UNIVERSITY OF SOUTHAMPTON

# The Computational Investigation Of Conformational Change

Adrian Peter Wiley

A thesis submitted for the qualification of
Doctor of Philosophy at the University of Southampton

School of Chemistry

November 2004

University of Southampton

ABSTRACT

FACULTY OF SCIENCE

SCHOOL OF CHEMISTRY

Doctor of Philosophy

# THE COMPUTATIONAL INVESTIGATION OF CONFORMATIONAL CHANGE

by Adrian Peter Wiley

Methods for investigating conformational motions are developed and applied to a range of protein systems. The motions of biological molecules can occur on timescales beyond that accessible to molecular dynamics (MD) simulation, and algorithms that enhance conformational sampling are therefore of great use. Reversible digitally filtered molecular dynamics (RDFMD) can amplify or suppress motions of specific frequencies as a simulation evolves. A method for the systematic parameterisation of RDFMD is presented and used to generate a protocol that maximises backbone dihedral angle change without overheating the target system. The recently developed Hilbert-Huang transform (HHT) is used to determine the most suitable frequencies for amplification by RDFMD. A computationally expensive parallel tempering (PT) simulation has been performed on the YPGDV pentapeptide, to determine the equilibrium distribution of accessible conformers. This has then been used to assist in the parameterisation of the RDFMD method. PT of YPGDV has also been used to study the population of *cis* and *trans* peptide bonds at a range of temperatures, sampling almost three thousand isomerisation events. Investigations into the conformational motions of T4 lysozyme, E. coli dihydrofolate reductase (EcDHFR), and human immunodeficiency virus-1 protease (HIV-1 PR) are also presented. In each case, conclusions from MD, PT, and RDFMD are in agreement with experimental data, and provide further insight into the dynamics of the system.

# Acknowledgements

I would like to thank all those who assisted me throughout my time at Southampton. In particular, I thank Jon Essex for his supervision and support which made the work presented in this thesis possible. Also, Jeremy Frey, my internal examiner, whose comments helped to keep me on track, and George Attard, for a lively transfer viva.

In some way I have been assisted by all the post-docs who contribute to Jon's support network: Steve Phillips originally created and taught me about the RDFMD method, Chris Woods kept my ideas in check with a significant contribution to the parallel tempering work in this thesis, Martin Swain and Rob Gledhill coded much of the software required for my research, and Richard Maurer assisted with the protonation of the protein systems.

Financial support was provided by EPSRC, and many thanks to all those who have created and supported the vital computer clusters that were essential for the quantity of simulation I have been able to perform over the years.

Finally, and most importantly, I thank Esther for her love, support and patience that backed me every step of the way.

# Contents

# Chapter 1

# Introduction

Knowledge of the conformations that a molecule can adopt is key to understanding its biological properties. Conformational changes that can occur in proteins vary from movements of small loop regions to large-scale domain motions in which areas of the protein move in a correlated fashion. These can take place on the millisecond time-scale, and may be difficult to fully characterise using experimental methods.

Protein simulations using molecular dynamics (MD) algorithms are frequently performed to provide atomistic detail that is otherwise unavailable.[1] Using traditional MD the computational expense of simulating the timescales required to see many conformational motions is too great. Even when a 'rare' conformational event is sampled by simulation, the distribution of accessible conformers is not converged, requiring even longer simulations and abundant sampling of all possible states.

Algorithms that improve the efficiency of conformational sampling are therefore of great use, and much research is put into their development.[2] Many make assumptions about the system to be studied, or use prior knowledge based on experimental evidence. For example, if two known conformers are obtained from X-ray crystallography, the simulation can be 'driven' from one to the next.[3] Other methods provide assistance over unknown energy barriers and use statistical meth-

ods to determine the validity of the generated conformers.[4]

Two methods of enhancing conformational sampling are presented in this thesis. Parallel tempering[5] (PT), constructs a generalised ensemble from a number of independent simulations running at different temperatures. The method is computationally expensive, and produces a random walk in temperature space by adjusting the kinetic energy of the system according to its position on the potential energy surface. Energy barriers can therefore be overcome, and statistically rigorous information is maintained for a range of temperatures.

In this project, PT has initially been implemented and applied to simple systems, developing the methodology in preparation for use with proteins of significant size. A technique of determining the distribution of temperatures that maximises the efficiency of the algorithm has been tested, and the conformational distributions generated by PT are shown to converge for systems that display quasi-ergodicity.

The second method, Reversible Digitally Filtered Molecular Dynamics[6] (RDFMD), overcomes conformational barriers by providing frequency-specific kinetic energy. The method has previously been successfully applied to simple systems and is not commonly used. Over the course of this project RDFMD has been rigorously tested and the methodology advanced to the stage where it can be systematically parameterised for a desired purpose. Insight into frequency information generated by MD has been gained using novel digital signal processing techniques including empirical mode decomposition[7] and the Hilbert transform.[8]

An RDFMD protocol designed to maximise dihedral angle motion, and thus induce conformational changes in flexible regions of proteins, has been developed. The distribution of conformers generated for a simple system have been compared with those from MD and PT simulations, assisting with the RDFMD parameterisation process.

Conformational investigations using MD, PT and RDFMD have been performed on three systems of significant size and interest: T4 lysozyme, E. coli dihydrofolate reductase (EcDHFR), and human immunodeficiency virus-1 protease

(HIV-1 PR). In each case, obtained results are in agreement with experimental data and add to the understanding of the system, as described in current literature. For T4 lysozyme, a stable closed conformation is observed, with infrequent opening events that can be rapidly induced using RDFMD. The investigation of E. coli dihydrofolate reductase reveals a transition between known closed and occluded conformations, without a more open intermediate previously thought to be required. Finally, an alternative mechanism of opening the protein flaps of HIV-1 PR is presented, with significant implications for drug design.

# Chapter 2

# Simulation Methodology

## 2.1 Computational Chemistry

Computational chemistry is a rapidly growing field due to the decreasing cost of computer power, and improvements of available techniques. The conformations, motions and properties of molecules can be predicted without the need for complex synthesis or difficult experiments. Methods are based upon models that approximate the chemical system to be studied, and generate required parameters from experimentally available information and current theoretical understanding. The adequacy of the model, and the reliability of parameter optimisation are limiting factors that must always be considered.

A range of methodology exists, designed to suit different chemical investigations. Reactions in which a bond is made or broken can be performed using quantum mechanics[9] (QM), in which the effects of electrons are explicitly included. Protein simulations are typically run using molecular mechanics[10] (MM), in which the Born–Oppenheimer approximation is invoked. Atoms are described by charge point centres with momentum. A range of further approximations can be invoked to reduce computational expense. For example, atoms can be grouped into structurally inflexible units,[11] and solvent molecules may be replaced by implicit solvent calculations.[12]

## 2.2 Force Fields

Central to molecular mechanics is the potential energy description of a system, termed the force field. Many force fields exist although protein simulation is dominated by AMBER,[13–15] CHARMM,[16,17] GROMOS,[18] and OPLS.[19] The most recent AMBER force field was published in 2003 (*ff03*),[15] building on the constantly refined force field of Cornell *et al.*,[13] which in turn replaced the work of Weiner *et al.*[14] The general form of the AMBER potential energy equation is a summation of component terms:

$$E_{potential} = E_{bonds} + E_{angles} + E_{dihedrals} + E_{coulomb} + E_{LJ} \tag{2.1}$$

Bond lengths, $r$, and angles, $\theta$, are described by simple harmonic expressions using force constants, $K_r$ and $K_\theta$, and equilibrium values, $r_{eq}$ and $\theta_{eq}$:

$$E_{bonds} = \sum_{bonds} K_r(r - r_{eq})^2 \tag{2.2}$$

$$E_{angles} = \sum_{angles} K_\theta(\theta - \theta_{eq})^2 \tag{2.3}$$

Dihedral terms are parameterised by a Fourier series with Fourier coefficients, $V_n$, dihedral angle, $\phi$ and phase difference, $\gamma$:

$$E_{dihedrals} = \sum_{dihedrals} \frac{V_n}{2}[1 + cos(n\phi - \gamma)] \tag{2.4}$$

Non-bonded interactions (i.e. interactions between atoms separated by at least three bonds) are described by electrostatic and van der Waals terms. Electrostatics are given by Coulomb's Law (Equation 2.5) in which the energy of interaction between two atoms is calculated from their partial atomic charges, $q$ (typically generated using *ab initio* methods), and the inter-atomic distance, $r$. Since $E_{coulomb}$ decays at the rate $\frac{1}{r}$, it is significant at lengths on the molecular scale and cannot be treated with a short-range cutoff. A number of methods exist that approximate and speed up the expensive electrostatics calculation, including particle mesh Ewald[20] used in this project. Other long-range treatments include the reaction field

method[21] and the cell multipole method.[22]

$$E_{coulomb} = \frac{q_i q_j}{4\pi\varepsilon_0 r} \tag{2.5}$$

van der Waals terms are represented by a 6–12 (or Lennard-Jones (LJ)) potential using the collision-diameter, $\sigma$, and well-depth, $\varepsilon$. 1–4 interactions (those between atoms separated by three bonds), are treated by a scaling factor. $E_{LJ}$ decays at a much faster rate than $E_{coulomb}$, and long-range LJ interactions can be treated with a cutoff, typically in the region of 10 – 14 Å. A simple truncation cutoff introduces a discontinuity onto the potential energy surface and so switching functions or shifted potentials are generally preferred.[23]

$$E_{LJ} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.6}$$

The CHARMM force field is also used in this work, the latest parameter set for proteins being equivalent in the last two releases CHARMM22 and CHARMM27. The functional form used does not differ from that of AMBER, except for the inclusion of 1–3 Urey–Bradley terms, introduced to improve comparison to vibrational spectra. These place a bond (using the same functional form as shown in Equation 2.2) between atoms on either side of a bond angle. A 1–4 scaling factor is not use by CHARMM force fields.

