M. Madrid, J. D. Evanseck and J. D. Madura. 2005. Effect of a bound non-nucleoside RT inhibitor on the dynamics of wild-type and mutant HIV-1 reverse transcriptase. *J Am Chem Soc*, **127**, 17253–17260. URL http://dx.doi.org/10.1021/ja053973d. (doi:10.1021/ja053973d)

[400] J. Wang, X. Kang, I. D. Kuntz and P. A. Kollman. 2005. Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA. *J Med Chem*, **48**, 2432–2444. URL http://dx.doi.org/10.1021/jm049606e. (doi:10.1021/jm049606e)

[401] P. Decha, P. Intharathep, T. Udommaneethanakit, P. Sompornpisut, S. Hannongbua, P. Wolschann and V. Parasuk. 2010. Theoretical studies on the molecular basis of HIV-1RT/NNRTIs interactions. *J Enzyme Inhib Med Chem*, pages 1–8. URL http://dx.doi.org/10.3109/14756360903563393. (doi:10.3109/14756360903563393)

[402] W. Treesuwan and S. Hannongbua. 2009. Bridge water mediates nevirapine binding to wild type and Y181C HIV-1 reverse transcriptase–evidence from molecular dynamics simulations and MM-PBSA calculations. *J Mol Graph Model*, **27**, 921–929. URL http://dx.doi.org/10.1016/j.jmgm.2009.02.007. (doi:10.1016/j.jmgm.2009.02.007)

[403] A. Jacobo-Molina, J. Ding, R. G. Nanni, A. D. Clark, X. Lu, C. Tantillo, R. L. Williams, G. Kamer, A. L. Ferris and P. Clark. 1993. Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 åresolution shows bent DNA. *Proc Natl Acad Sci U S A*, **90**, 6320–6324.

[404] J. Wang, S. J. Smerdon, J. Jäger, L. A. Kohlstaedt, P. A. Rice, J. M. Friedman and T. A. Steitz. 1994. Structural basis of asymmetry in the human immunodeficiency virus type 1 reverse transcriptase heterodimer. *Proc Natl Acad Sci U S A*, **91**, 7242–7246.

[405] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, **103**, 8577–8593.

[406] T. Ichiye and M. Karplus. 1991. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, **11**, 205–217. URL http://dx.doi.org/10.1002/prot.340110305. (doi:10.1002/prot.340110305)

[407] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon and L. S. D. Caves. 2006. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, **22**, 2695–2696. URL http://dx.doi.org/10.1093/bioinformatics/btl461. (doi:10.1093/bioinformatics/btl461)

[408] J. Ren, R. Esnouf, E. Garman, D. Somers, C. Ross, I. Kirby, J. Keeling, G. Darby, Y. Jones and D. Stuart. 1995. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat Struct Biol*, **2**, 293–302.

[409] S. G. Sarafianos, K. Das, C. Tantillo, A. D. Clark, J. Ding, J. M. Whitcomb, P. L. Boyer, S. H. Hughes and E. Arnold. 2001. Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *EMBO J*, **20**, 1449–1461. URL http://dx.doi.org/10.1093/emboj/20.6.1449. (doi:10.1093/emboj/20.6.1449)

[410] Q. Xia, J. Radzio, K. S. Anderson and N. Sluis-Cremer. 2007. Probing nonnucleoside inhibitor-induced active-site distortion in HIV-1 reverse transcriptase by transient kinetic analyses. *Protein Sci*, **16**, 1728–1737. URL http://dx.doi.org/10.1110/ps.072829007. (doi:10.1110/ps.072829007)

[411] L. Tambuyzer, H. Azijn, L. T. Rimsky, J. Vingerhoets, P. Lecocq, G. Kraus, G. Picchio and M. de Béthune. 2009. Compilation and prevalence of mutations

associated with resistance to non-nucleoside reverse transcriptase inhibitors. *Antivir Ther*, **14**, 103–109.