## 2.3 Molecular Dynamics

Molecular Dynamics (MD) is generally performed on systems for which at least each heavy atom (i.e. non-hydrogen) is considered as an individual entity. Forces, $\mathbf{F}$, are calculated from the gradient of the potential energy surface, $\mathbf{U}$, for a specific arrangement of atoms, $\mathbf{r}$ (Equation 2.7). Once the forces are known, the acceleration, $\mathbf{a}$ or $\frac{d^2\mathbf{r}}{dt^2}$, acting upon each atom with mass $m$ can be calculated using Newton's second law of motion, $\mathbf{F} = m\mathbf{a}$.

$$\mathbf{F}(\mathbf{r}) = -(\frac{d}{dx}, \frac{d}{dy}, \frac{d}{dz})U(\mathbf{r}) \tag{2.7}$$

Since the force acting upon each atom depends upon the coordinates of all other atoms, a set of simultaneous, differential equations are generated that can not be integrated analytically. Numerical solutions are calculated using finite difference techniques that approximate positions and dynamic properties as Taylor expansions. Equation 2.8 shows this series for the progression of coordinates over a small advancement of time (or time-step), $\delta t$.

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) + \ldots \tag{2.8}$$

From these expansions several algorithms have been developed, including the velocity Verlet method[24] shown in Equations 2.9 to 2.11. The velocity Verlet algorithm deals explicitly with velocities and is time reversible, a requirement of the RDFMD method.

$$\mathbf{v}(t+\frac{\delta t}{2}) = \mathbf{v}(t) + \frac{\delta t}{2}\mathbf{a}(t) \tag{2.9}$$

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

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) \tag{2.11}$$

The stability of the integrator algorithm is dependent on the length of the time-step. If this is too long, forces do not remain constant during the duration of the step and energy will not be conserved. If the time-step is needlessly small, the method is inefficient. Typically a time-step of 1 fs is used, at which the fastest motions in the system (hydrogen-containing bond vibrations) are adequately sampled. Constraining the fastest degrees of freedom can therefore significantly save computational expense, and the iterative SHAKE[25] algorithm is typically employed to constrain either all bond lengths, or the lengths of all bonds containing hydrogen atoms. A 2 fs time-step is then usually sufficient to conserve

energy, halving computational expense for the overhead of the constraint algorithm.

## 2.4 Modelling Solvent

To model a solvated system either sufficient solvent molecules must be explicitly included for simulation, or some method of implicitly modelling solvent is required.

Common explicit solvent models include TIP3P[26] and SPC/E,[27] both with 3 atomic sites and rigid geometry to reduce computational expense. Models exist that use more sites, included to improve certain physical properties such as tetragonal behaviour and replication of radial distribution functions.[28] Ideally the solvent should be cheap to implement when the solute is the subject of investigation, and the 3 site models dominate explicit solvent simulations of biological systems.

Surrounding a solute with sufficient solvent molecules greatly increases the size and cost of the MD calculation. Implicit models are therefore of use if their implementation is cheaper than simulation of explicit solvent, and if results produced are as reliable. There are several methods used for biological systems, the most popular being the Poisson-Boltzmann model and the generalised Born model.

The Poisson-Boltzmann (PB) model[29–31] is so named due to the use of the linearised PB equation to describe the change in electrostatic potential, $\nabla\phi(\mathbf{r})$, with respect to the change in dielectric constant, $\nabla\varepsilon(\mathbf{r})$ (Equation 2.12). $\rho$ represents the solvent charge distribution and $\kappa(\mathbf{r})$ is the Debye-Huckel parameter- zero within the interior of the solute, and within an ion exclusion radius of the molecular surface. The dielectric, $\varepsilon(\mathbf{r})$, takes a interior or exterior value depending on whether the atom lies inside or outside the solute surface.[32] Solvent is therefore represented as a dielectric continuum; a macroscopic concept being applied on the microscopic scale.

$$\nabla\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}) - \kappa^2(\mathbf{r})\varepsilon(\mathbf{r})\phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{2.12}$$

Owing to its complexity, the PB calculation must be solved numerically, generally through the use of a finite difference grid. Many studies have been performed developing parameters for the model,[33,34] a step that remains critical for the method's success.

The generalised Born (GB) model[35–37] calculates the electrostatic contribution to solvation based upon the GB equation (Equation 2.13); a combination of the pairwise sum of interacting charges in the solute representing the induced reaction field energy, and the Born-equation. $a_{ij}$ represents the Born radii which must be derived for each polar atom in the molecule via a non-trivial procedure. The factor of $\left(1 - \frac{1}{\varepsilon}\right)$ describes the change of dielectric constant when moving from vacuum to solvent, again representing solvent as a dielectric continuum. The GB method is often used in conjunction with methods to calculate the solvent-accessible surface area.

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

Much work has been done validating the use of implicit solvent models,[12,38–40] however the methodology is still in development and no single method has emerged with sufficient confidence to replace conventional explicit solvent models. Although implicit solvent dynamics are often similar to those reported for explicit solvent simulations, deviations from known starting structures are generally larger when using an implicit model,[38] and less kinetic energy is required to overcome conformational transitions.[39]

## 2.5 Periodic boundary conditions

If a cube were created centred on a protein and filled with explicit solvent molecules, then not only would the outer molecules require some from of

constraint to stop them drifting into the surrounding vacuum, but the long-range forces seen by the protein would not resemble those representing bulk solvent. One solution to this problem is to impose periodic boundary conditions where the cube is replicated infinitely in every direction, creating a periodic array. If an atom moves out of the cube during simulation, it appears on the opposite side, and each atom is allowed to interact only with the closest image of each other atom (the minimum image convention). If the surrounding solvent depth is sufficient, the protein will experience solution phase conditions as desired.

Although cuboid systems are still the most common, and generally simplest to use in available MD packages, they many not be the most efficient for all solute shapes. If the goal is assumed to be maintaining a minimum solute-solute distance whilst requiring the lowest number of solvent molecules, then, for example, a truncated octahedron or hexagonal prism may prove the most efficient periodic cell.

## 2.6 Physical properties

Various thermodynamic properties, such as pressure, temperature, energy and heat capacity, can be calculated from simulation averages. For example, the temperature, $T$, and average kinetic energy, $\langle K \rangle$, can be found as shown in Equation 2.14. Angular brackets denote simulation averages. The number of degrees of freedom, $D$, is three times the number of atoms, less the number of constraints such as fixed bond lengths, or removal of external motions.

$$\left\langle \sum_{k=1}^{N} \frac{m_k \mathbf{v}_k^2}{2} \right\rangle = \frac{Dk_BT}{2} = \langle K(\mathbf{p}) \rangle_T \tag{2.14}$$

Molecular dynamics is typically run to replicate real-life conditions, for which a molecular system is subject to physical properties dictated by its macromolecular surrounding. An MD integrator will produce constant energy conditions, termed the microcanonical, or NVE, ensemble, in which the number of atoms (N), volume (V) and energy (E) are all kept constant. Other ensembles commonly used include the canonical ensemble (NVT), in which temperature (T) is constrained to a target

value, and the isothermal-isobaric ensemble (NPT) in which pressure (P) is also controlled. In traditional molecular dynamics, it is usual that the NPT and NVT ensembles are used to equilibrate a system to desired conditions, and then the temperature and pressure constraints are relaxed to minimise effects on the system.

The simplest method of controlling the temperature is to scale all simulation velocities.[41] Manipulation of Equation 2.14 reveals a velocity scaling factor, $\lambda$, that modifies the temperature to a desired value, $T_{desired}$:

$$T_{desired} = \frac{1}{Dk_BT}\sum_{k=1}^{N}\frac{m_k(\lambda \mathbf{v}_k)^2}{2} \tag{2.15}$$

$$= \lambda^2 T \tag{2.16}$$

$$\lambda = \sqrt{\frac{T_{desired}}{T}} \tag{2.17}$$

Although velocity rescaling reaches the desired temperature quickly, temperature differences between parts of the system are maintained and artefacts can be introduced by sudden changes of velocities. A commonly used thermostat involves coupling to an external heat bath,[42] but this method, like velocity rescaling, does not produce a rigorous canonical ensemble. One method that does is the application of a Langevin thermostat,[43] that introduces stochastic collisions to a chosen atom in the system, as shown in Equation 2.18. $\gamma$ indicates a collision frequency and $\mathbf{R}$ is a random force. Forces are manipulated, adjusting the velocities of the next step, and therefore the temperature.

$$m\frac{d^2\mathbf{r}}{dt^2} = \mathbf{F}(\mathbf{r}) - \gamma\frac{d\mathbf{r}}{dt}m + \mathbf{R} \tag{2.18}$$

Pressure control is generally more complex than that of temperature, since pressure fluctuations during simulation must be averaged over a large number of steps to obtain a reliable value. Pressure is manipulated by adjusting the volume of the system, either by scaling each of the the atomic directions together (isotropic scaling) or independently (anisotropic scaling).

## 2.7 Monte Carlo

Monte Carlo (MC) is an alternative form of molecular simulation which randomly selects and attempts to change a degree of freedom, such as the location of an atom or a bond length.[44] Whether the change is accepted relies on the outcome of a test based upon the difference in potential energies, $\Delta E$, between the original and generated coordinate sets, and a random number (Equation 2.19). If the test is satisfied, simulation continues from the new coordinate set. If failed, a new move is attempted from the original configuration. In this manner an ensemble is generated from each coordinate set from which a move is attempted. MC must limit the size of each random move so that a suitable probability of accepting moves is achieved, otherwise simulation progression is limited and the method inefficient. Unlike MD, the MC method includes no concept of time.

$$e^{\left(-\frac{\Delta E}{k_B T}\right)} \geq rand(0,1) \tag{2.19}$$

MC simulations of biological systems encounter significant difficulties due to the nature of motions required to move from one area of the potential energy surface to another. To simulate a dihedral angle motion using MC moves in cartesian space requires each atom on one side of the dihedral angle to move in a correlated fashion; an unlikely conclusion of random events. An alternative method would be to perform MC moves of the dihedral angle itself, however the effects of even a small rotation about a bond would lead to significant motions far from the site of the move. The resulting high energies from atomic overlaps would lead to very few moves being accepted. Very small moves must therefore be used, and conformational sampling is slow.

## 2.8 Summary

In this chapter, a summary of simulation methodology has been presented, including the potential energy description of the system, algorithms required for molecular dynamics, methods for modelling solvent, and periodic boundary

conditions. Most of the included methods are implemented in current simulation packages, and are widely used and accepted.

# Chapter 3

# Conformational change

## 3.1 The nature of conformational change

Proteins are polymer chains made up from twenty possible monomer units (amino acids), each sharing a common backbone structure with differing functional side chains. It is convenient to consider the amino acid chains (the primary structure) in terms of relatively rigid and stable secondary structure units, held together by hydrogen bonds. These are classified as either helices, where a section of coiled amino acids forms a rod–like helix, or sheets, where amino acid strands form layered structures. These units of secondary structure are connected by flexible loop regions, and secondary structure units can themselves associate to form compact globular units, or domains, that move in a correlated fashion.[45]

Understanding the functions and mechanisms of biological molecules requires detailed knowledge of their structure and internal motions. The arrangement, or conformation, of a protein is of considerable importance to its biological function.[46,47] For example, Creutzfeldt–Jakob Disease (CJD) is a fatal neurodegenerative disorder that is believed to be caused by conformational change in the prion protein.[48] The prion ('proteinaceous infectious particle') theory proposes that the prion protein (a normal constituent of mammalian cells) undergoes a conformational change to an active form that induces the same change in further prion proteins. Protein aggregation then occurs, which is associated with cell death.

Large scale conformational motion occurs predominantly by rotations about dihedral angles, since a small change in these can have a significant effect on the structure of the protein. It is believed that anharmonic, low frequency motions in dihedral angles are responsible for conformational changes.[49,50]

## 3.2 Experimental Methods

There are several experimental methods for determining 3–dimensional structures of molecules, of which X-ray crystallography[51] is undoubtedly the most common. When X-rays are shone through a protein crystal (generally cooled to very low temperatures to limit atomic motions), the rays are deflected forming diffraction patterns. These can be interpreted using various computer packages and known information about the protein. Heavier atoms scatter X-rays more effectively and for larger systems, the more complex diffraction patterns mean it becomes increasingly difficult to locate hydrogen atoms, and to differentiate between atoms of similar electron density.

The molecular arrangements seen in X-ray crystallography are 'snapshots' in which the protein is not in its natural environment. Often, proteins can be crystallised in a variety of arrangements that differ significantly[45,52] and it is believed that crystal packing forces can have a significant influence on the observed conformation.[53]

Another prominent experimental technique for studying protein structures is Nuclear Magnetic Resonance (NMR).[54] Certain atomic nuclei (for example $^{1}H$, $^{2}H$, $^{13}C$, $^{14}N$, and $^{15}N$) adopt a specific orientation, or spin, in the presence of a magnetic field. The spin can be changed by the absorption of electromagnetic radiation, the frequency of which is highly dependent on the nuclei's local environment. Information about the protein structure can also be determined from spin coupling between magnetic nuclei.

NMR gathers data about the protein when it is in solution, and crystal packing

artefacts associated with X-ray crystallography are avoided. Unfortunately, application of the method to large proteins can become extremely complex and time consuming owing to the need for significant isotopic replacements to allow for the spectra to be assigned. The resolution of the technique is a limiting factor to the size of system that can have its structure resolved.

Structures obtained using NMR or X-ray crystallography generally require refinement using computational methods such as molecular mechanics minimisation or torsional dynamics MD.[55] A range of techniques have been developed to extend the basic methods that can provide information on dynamical properties of the protein.[56] However, experimental methods are generally unable to follow atomic coordinates with sufficient time resolution to describe the nature of many conformational events.[1]

## 3.3 Computational methods

The conformational changes in protein systems can vary from small movements of key residues to large domain-scale hinges that dramatically alter the protein's structure.[45] The time scales associated with such events can be of the order of milli-seconds and therefore are not adequately sampled by traditional MD.

In this section, current computational methods used to predict and analyse conformational changes in proteins are reviewed. These are categorised as methods that infer motions from known structures, pathway refinement and exploration techniques, methods that increase conformational sampling, and equilibrium methods that aim to generate the correct distribution of conformers.

### 3.3.1 Methods that infer motions from known structures

A range of computational methods designed to extract dynamics from experimentally obtained structures exist. In Normal Mode Analysis[57] (NMA) the potential energy function of a system is differentiated twice with respect to coordinates to give a Hessian matrix. This can be diagonalised to yield a set of eigenvectors

and eigenvalues; the eigenvectors represent directions of oscillators with a force constant given by the corresponding eigenvalue. NMA requires the Hessian to be symmetrical which only occurs at energy minima. This is a major limitation of this quick and simple technique, and NMA is often applied in cases where the system is not in a minimum energy configuration.

Essential Dynamics[58] (ED) separates and orders modes of motion from several coordinate sets, from either simulation snapshots, or from experimentally obtained structures. By using Principal Component Analysis (PCA), the coordinate covariance matrix is diagonalised in a similar fashion to NMA, yielding eigenvectors that represent directions whose statistical relevance is given by the corresponding eigenvalues. By ignoring low-valued eigenvalues (in other words directions which are responsible for little variance between the set of coordinates), the essential modes are identified. In this manner the motions required to move between several coordinate sets that, for example, represent reaction intermediates or conformers crystallised under different conditions, can be determined, without use of a potential energy function.

A similar approach to ED involves the grouping of atoms into quasi-rigid bodies that move in a correlated fashion. For example, the HingeFind[59] program searches for rigid domains using a multiple least-squares fitting procedure. Alternatively, dynamic domains can be located using clusters of rigid-body parameters, as performed by the DynDom[60] and DomainFinder[61] packages.

The validity of inferring motions between known structures depends upon the connection between the coordinate sets used. If these are too different, it is likely that a motion describing the transition between them will be non-physical, possibly involving high energy atomic overlaps. Here the limitations of methods that do not allow structure reorganisation due to energy constraints are reached. However, the ability to gain insight into possible motions using a single coordinate set and potential energy function, or several coordinate sets, make NMA and ED useful and widely applied techniques.

### 3.3.2 Pathway refinement and exploration techniques

Several methods that refine the pathway between experimental structures exist. For example, Path Energy Minimisation[62] (PEM) separates the path between known states as a discrete set of conformations. A function that approximates the peak potential energy of the path is then minimised by adjusting the coordinates of the conformer set.

The use of molecular dynamics can be coupled to analysis of a reaction path. For example, CONTRA MD[63] (Conformational Transitions using MD with minimum biasing) defines a conformational variable that describes motion from the starting to final coordinates. A number of conformers are extrapolated from the path, and these are run in short MD simulations. Trajectories are then tested to see whether they have evolved along the conformational variable, and if so, they are pieced together to form a continuous path.

It is also possible to bias MD simulation to move between known conformations. For example, Path Exploration with Distance Constraints[3] (PEDC) applies a restraint potential based upon the mass weighted root mean square distance (MRMSD) from a reference structure. This flexible technique can be used to explore possible conformational pathways with the restraint 'pushing' from a starting conformer and 'pulling' towards the target.

Steered MD (SMD) applies an external force to influence the simulation. No target conformer is required, but the direction and amplitude of the force must be defined by the user. SMD forces have been used to represent atomic force microscopy (AFM) experiments,[64] simulations of unbinding processes[65] and simulations showing domain unfolding.[66]

The techniques discussed here are suitable only when a target conformation or response is known, and are of limited use in many computational investigations. Methods that are designed to minimise the use of *a priori* knowledge are more desirable, and are able to probe unknown regions of the potential energy surface.

### 3.3.3 Methods that increase conformational sampling

As discussed, traditional MD simulation starting from a single structure can become trapped in an energy minima. The simplest way to increase the conformational sampling of a system is to raise the temperature and therefore provide kinetic energy to overcome potential energy barriers. However, many properties of molecules change at higher temperatures and proteins may begin to unfold. Some success has been achieved using high temperature molecular dynamics followed by the minimisation of resulting structures.[67] This method does not search the potential energy surface at lower temperatures and thus information is gained only at the minima.

Locally enhanced sampling (LES)[68] is a method that duplicates a region of the system for which enhanced sampling is required. Simulation is then performed with the rest of the system seeing an average force from all the duplicates, in effect generating several trajectories for the chosen region, with a lower computational expense compared to several simulations of the entire system. Conformational events are further enhanced by the use of the average force. For example, if the LES region is duplicated twice, with one region close to an energy minima, and the other approaching a conformational change (and thus at a higher potential energy), the rest of the system would see only an average force, reducing the energy penalty required to overcome a saddle-point on the potential energy surface. However, the LES method can only be performed in conjunction with implicit solvent. Also, the method is not suitable for simulating domain motions, that require rearrangements of large portions of the system.