[412] L. T. Bacheler, E. D. Anton, P. Kudish, D. Baker, J. Bunville, K. Krakowski, L. Bolling, M. Aujay, X. V. Wang, D. Ellis, M. F. Becker, A. L. Lasut, H. J. George, D. R. Spalding, G. Hollis and K. Abremski. 2000. Human immunodeficiency virus type 1 mutations selected in patients failing efavirenz combination therapy. *Antimicrob Agents Chemother*, **44**, 2475–2484.

[413] J. Ren, C. E. Nichols, P. P. Chamberlain, K. L. Weaver, S. A. Short and D. K. Stammers. 2004. Crystal structures of HIV-1 reverse transcriptases mutated at codons 100, 106 and 108 and mechanisms of resistance to non-nucleoside inhibitors. *J Mol Biol*, **336**, 569–578. URL http://dx.doi.org/10.1016/j.jmb.2003.12.055. (doi:10.1016/j.jmb.2003.12.055)

[414] L. Bacheler, S. Jeffrey, G. Hanna, R. D'Aquila, L. Wallace, K. Logue, B. Cordova, K. Hertogs, B. Larder, R. Buckery, D. Baker, K. Gallagher, H. Scarnati, R. Tritch and C. Rizzo. 2001. Genotypic correlates of phenotypic resistance to efavirenz in virus isolates from patients failing nonnucleoside reverse transcriptase inhibitor therapy. *J Virol*, **75**, 4999–5008. URL http://dx.doi.org/10.1128/JVI.75.11.4999-5008.2001. (doi:10.1128/JVI.75.11.4999-5008.2001)

[415] V. Soriano and C. de Mendoza. 2002. Genetic mechanisms of resistance to NRTI and NNRTI. *HIV Clin Trials*, **3**, 237–248.

[416] R. Silvestri and G. Maga. 2006. Current state-of-the-art in preclinical and clinical development of novel non-nucleoside HIV-1 reverse transcriptase inhibitors. *Expert Opin. Ther. Patients*, **17**, 939–962.

[417] S. Wan and P. V. Coveney. 2011. Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs. *J R Soc Interface*. URL http://dx.doi.org/10.1098/rsif.2010.0609. (doi:10.1098/rsif.2010.0609)

[418] E. Callaway. 2010. Cancer-gene testing ramps up. *Nature*, **467**, 766–767. (doi:10.1038/467766a)

[419] S. Wan and P. V. Coveney. In Press. Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs. *J. R. Soc. Interface*.

[420] M. Raba, K. Limburg, M. Burghagen, J. R. Katze, M. Simsek, J. E. Heckman, U. L. Rajbhandary and H. J. Gross. 1979. Nucleotide sequence of three isoaccepting lysine tRNAs from rabbit liver and SV40-transformed mouse fibroblasts. *Eur J Biochem*, **97**, 305–318.

[421] D. Baltimore. 1992. Viral RNA-dependent DNA polymerase. 1970. *Biotechnology*, **24**, 3–5.

[422] H. M. Temin and S. Mizutani. 1992. RNA-dependent DNA polymerase in virions of rous sarcoma virus. 1970. *Biotechnology*, **24**, 51–56.

[423] H. Varmus. 1987. Reverse transcription. *Sci Am*, **257**, 56–9, 62–4.

[424] C. Barat, V. Lullien, O. Schatz, G. Keith, M. T. Nugeyre, F. Grüninger-Leitch, F. Barr-Sinoussi, S. F. LeGrice and J. L. Darlix. 1989. HIV-1 reverse transcriptase specifically interacts with the anticodon domain of its cognate primer tRNA. *EMBO J*, **8**, 3279–3285.

[425] C. Isel, J. M. Lanchy, S. F. Le Grice, C. Ehresmann, B. Ehresmann and R. Marquet. 1996. Specific initiation and switch to elongation of human immunodeficiency virus type 1 reverse transcription require the post-transcriptional modifications of primer tRNA3Lys. *EMBO J*, **15**, 917–924.

[426] O. Schatz, J. Mous and S. F. Le Grice. 1990. HIV-1 RT-associated ribonuclease H displays both endonuclease and 3'—-5' exonuclease activity. *EMBO J*, **9**, 1171–1176.

[427] E. S. Furfine and J. E. Reardon. 1991. Human immunodeficiency virus reverse transcriptase ribonuclease H: specificity of tRNA(Lys3)-primer excision. *Biochemistry*, **30**, 7041–7046.