By extracting information from a simulation in a time-evolving fashion, motions existing in the system can be determined and manipulated. Selectively Enhanced Molecular Dynamics[69] (SEMD) replaces velocities by an average from past time steps so that high frequency motions, which will change sign frequently, can be reduced. A set of coefficients, $C_k$, define the filtering function as shown in Equation 3.1; for high values of $k$, longer time averages are taken corresponding to lower frequency motions. The only published applications of SEMD have taken

all values of $C_k$ equal to $\frac{1}{M}$. Since the velocities are replaced by an average over previous simulation time, $\mathbf{v}'(t)$, there is no phase correlation between coordinates and velocities.

$$\mathbf{v}'_i(t_n) = \sum_{k=0}^{M} C_k \mathbf{v}_i(t_{n-k}) \tag{3.1}$$

Self-Guided Molecular Dynamics[70] (SGMD) is similar to SEMD and applies a guiding force based upon averaged non-bonded forces. This time-evolving force is coupled into the equations of motion, giving continuous simulation properties. Unfortunately, since forces corresponding to fast-moving degrees of freedom are of larger magnitude than those corresponding to slow motions, the guiding force does not necessarily promote the desired low frequency motions.

An advance of SGMD, momentum-enhanced hybrid monte carlo method[71] (MEHMC) addresses this issue by applying a guiding force to momenta rather than forces. This method uses hybrid monte carlo (HMC) in which momenta are randomly assigned from a distribution, and a small number of MD steps are performed. The total energy of the generated conformation is then tested in a similar fashion to MC. The MEHMC method assigns momenta from a distribution that has been skewed to promote momenta corresponding to slowly varying degrees of freedom. One concern highlighted by the authors of this method, is that MEHMC may not be able to overcome energy barriers faced along the degrees of freedom promoted, as momenta are not amplified outside a distribution designed to produce a canonical ensemble. To do this, some method of deforming the potential energy surface, or increasing the system's energy, is likely to be required.

Methods based upon SGMD described so far attempt to amplify slow moving degrees of freedom, without parameterisation based upon frequency analysis. The extraction of frequency information from a trajectory is trivial, and methods that perform post-simulation frequency analysis are simple to implement. For example, by Fourier transforming atomic trajectories to the frequency domain, filtering the result and inverse Fourier transforming to the time domain[72–77] high frequency

motion may be removed and the low frequency motions can be viewed.

Digitally filtered molecular dynamics[78] (DFMD) makes use of frequency information in the time domain to promote existing motions as a simulation evolves. More detail on digitally filtered MD techniques is given in Chapter 5, and a brief summary is presented here. In DFMD, a designed digital filter is applied to a list (or buffer) of atomic velocities from a simulation, suppressing or amplifying motions corresponding to a chosen frequency target. The problem of out of phase coordinates and velocities encountered in SEMD is solved in DFMD by the use of linear phase filters which are symmetrical around the central coefficient. The velocities created by applying a linear phase filter to a buffer are in phase with the coordinates corresponding to the central point of the buffer. It is from this central point that simulation continues after a filter application.

To increase the energy put into the system, a sequence of filter applications is used. However, each filter should be applied before the effects of the last have dissipated. If this is not the case, a set of largely independent filter applications are performed, which cannot progressively manipulate the system. This is a significant limitation of DFMD, as a filter cannot be applied until enough time has been simulated to fill a velocity buffer. The delay between filters is therefore at least half the buffer length.

Reversible digitally filtered molecular dynamics[6] (RDFMD) addresses this limitation and allows delays of any length between filter applications. After applying a filter to a buffer as described with DFMD, and starting with coordinates from the centre of the buffer, a new buffer is filled using simulation both backward and forward in time. The method is not reversible in the sense normally ascribed to an integrator, since velocities created by a filter application cannot be inversely transformed to yield velocities from the buffer from which they were created.

RDFMD is useful for increasing conformational sampling but, since motions are inherently biased, the method is non-equilibrium and will never generate the correct distribution of conformers. The effectiveness of the technique depends upon its parameterisation, particularly on the frequencies of motions to be targeted.

### 3.3.4 Equilibrium methods

If an MD simulation is allowed to evolve in time indefinitely, it will sample all possible phase space, and generate a converged ensemble, or distribution, of states, regardless of the starting conformation. However, available resources limit simulation length, and trajectories often sample one side of a conformational barrier. Techniques, such as those previously described in this section, can be used to increase sampling over such a barrier, but these methods bias simulation so that the correct distribution of states may never be achieved. Generalised or expanded ensemble methods are algorithms that incorporate MC swaps of simulation properties into traditional MC or MD conformational searches.[4] The target is to allow a random walk in potential energy space, and to maintain an equilibrium ensemble of states at a desired set of conditions.

Simulated tempering[79] (ST), also referred to as Method of Expanded Ensemble,[80] allows free random walk in temperature space in a single simulation. The temperature is weighted during the simulation to give a flat probability distribution, thus inducing random walk in potential energy space allowing energy barriers to be overcome.

The Multi-Canonical Algorithm[81,82] (MUCA), also referred to as Adaptive Umbrella Sampling of Potential Energy,[83] is a generalised-ensemble method. Each state is weighted by a non-Boltzmann weight factor (multi-canonical weight factor) giving a uniform energy distribution. The uniformity of this distribution allows a direct random walk in energy space. The weight factor must be derived however, usually from previous conformational analysis work, or from nontrivial iterative histogram-fitting methods. The refinement of these methods has been the source of much research.

The deficit of MUCA and ST lies in the weighting factors required; typically generated through iterations of short trial runs; these are expensive to calculate and often slow to converge. If incorrect or incompletely converged, a random walk is

not achieved.

The Replica-Exchange Method[82,84] (REM), with similar published variants parallel tempering[5] (PT), Multiple Markov Chain Method, Replica Monte Carlo Method and Replica-Exchange Simulated Tempering Method,[85] constructs a generalised ensemble from a number of fully independent simulations running with different physical properties. The most researched MD REM has been performed across kinetic energy space (PT), with a number of independent replicas running NVT simulation at different temperatures. The NPT Metropolis criteria are nontrivial in derivation and have been used in MC simulations,[86] but there are extra barriers to their application in MD due to the slow convergence of pressure in MD simulations.

With parallel tempering, each replica is run for a set period of time after which an MC test is applied against the temperatures and potential energies of the systems. Coordinates can therefore be swapped with those of neighbouring temperatures. By repeating this exchange process, movement in energy space is achieved whilst maintaining statistical ensembles at each temperature. Each temperature therefore gains the information of correctly weighted conformer distributions, and each coordinate replica is available for pathway and frequency analysis. More detail on the PT method is given in Chapter 6.

The potential energy surface does not change at different temperatures, and presently used force fields are optimised to give correct results at a single temperature (typically 298 or 300 K). Raising the temperature has the desired effect of allowing greater motion across the potential energy surface, however it is important to be clear that motion is not occurring across a different surface for the new temperature. The high energy information in a potential energy surface is not optimised in the same fashion as at the minima. At higher temperatures the energy surface is therefore likely to be less accurate. For example, different force fields are suspected to yield different transition temperatures for β hairpin folding.[87] Various force fields have been tested with REM methods and conformer distributions at the temperature for which the force fields are optimised agree with experimental data.[88]

The efficiency of parallel tempering is dictated by the distribution of temperatures at which replicas are simulated. If these temperatures are closely spaced then many replicas are required to reach temperatures at which conformational barriers can be overcome. If too widely spaced, the probability of swapping replicas is low, and the parallelisation of the method is decreased. Since the test to swap replicas is based upon potential energy, a system with a large number of degrees of freedom, which will have a narrower potential energy distribution, requires replicas to be spaced closer together to generate the same acceptance rate of swapping moves as a system with fewer degrees of freedom. For this reason implicit solvent simulations are favoured for use with the REM method.[89,90]

## 3.4 Conclusions

A vast range of methods that probe possible protein conformations and the dynamics required to move between these, have been discussed. No single method is suitable for all conformational analysis applications, and the methodology for many of the computational techniques is still evolving.

Reversible digitally filtered MD is a non-equilibrium technique for the amplification or suppression of motions with a specific frequency in simulation. Two targets must be met for the development of RDFMD as a tool to increase the rate of conformational change. Firstly, a 'recipe' for the parameterisation of the technique to any application is required. Secondly, parameter sets suitable for specific uses in conformational investigations must be produced and validated. Both these targets are addressed and met in this thesis.

The parallel tempering algorithm targets an equilibrium distribution of conformations, allowing energy barriers to be overcome by changes in system temperature. The method is computationally expensive, but has been widely adopted in conformational investigations.

# Chapter 4

# The application of signal analysis methods to molecular dynamics

## 4.1 Introduction

From a molecular dynamics simulation a number of frequency signals can be extracted and monitored for comparison with experimental data[14,17] or to provide insight into the nature of motions within the system.

The majority of intramolecular motions are wave–like in nature and exist either as localised oscillations in single bonds or angles, or as collective motions (or modes) characterised by coupled vibrations. For example, the backbone amide I mode in proteins that originates from the carbonyl stretching vibration also includes a contribution from the backbone carbon–nitrogen bond.[91] There is also abundant evidence from experimental sources indicating the existence of low frequency motions occurring on molecular scales.[49,50] By increasing understanding about the range of motions, and their relevance to conformational change, the frequencies most suitable as targets for the RDFMD method should be revealed.