[428] G. J. Klarmann, C. A. Schauber and B. D. Preston. 1993. Template-directed pausing of DNA synthesis by HIV-1 reverse transcriptase during polymerization of HIV-1 sequences in vitro. *J Biol Chem*, **268**, 9793–9802.

[429] E.J. Arts, Z. Li and M.A. Wainberg. 1995. Analysis of primer extension and the first template switch during human immunodeficiency virus reverse transcription. *J Biomed Sci*, **2**, 314–321.

[430] J. A. Peliska and S. J. Benkovic. 1992. Mechanism of DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase. *Science*, **258**, 1112–1118.

[431] Y. Chen, M. Balakrishnan, B. P. Roques, P. J. Fay and R. A. Bambara. 2003. Mechanism of minus strand strong stop transfer in HIV-1 reverse transcription. *J Biol Chem*, **278**, 8006–8017. URL http://dx.doi.org/10.1074/jbc.M210959200. (doi:10.1074/jbc.M210959200)

[432] J. L. van Wamel and B. Berkhout. 1998. The first strand transfer during HIV-1 reverse transcription can occur either intramolecularly or intermolecularly. *Virology*, **244**, 245–251. URL http://dx.doi.org/96. (doi:96)

[433] H. Ben-Artzi, J. Shemesh, E. Zeelon, B. Amit, L. Kleiman, M. Gorecki and A. Panet. 1996. Molecular analysis of the second template switch during reverse transcription of the HIV RNA template. *Biochemistry*, **35**, 10549–10557. URL http://dx.doi.org/10.1021/bi960439x. (doi:10.1021/bi960439x)

[434] G. Yusupova, J. M. Lanchy, M. Yusupov, G. Keith, S. F. Le Grice, C. Ehresmann, B. Ehresmann and R. Marquet. 1996. Primer selection by HIV-1 reverse transcriptase on RNA-tRNA(3Lys) and DNA-tRNA(3Lys) hybrids. *J Mol Biol*, **261**, 315–321. URL http://dx.doi.org/10.1006/jmbi.1996.0463. (doi:10.1006/jmbi.1996.0463)

[435] P. Charneau, G. Mirambeau, P. Roux, S. Paulous, H. Buc and F. Clavel. 1994. HIV-1 reverse transcription. a termination step at the center of the genome. *J Mol Biol*, **241**, 651–662. URL http://dx.doi.org/10.1006/jmbi.1994.1542. (doi:10.1006/jmbi.1994.1542)

[436] V. Zennou, C. Petit, D. Guetard, U. Nerhbass, L. Montagnier and P. Charneau. 2000. HIV-1 genome nuclear import is mediated by a central DNA flap. *Cell*, **101**, 173–185. URL http://dx.doi.org/10.1016/S0092-8674(00)80828-4. (doi:10.1016/S0092-8674(00)80828-4)

[437] P. O. Brown, B. Bowerman, H. E. Varmus and J. M. Bishop. 1989. Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proc Natl Acad Sci U S A*, **86**, 2525–2529.

[438] M. J. Roth, P. L. Schwartzberg and S. P. Goff. 1989. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. *Cell*, **58**, 47–54.

[439] S. G. Sarafianos, K. Das, J. Ding, P. L. Boyer, S. H. Hughes and E. Arnold. 1999. Touching the heart of HIV-1 drug resistance: the fingers close down on the dNTP at the polymerase active site. *Chem Biol*, **6**, R137–R146.

[440] J. Ding, S. H. Hughes and E. Arnold. 1997. Protein-nucleic acid interactions and DNA conformation in a complex of human immunodeficiency virus type 1 reverse

transcriptase with a double-stranded DNA template-primer. *Biopolymers*, **44**, 125–138. URL http://dx.doi.org/3.0.CO;2-X. (doi:3.0.CO;2-X)

[441] M. Ghosh, J. Williams, M. D. Powell, J. G. Levin and S. F. Le Grice. 1997. Mutating a conserved motif of the HIV-1 reverse transcriptase palm subdomain alters primer utilization. *Biochemistry*, **36**, 5758–5768. URL http://dx.doi.org/10.1021/bi963045e. (doi:10.1021/bi963045e)

[442] M. D. Powell, M. Ghosh, P. S. Jacques, K. J. Howard, S. F. Le Grice and J. G. Levin. 1997. Alanine-scanning mutations in the "primer grip" of p66 HIV-1 reverse transcriptase result in selective loss of RNA priming activity. *J Biol Chem*, **272**, 13262–13269.