In this chapter, a range of signal analysis methods are explored, in the context of molecular dynamics. Methods later used to assist the parameterisation of RDFMD are introduced, and examples given. Many of the results presented here have included in a recently accepted publication (Wiley *et al.*)[92]

## 4.2 Digital signal processing

Digital signal processing (DSP) refers to the investigation of a digital (i.e. discrete) or continuous (i.e. analogue) signal via digital sampling. Such signals are abundant in nature and technology, and much work has been performed on their analysis.

DSP is dominated by well known Fourier techniques, however the Fourier transform (FT) has several major deficiencies that limit its use. The Hilbert transform (HT)[7] has only recently been proposed for application to molecular dynamics[93] and it is explored here. The HT is only suitable for application to signals of a very specific nature, and the method must usually be coupled to a technique that decomposes an input so that it satisfies the necessary criteria. The Empirical Mode Decomposition is a technique capable of doing this, and its use with the Hilbert transform has been recently named the Hilbert-Huang transform[94] (HHT).

## 4.3 Linear and stationary signals

It is first useful to define two commonly used terms that help classify the nature of signals: linear and stationary.

A signal may be described as linear if it has linear phase; in other words the signal obeys $\frac{d\phi}{dt} = k$, where $k$ is a constant and $\phi$ is the phase. The definitions of frequency and amplitude for a sinusoidal linear signal are simple, as they must be constant throughout time. Frequency can be measured as the number of complete cycles per unit time, and amplitude by the maximum value of the periodic curve.

A stationary signal is, strictly speaking, one for which all statistical properties are invariant to a shift of the time origin. Generally a stationary signal can be defined as one whose frequency components do not change with time. Thus the linear sine wave $y = \sin(t)$ is stationary, as is the nonlinear combination of

two linear sine waves, $y = \sin(t) + \sin(2t)$, since the frequency description of the signals do not change with time. A wave with varying frequency such as $y = \sin(t^2)$ is both nonstationary and nonlinear. Loss of stationarity is an important sampling issue when considering a portion of a stationary signal that has been replicated to infinity. If, and only if, the beginning of the signal is in phase with the end, will stationarity be maintained. Otherwise, a phase discontinuity is introduced, and frequency domain behaviour can no longer be regarded as time-invariant.

## 4.4 Fourier methods

The Fourier transform requires input data to be both linear and stationary. An input signal $x(t)$, is initially reproduced to infinity, so that the beginning of the data is added to the end (i.e. $x(t + nT_0) = x(t)$ where $T_0$ is the length of the input signal and $n$ is any integer). Even if the initial signal was stationary, once replicated to infinity, this may no longer be the case as previously described.

The signal, regardless of its form, is then decomposed into a set of sine waves with fixed frequency and amplitude. The contribution of a specific frequency, $\omega$, to the initial signal can be found using the Fourier transform (Equation 4.1). The Fourier transform multiplies an input signal by an exponential function, $e^{-i\omega t}$, which can also be represented by a complex waveform, $\cos(\omega t) + i\sin(\omega t)$. By integrating across time, the relevance (or spectral component), $F(\omega)$, of the frequency, $\omega$, is calculated. The result of a Fourier transform is usually represented as a frequency, or Fourier, spectrum, in which the magnitude of $F(\omega)$ is plotted against $\omega$.

$$F(\omega) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt \tag{4.1}$$

The initial signal can be recreated from its Fourier spectrum by an inverse Fourier transform, as shown in Equation 4.2.

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}d\omega \tag{4.2}$$

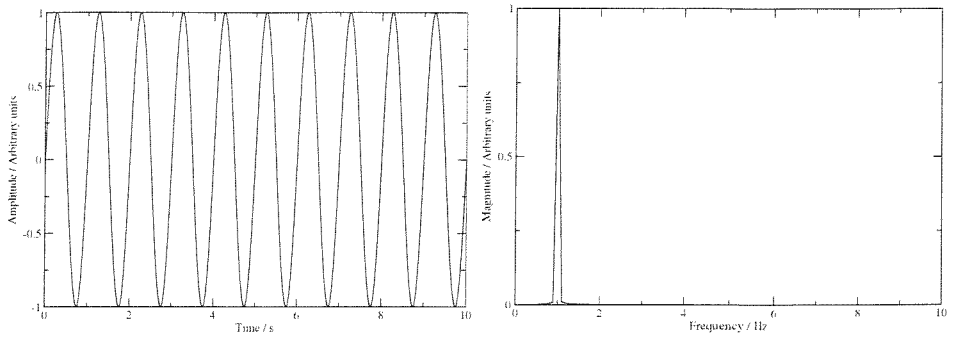
The Fourier transform method therefore calculates the relevance of a frequency to an input signal but gives no time localised information. If a nonstationary signal with an abrupt change of frequency is defined, such as that in Equation 4.3, then as the Fourier transform integrates across time, two dominant spectral components will be located at frequencies corresponding to those found in the signal. A number of other components with lower amplitudes will also be seen that attempt to describe the sudden change of frequency. These are mathematically present in the signal, in the sense that Equation 4.2 still holds, however they have no physical relevance, having been created to constructively and destructively combine to create a non-sinusoidal function.

$$\begin{cases} y = \sin(2\pi t) & if\ t < 0 \\ y = \sin(4\pi t) & if\ t \geq 0 \end{cases} \tag{4.3}$$

The limitations of the Fourier transform are best described using a few examples. Figure 4.1 (a) shows a sine wave with a period of 1 second (i.e. a frequency of 1 Hz) that has been sampled every 0.01 s. 10 s of data have been sampled and a discrete Fourier transform applied. The result is a single peak, as expected, at 1 Hz, with excellent frequency resolution, shown in Figure 4.1 (b). The signal is both linear and stationary so no problems are encountered with the Fourier transform.

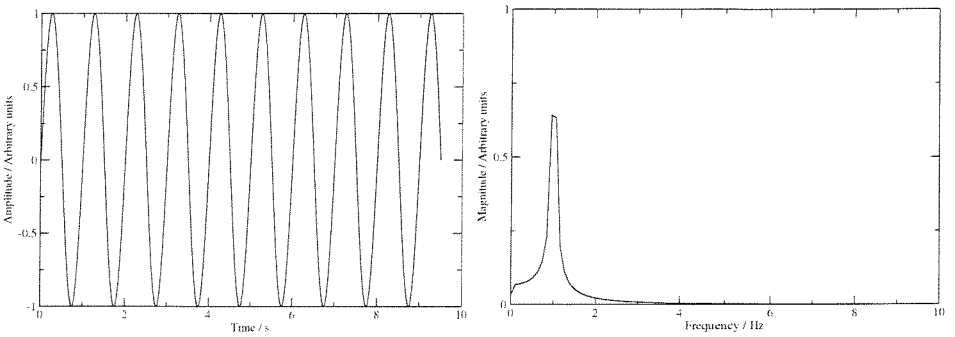
Figure 4.2 (a) shows a similar sine wave, with a frequency of 1 Hz, however only 9.5 s have been sampled. The linear wave is now nonstationary when reproduced to infinity as a discontinuity has been introduced. The frequency resolution shown in 4.2 (b) is clearly degraded, and the magnitude of the peak at 1 Hz has been reduced.

Figure 4.3 (a) shows an initially nonstationary signal as described in Equation 4.3 in which two frequencies are included in the data. The FT shown in Figure 4.3 (b) clearly locates the two frequencies as dominant spectral components. However the frequency resolution is severely lowered, with classic 'lobe' effects




(a) Signal (b) Fourier transform

Figure 4.1: Signal and Fourier transform of 10 periods of a 1 Hz sine wave.



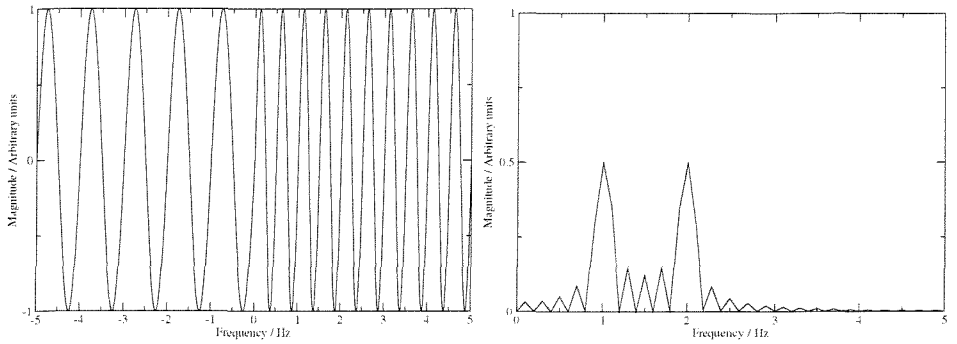

(a) Signal (b) Fourier transform

Figure 4.2: Signal and Fourier transform of 9.5 periods of a 1 Hz sine wave.

that are often seen in FTs.

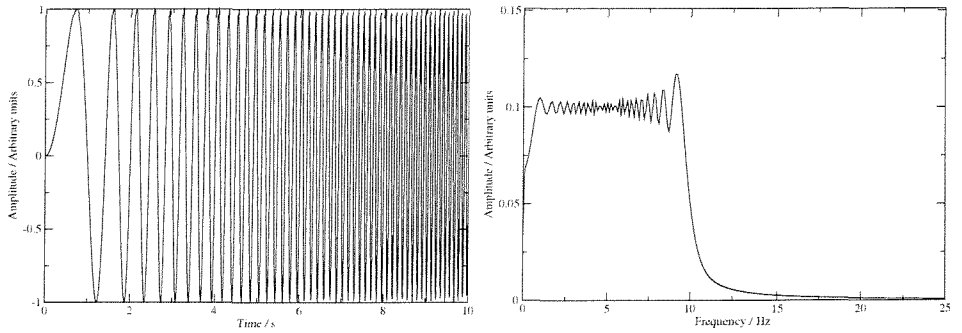
A type of signal that is poorly analysed by FT is one in which the frequency is changed as the signal is sampled. For example, Figure 4.4 (a) shows a sine wave that starts with a frequency close to 0 Hz, which is then linearly increased in frequency to 10 Hz over 10 s. As Figure 4.4 (b) shows, the Fourier transform provides very limited information. When a signal is spread across a spectrum, and its amplitude reduced, it can become difficult to discern from background 'noise'.

A variety of Fourier methods exist to address problems of time-varying signals. If a signal is nonstationary, but can be assumed to be stationary over a length $w$, then a windowing function of width $w$ can be applied across the data. The window is a mathematical construct that typically biases the data toward the central point, so that the contribution of the end regions is reduced, without




(a) Signal (b) Fourier transform

Figure 4.3: Signal and Fourier transform of signal described by Equation 4.3 and sampled from $t$ = -5 s to $t$ = 5 s.




(a) Signal (b) Fourier transform

Figure 4.4: Signal and Fourier transform of 10 s signal, starting with a frequency close to 0 Hz, and increasing to 10 Hz in a linear fashion.

introducing a sudden phase discontinuity. The window is applied to the first $w$ points of the data, and an FT calculated. The window is then moved along the data, and applied again. This method is known as the short-time Fourier transform (STFT) and the resulting spectra is referred to as a spectrogram. If the window is applied to a region of stationary data, frequencies are well resolved. The effects of a time-varying event are localised to windows applied close to the event. Naturally if the signal is never stationary, as frequencies in the system are constantly evolving, a spectrogram of the data will always contain inaccuracies.

The main deficit of STFT lies in the reduction of the data set's length. Since only a subset of points are analysed, discontinuities introduced into the infinitely replicated data set will be more frequent, and the lowest possible frequency

sampled by the method will be increased. Identification of the significant spectral components becomes less clear, thus decreasing frequency resolution whilst delivering superior time resolution. If $w$ is increased, the inverse is true, and greater frequency resolution is gained, at the expense of time resolution. The balance of frequency and time resolution is key to digital signal processing.

To address this issue, the wavelet transform[95] (WT) was developed in which the width of the applied window is changed as the transform is computed. Wavelet transformations require parameterisation, and are a significant field in themselves. The WT has previously been compared to FT and HHT[93] and will not be further discussed here.

Fourier-based methods are used for most DSP analysis, however the issue of frequency-time resolution has not been adequately resolved for nonlinear and nonstationary data. Signals extracted from molecular dynamics simulations are rarely stationary or linear, although some examples can be found. Physical assignment of frequencies to conformational motions, as shall be described later, challenge the resolution limits of FT and an alternative approach is required.

## 4.5 Hilbert-Huang techniques

### 4.5.1 The Hilbert transform

The Hilbert transform (HT) takes a time-domain signal and transforms it into another time-domain signal, unlike the Fourier transform which transforms from the time to the frequency domain. The HT of a real-valued function, $x(t)$, over the range $-\infty < t < +\infty$ is another real-valued function, $h(t)$, which is the convolution (see Appendix A.1) of $x(t)$ with $\frac{1}{\pi t}$ (Equation 4.4). A function of this form carries a discontinuity at the limit $t = u$ however $P$ signifies the Cauchy Principal Value for which a value of the transform can be calculated (see Appendix A.2). The rapid decay of $\frac{1}{t-u}$ biases the transform heavily to points close to t, and thus the HT acts on time-localised data.

$$h(t) = P\int_{-\infty}^{\infty} \frac{x(u)}{\pi(t-u)}du \tag{4.4}$$

In practice, the Hilbert transform can be computed by taking the Fourier transform of the data, setting all of the negative frequency components to zero, doubling the positive frequency components, and performing an inverse Fourier transform.[7] This method relies on the infinite replication of the data set, as for the Fourier transform, yielding unreliable regions at the start and end of the transformed data due to the discontinuities introduced.

It can be shown that the signal magnitude is unchanged by the Hilbert transform, but the phase is adjusted by $\frac{\pi}{2}$.[7] The original signal, $x(t)$, and its Hilbert transform, $h(t)$, may be considered part of an analytic signal, $z(t)$, represented by a complex function as shown in Equation 4.5.

$$z(t) = x(t) + ih(t) \tag{4.5}$$

The analytic signal can also be written in terms of amplitude, $A$, and phase, $\phi$:

$$z(t) = A(t)e^{i\phi(t)} \tag{4.6}$$

where

$$A(t) = \sqrt{x^2(t) + h^2(t)} \tag{4.7}$$

$$\phi(t) = tan^{-1}\left(\frac{h(t)}{x(t)}\right) \tag{4.8}$$

The analytic signal can therefore be used calculate a phase angle as a function of time. The rate of change of this phase is the instantaneous frequency, $f(t)$, as shown in Equation 4.9. The validity of a calculated instantaneous frequency is highly dependent upon the nature of the signal. If, for example, the phase of a sum of linear signals were differentiated, then the instantaneous frequency would be

meaningless.

$$f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{4.9}$$

An instantaneous frequency is meaningful for signals that are subject to the following constraints:[7]

- The signal must represent a single motion with no carrier or riding waves
- The signal must be sinusoidal with a slowly varying frequency
- The amplitude of the signal must vary slowly with time

An alternative way to consider the last two criteria would be to say 'A signal must be locally symmetrical about its mean point.'

Signals that meet these criteria are rare in real-life, a fact that has significantly limited the use of the Hilbert transform. Fortunately, it is possible to split a signal into components that are suitable for the HT, using Empirical Mode Decomposition.

### 4.5.2 Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) is a recently developed signal processing method.[8] It decomposes a signal, that may be nonstationary, into a set of intrinsic mode functions (IMFs). An IMF is defined as a sinusoidal wave in which the number of extrema and the number of zero-crossings differ by a maximum of one and where, at any point, the mean of the envelope defined by the local maxima and minima is zero. Thus IMFs meet the criteria required to apply the Hilbert transform.

The EMD algorithm can be split into two parts: one that extracts an IMF from the data set, and one that controls the decomposition. The control of the decomposition is termed 'sifting,' and can be described as follows:

1. Find an IMF from the data set.

2. Subtract the IMF from the data set.

3. If there is only one maximum and one minimium remaining in the data set, decomposition is complete; stop.

4. If the magnitude of the recovered IMF is below a predefined cutoff, further decomposition is unnecessary; stop.

To extract an IMF from the data set, the following sequence is repeated:

1. Cubic spline curves are fitted to the maxima and minima of the data set.

2. The difference of the two spline curves is calculated and subtracted from the data set.

3. A check is made against possible stopping criteria. If either are met, an IMF has been created.

    - If the number of extrema and zero-crossings differ by 1 or 0, have the numbers been unchanged for the last $S$ iterations?

    - Is the integral over the data set below a cutoff?

Terminating the IMF creation procedure is an important step. If no stopping criteria are imposed, the result will be an IMF with constant amplitude and therefore no physical meaning. Personal communications with Prof. N. Huang indicated that the best stopping criteria involves testing only the number of extrema and zero-crossings. If these are unchanged for 5 iterations, the procedure should be stopped. This value for $S$ was determined on data covering a range of applications

and measuring the orthogonality of the produced IMFs, a topic that shall not be discussed here. All EMD applications presented in this thesis use this stopping criteria.

When no further IMFs can be extracted from a data set, a data residual remains. This is similar in concept to the constant (zero frequency) component of a Fourier transform. However, the residual contains the overall trend of the data, and may not necessarily be a constant.

Once a signal has been decomposed into a set of IMFs and a residual, each IMF is separately Hilbert transformed, giving the amplitude and instantaneous frequency at discrete points in time. If the IMF represented a signal from a harmonic oscillator of variable amplitude and frequency, then its signal energy at time $t$ can be written in terms of the signal amplitude, $A$, and frequency, $\nu$ (see Appendix B).

$$E(t) \propto \nu(t)^2 A(t)^2 \tag{4.10}$$

The IMF extraction procedure involves fitting a cubic spline up to the ends of the data set. This introduces numerical errors, as points must be extrapolated beyond the edges of the signal. If viewing IMFs prior to HT, the end-effects should be considered. The HT of IMFs also contains inaccuracies in the end regions of data due to the discontinuities introduced by the infinite replication of data (since FTs are used to implement the HT), and these regions are generally excluded from consideration.

Although the EMD algorithm does not require an input of stationary data, the input signal should not contain components that overlap in frequency. If it does then since the algorithm attempts to create IMFs that follow the highest frequency remaining in the data set, a generated IMF will pass from following one component, to another component that overtakes the frequency of the first. At the point of overlap, artefacts are introduced and frequency resolution is reduced. The algorithm does recover however, and this limitation does not prove a serious flaw of the method.[93]

### 4.5.3 The application of Empirical Mode Decomposition and the Hilbert transform to test signals

While it is trivial to show the success of EMD and HT to linear and stationary data, analysis of the signals for which the limitations of the Fourier transform were shown shall be presented here. Figure 4.3 shows a signal that suddenly changes frequency at $t = 0$, leading to lobes on the Fourier spectrum. These are introduced due to the method attempting to assign time-invarient sine waves to the entire data set. EMD, which produces IMFs with frequencies that can change with time, follows the signal well with a single IMF shown in Figure 4.5. Several IMFs of low amplitude were generated due to the end-effects previously discussed. Figure 4.6 shows the Hilbert transform for all IMFs, with excellent resolution of the signal's change in frequency. The scale, or relevance, of the frequency to the original signal is shown using colour. Since the IMF's frequency must vary 'slowly' according to definition, there are ripples in the frequency after the sudden jump from 1 to 2 Hz. These quickly settle and are not accompanied by significant introduced artefacts. Several IMFs are introduced by the various sources of error (discontinuities in the replicated data set, extrapolation of the spline procedure beyond the edges of the data, and the artefacts introduced by the sudden change of frequency) and can be seen at low frequencies. The relevance of these IMFs, when compared on the scale used to describe the signal's motion, is very small.