[443] T. Hermann, T. Meier, M. Gtte and H. Heumann. 1994. The 'helix clamp' in HIV-1 reverse transcriptase: a new nucleic acid binding motif common in nucleic acid polymerases. *Nucleic Acids Res*, **22**, 4625–4633.

[444] P. H. Patel, A. Jacobo-Molina, J. Ding, C. Tantillo, A. D. Clark, R. Raag, R. G. Nanni, S. H. Hughes and E. Arnold. 1995. Insights into DNA polymerization mechanisms from structure and function analysis of HIV-1 reverse transcriptase. *Biochemistry*, **34**, 5351–5363.

[445] S. G. Sarafianos, A. D. Clark, K. Das, S. Tuske, J. J. Birktoft, P. Ilankumaran, A. R. Ramesha, J. M. Sayer, D. M. Jerina, P. L. Boyer, S. H. Hughes and E. Arnold. 2002. Structures of HIV-1 reverse transcriptase with pre- and post-translocation AZTMP-terminated DNA. *EMBO J*, **21**, 6614–6624.

[446] Y. Hsiou, K. Das, J. Ding, A. D. Clark, J. P. Kleim, M. Rösner, I. Winkler, G. Riess, S. H. Hughes and E. Arnold. 1998. Structures of Tyr188Leu mutant and wild-type HIV-1 reverse transcriptase complexed with the non-nucleoside inhibitor HBY 097: inhibitor flexibility is a useful design feature for reducing drug resistance. *J Mol Biol*, **284**, 313–323. URL http://dx.doi.org/10.1006/jmbi.1998.2171. (doi:10.1006/jmbi.1998.2171)

[447] J. Ren, R. M. Esnouf, A. L. Hopkins, D. I. Stuart and D. K. Stammers. 1999. Crystallographic analysis of the binding modes of thiazoloisoindolinone non-nucleoside inhibitors to HIV-1 reverse transcriptase and comparison with modeling studies. *J Med Chem*, **42**, 3845–3851.

[448] A. L. Hopkins, J. Ren, H. Tanaka, M. Baba, M. Okamato, D. I. Stuart and D. K. Stammers. 1999. Design of MKC-442 (emivirine) analogues with improved activity against drug-resistant HIV mutants. *J Med Chem*, **42**, 4500–4505.

[449] J. Ren, J. Diprose, J. Warren, R. M. Esnouf, L. E. Bird, S. Ikemizu, M. Slater, J. Milton, J. Balzarini, D. I. Stuart and D. K. Stammers. 2000. Phenylethylthiazolylthiourea (PETT) non-nucleoside inhibitors of HIV-1 and HIV-2 reverse transcriptases. Structural and biochemical analyses. *J Biol Chem*, **275**, 5633–5639.

[450] J. Ren, C. Nichols, L. E. Bird, T. Fujiwara, H. Sugimoto, D. I. Stuart and D. K. Stammers. 2000. Binding of the second generation non-nucleoside inhibitor S-1153 to HIV-1 reverse transcriptase involves extensive main chain hydrogen bonding. *J Biol Chem*, **275**, 14316–14320.

[451] J. Ren, J. Milton, K. L. Weaver, S. A. Short, D. I. Stuart and D. K. Stammers. 2000. Structural basis for the resilience of efavirenz (DMP-266) to drug resistance mutations in HIV-1 reverse transcriptase. *Structure*, **8**, 1089–1094.

[452] J. Ding, K. Das, C. Tantillo, W. Zhang, A. D. Clark, S. Jessen, X. Lu, Y. Hsiou, A. Jacobo-Molina and K. Andries. 1995. Structure of HIV-1 reverse transcriptase in a complex with the non-nucleoside inhibitor alpha-APA R 95845 at 2.8 å resolution. *Structure*, **3**, 365–379.