The Fourier transform is unsuitable for analysis of frequency-modulated signals as shown in Figure 4.4. Again however, EMD produces a single, dominant IMF that follows the signal, shown in Figure 4.7, and several low amplitude IMFs due to end-effects of the spline-fitting procedure. The Hilbert transform of all IMFs again shows excellent time and frequency resolution, as can be seen in Figure 4.8. Whilst this signal is the hardest to analyse with Fourier methods, analysis using the HHT introduces no significant errors.

As previously discussed, it is a requirement of EMD that the frequency scales of motions present in the signal are separable. If this is not met, IMFs must transfer from one motion to another, as they always follow the highest frequency signal.

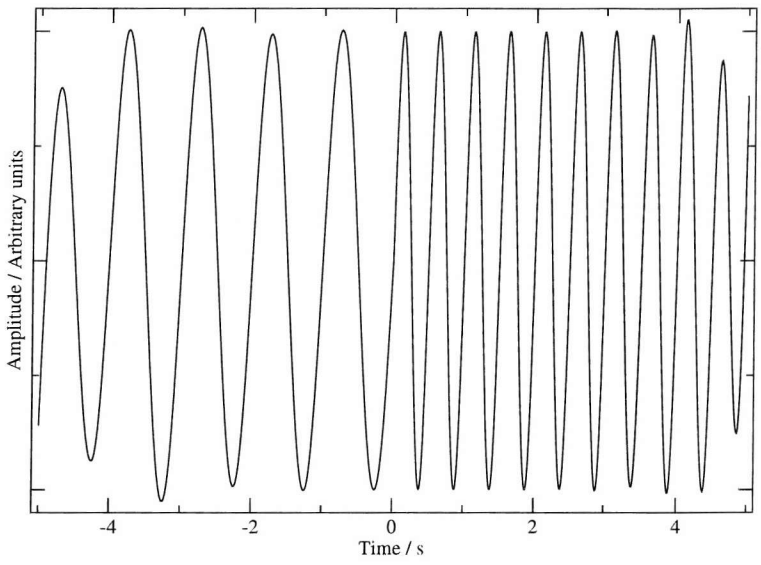



Figure 4.5: The highest frequency IMF produced by EMD on the signal shown in Figure 4.3 (a).

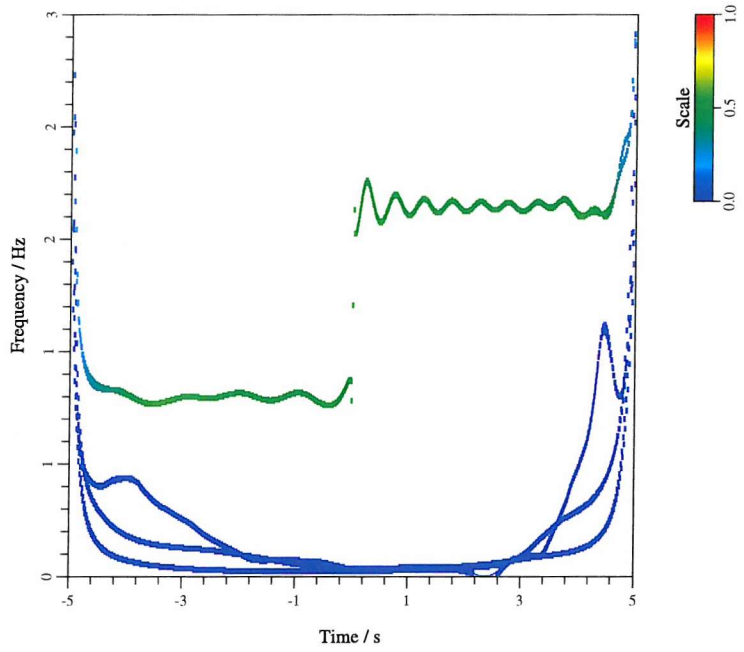



Figure 4.6: The Hilbert transform of the signal shown in Figure 4.3 (a). The colour scale represents signal amplitude.

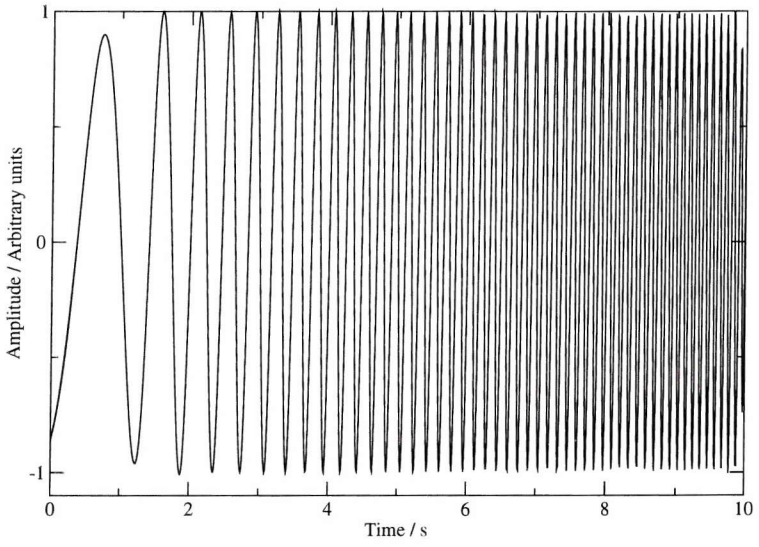



Figure 4.7: The highest frequency IMF produced by EMD on the signal shown in Figure 4.4 (a).

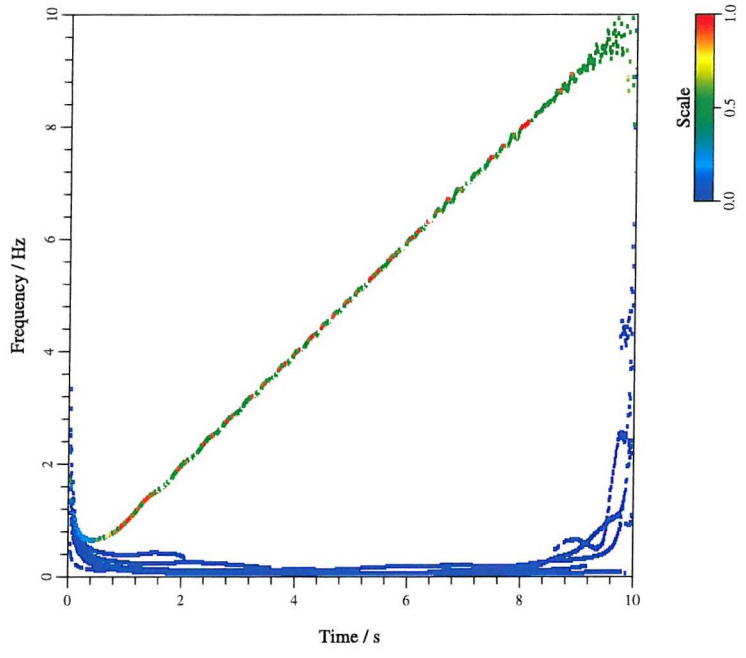



Figure 4.8: The Hilbert transform of the signal shown in Figure 4.4 (a). The colour scale represents signal amplitude.

Figure 4.9 shows a 5 Hz sine wave that has been added to the frequency-modulating signal shown in Figure 4.4. Two dominant IMFs, produced by EMD, are also shown in Figure 4.9 and the Hilbert transform of all IMFs is shown in Figure 4.10. Even when the signals cross, the dominant spectral components lie in the correct frequency region. On either side of the crossing the transform is well resolved.

The end effects of the EMD algorithm and the Hilbert transform are noticable in Figures 4.6, 4.8, and 4.10, where low amplitude signals appear to 'climb' up the frequency axis at the beginning and ends of the data set. Only small regions of data are affected, and hereafter these regions are excluded from analysis.