[453] J. Ding, K. Das, H. Moereels, L. Koymans, K. Andries, P. A. Janssen, S. H. Hughes and E. Arnold. 1995. Structure of HIV-1 RT/TIBO R 86183 complex reveals similarity in the binding of diverse nonnucleoside inhibitors. *Nat Struct Biol*, **2**, 407–415.

[454] J. Jaeger, T. Restle and T. A. Steitz. 1998. The structure of HIV-1 reverse transcriptase complexed with an RNA pseudoknot inhibitor. *EMBO J*, **17**, 4535–4542. URL http://dx.doi.org/10.1093/emboj/17.15.4535. (doi:10.1093/emboj/17.15.4535)

[455] S. G. Sarafianos, K. Das, A. D. Clark, J. Ding, P. L. Boyer, S. H. Hughes and E. Arnold. 1999. Lamivudine (3TC) resistance in HIV-1 reverse transcriptase involves steric hindrance with beta-branched amino acids. *Proc Natl Acad Sci U S A*, **96**, 10027–10032.

[456] J. H. Chan, J. S. Hong, R. N. Hunter, G. F. Orr, J. R. Cowan, D. B. Sherman, S. M. Sparks, B. E. Reitter, C. W. Andrews, R. J. Hazen, M. St Clair, L. R. Boone, R. G. Ferris, K. L. Creech, G. B. Roberts, S. A. Short, K. Weaver, R. J. Ott, J. Ren, A. Hopkins, D. I. Stuart and D. K. Stammers. 2001. 2-Amino-6-arylsulfonylbenzonitriles as non-nucleoside reverse transcriptase inhibitors of HIV-1. *J Med Chem*, **44**, 1866–1882.

[457] P. P. Chamberlain, J. Ren, C. E. Nichols, L. Douglas, J. Lennerstrand, B. A. Larder, D. I. Stuart and D. K. Stammers. 2002. Crystal structures of Zidovudine-

or Lamivudine-resistant human immunodeficiency virus type 1 reverse transcriptases containing mutations at codons 41, 184, and 215. *J Virol*, **76**, 10015–10019.

[458] E. N. Peletskaya, A. A. Kogon, S. Tuske, E. Arnold and S. H. Hughes. 2004. Nonnucleoside inhibitor binding affects the interactions of the fingers subdomain of human immunodeficiency virus type 1 reverse transcriptase with DNA. *J Virol*, **78**, 3387–3397.

[459] J. Ren, R. Esnouf, A. Hopkins, C. Ross, Y. Jones, D. Stammers and D. Stuart. 1995. The structure of HIV-1 reverse transcriptase complexed with 9-chloro-TIBO: lessons for inhibitor design. *Structure*, **3**, 915–926.

[460] A. L. Hopkins, J. Ren, R. M. Esnouf, B. E. Willcox, E. Y. Jones, C. Ross, T. Miyasaka, R. T. Walker, H. Tanaka, D. K. Stammers and D. I. Stuart. 1996. Complexes of HIV-1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent non-nucleoside inhibitors. *J Med Chem*, **39**, 1589–1600. URL http://dx.doi.org/10.1021/jm960056x. (doi:10.1021/jm960056x)

[461] J. Ren, R. M. Esnouf, A. L. Hopkins, E. Y. Jones, I. Kirby, J. Keeling, C. K. Ross, B. A. Larder, D. I. Stuart and D. K. Stammers. 1998. 3'-Azido-3'-deoxythymidine drug resistance mutations in HIV-1 reverse transcriptase can induce long range conformational changes. *Proc Natl Acad Sci U S A*, **95**, 9518–9523.

[462] J. Ren, R. M. Esnouf, A. L. Hopkins, J. Warren, J. Balzarini, D. I. Stuart and D. K. Stammers. 1998. Crystal structures of HIV-1 reverse transcriptase in complex with carboxanilide derivatives. *Biochemistry*, **37**, 14394–14403. URL http://dx.doi.org/10.1021/bi981309m. (doi:10.1021/bi981309m)

[463] S. Tuske, S. G. Sarafianos, A. D. Clark, J. Ding, L. K. Naeger, K. L. White, M. D. Miller, C. S. Gibbs, P. L. Boyer, P. Clark, G. Wang, B. L. Gaffney, R. A. Jones, D. M. Jerina, S. H. Hughes and E. Arnold. 2004. Structures of HIV-1 RT-DNA complexes before and after incorporation of the anti-AIDS drug tenofovir. *Nat Struct Mol Biol*, **11**, 469–474. URL http://dx.doi.org/10.1038/nsmb760. (doi:10.1038/nsmb760)

[464] A. L. Hopkins, J. Ren, J. Milton, R. J. Hazen, J. H. Chan, D. I. Stuart and D. K. Stammers. 2004. Design of non-nucleoside inhibitors of HIV-1 reverse transcriptase with improved drug resistance properties. 1. *J Med Chem*, **47**, 5912–5922. URL http://dx.doi.org/10.1021/jm040071z. (doi:10.1021/jm040071z)

[465] G. A. Freeman, C. W. Andrews III, A. L. Hopkins, G. S. Lowell, L. T. Schaller, J. R. Cowan, S. S. Gonzales, G. W. Koszalka, R. J. Hazen, L. R. Boone, R. G.

Ferris, K. L. Creech, G. B. Roberts, S. A. Short, K. Weaver, D. J. Reynolds, J. Milton, J. Ren, D. I Stuart, D. K. Stammers and J. H. Chan. 2004. Design of non-nucleoside inhibitors of HIV-1 reverse transcriptase with improved drug resistance properties. 2. *J Med Chem*, **47**, 5923–5936. URL http://dx.doi.org/10.1021/jm040072r. (doi:10.1021/jm040072r)

[466] K. Das, J. Ding, Y. Hsiou, A. D. Clark, H. Moereels, L. Koymans, K. Andries, R. Pauwels, P. A. Janssen, P. L. Boyer, P. Clark, R. H. Smith, M. B. Kroeger Smith, C. J. Michejda, S. H. Hughes and E. Arnold. 1996. Crystal structures of 8-Cl and 9-Cl TIBO complexed with wild-type HIV-1 RT and 8-Cl TIBO complexed with the Tyr181Cys HIV-1 RT drug-resistant mutant. *J Mol Biol*, **264**, 1085–1100.

[467] D. M. Himmel, K. Das, A. D. Clark, S. H. Hughes, A. Benjahad, S. Oumouch, J. Guillemont, S. Coupa, A. Poncelet, I. Csoka, C. Meyer, K. Andries, C. H. Nguyen, D. S. Grierson and E. Arnold. 2005. Crystal structures for HIV-1 reverse transcriptase in complexes with three pyridinone derivatives: a new class of non-nucleoside inhibitors effective against a broad range of drug-resistant strains. *J Med Chem*, **48**, 7582–7591. URL http://dx.doi.org/10.1021/jm0500323. (doi:10.1021/jm0500323)

[468] J. Ren, C. E. Nichols, A. Stamp, P. P. Chamberlain, R. Ferris, K. L. Weaver, S. A. Short and D. K. Stammers. 2006. Structural insights into mechanisms of non-nucleoside drug resistance for HIV-1 reverse transcriptases mutated at codons 101 or 138. *FEBS J*, **273**, 3850–3860. URL http://dx.doi.org/10.1111/j.1742-4658.2006.05392.x. (doi:10.1111/j.1742-4658.2006.05392.x)

[469] D. M. Himmel, S. G. Sarafianos, S. Dharmasena, M. M. Hossain, K. McCoy-Simandle, T. Ilina, A. D. Clark, J. L. Knight, J. G. Julias, P. K. Clark, K. Krogh-Jespersen, R. M. Levy, S. H. Hughes, M. A. Parniak and E. Arnold. 2006. HIV-1 reverse transcriptase structure with RNase H inhibitor dihydroxy benzoyl naphthyl hydrazone bound at a novel site. *ACS Chem Biol*, **1**, 702–712. URL http://dx.doi.org/10.1021/cb600303y. (doi:10.1021/cb600303y)