## 4.6 The application of HHT to molecular dynamics

### 4.6.1 The YPGDV test case

YPGDV has been used as a test case for RDFMD and the digital signal analysis methods described here. Figure 4.11 shows the structure of the pentapeptide with dihedral angles of importance labelled. Previous studies of YPGDV include conformational analysis,[96] and self-guided molecular dynamics.[97] *Trans*-proline NMR data in water is known to show an approximately equal proportion of reverse turn and extended conformations at 273 K.[98]

A YPGDV starting structure was obtained that had been set up from an all-trans Z-matrix, generated and solvated within the MCPRO package[99] using 805 water molecules and 1 sodium ion. From this, simulation was performed using the NAMD package[100] with a switching function applied to the Lennard Jones interactions between 8 and 12 Å, a particle mesh Ewald treatment of electrostatics,[20] and SHAKE[25] was applied to all bonds involving a hydrogen atom with a tolerance of $10^{-8}$ Å. Cubic periodic boundary conditions were used throughout. Explicit water was modelled by the TIPS3P water model as implemented by CHARMM and the protein described by the CHARMM22 force field.[17]

Initially, minimization was performed with the conjugate gradient line-search

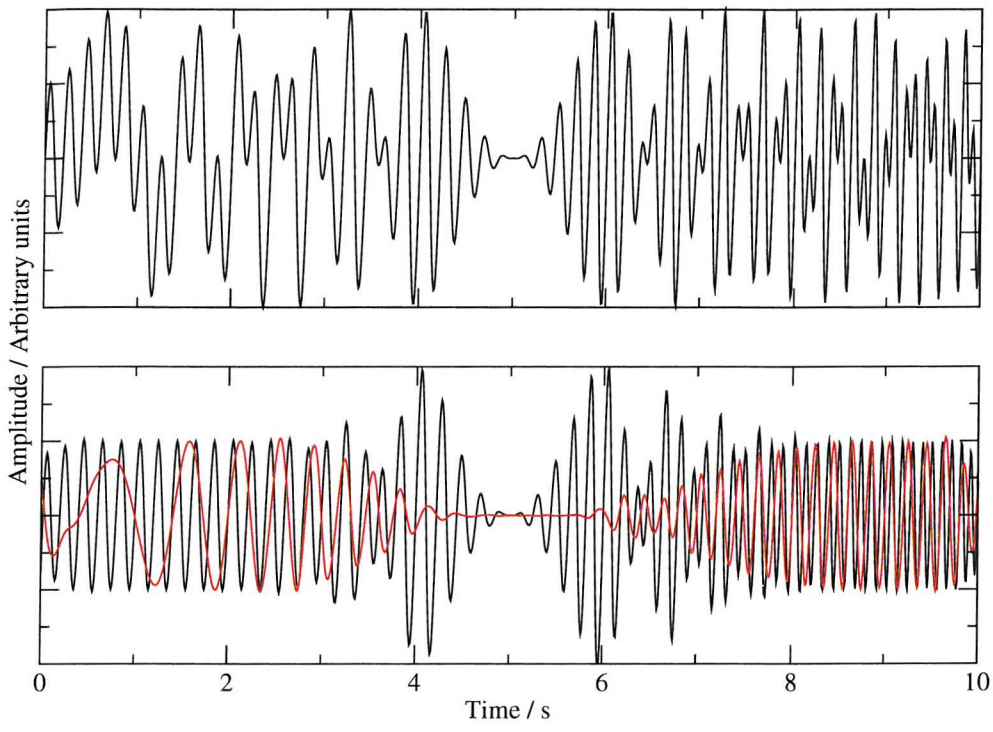



Figure 4.9: Data produced from the sum of a 5 Hz signal with the frequency-modulating signal shown in Figure 4.4 (a). Top: Signal, Bottom: Dominant IMFs

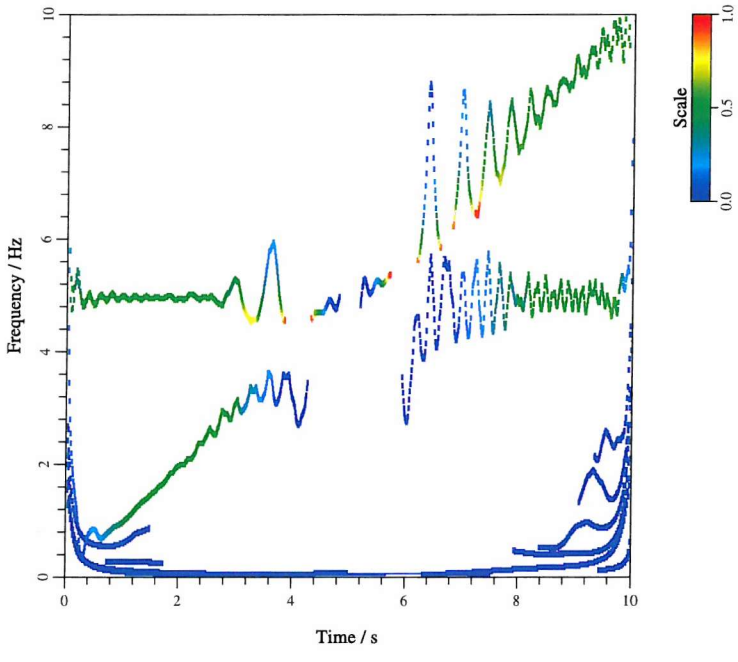



Figure 4.10: The Hilbert transform of the signal shown in Figure 4.9 (a). The colour scale represents signal amplitude.

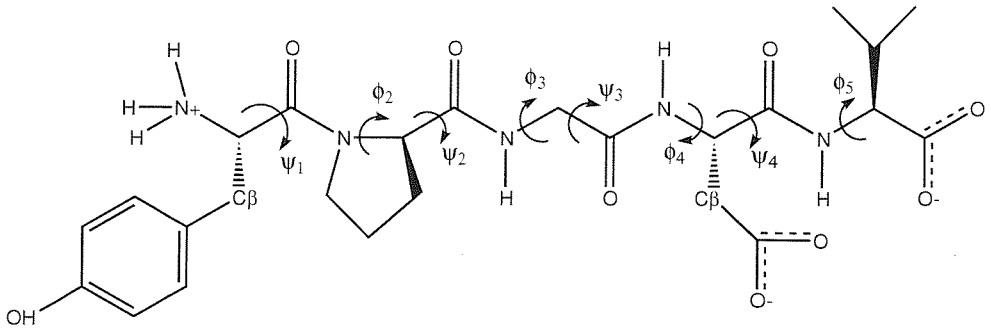



Figure 4.11: Pentapeptide YPGDV (Tyr–Pro–Gly–Asp–Val) with backbone dihedral angles $\psi_i$ and $\phi_i$ labelled.

algorithm[100] for a total of 22 000 steps. The system was then gradually heated with a 20 000 step canonical simulation at each temperature between 50 and 300 K at 50 K intervals using a 2 fs time step. A Langevin thermostat[101] was used with a damping parameter of 10 $ps^{-1}$. 80 000 steps were then performed at the target 300 K followed by 500 000 steps (1 ns) in the isothermal-isobaric ensemble. A Nosé-Hoover Langevin piston barostat[102] with a pressure target of 1 atm, a piston temperature of 300 K, a damping decay parameter of 200 fs and an oscillation period of 400 fs was used. A further 200 000 steps (400 ps) of isothermal-isobaric MD were then performed with a 1 $ps^{-1}$ thermostat damping parameter, 300 fs barostat damping decay and 500 fs piston oscillation period. Several simulations from this point have been performed, and the system state shall hereafter be referred to as the equilibrated YPGDV system.

Results reported in this chapter have been generated from a 16 678 step microcanonical (NVE) ensemble simulation performed from the equilibrated YPGDV system. A timestep of 2 fs is used, and the length of the simulation allows for a minimum frequency of 1 $cm^{-1}$ to be sampled. This trajectory has contributed to a publication[93] in which some frequency analysis was performed. The results here are presented in greater detail and are entirely the work of the author.

### 4.6.2 HHT analysis of YPGDV

The versatility of computer simulation allows the selection of specific internal coordinates for analysis. For example, it is trivial to extract the progression of

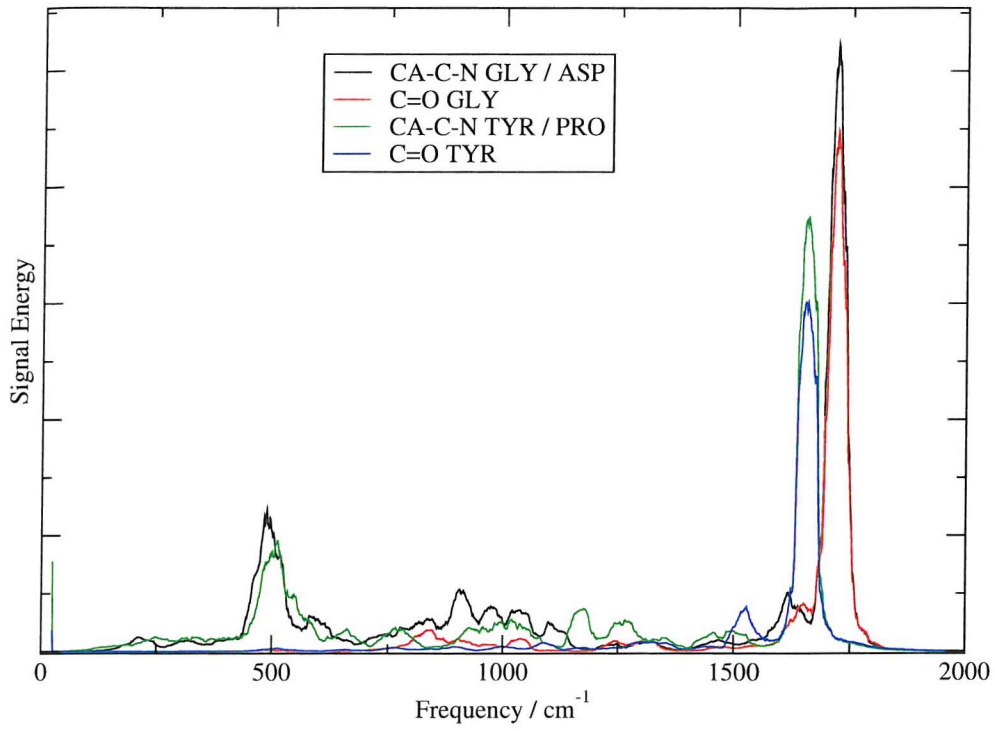



Figure 4.12: Fourier transform of backbone dihedral angles using a Hanning window and a 50 point running average.

an angle or bond, and locate the frequencies of motion with a windowed Fourier transform. In Figure 4.12 the Fourier transforms of the glycine–aspartate CA–C–N and associated C=O signals from the NVE simulation trajectory are shown. These harmonic motions are reasonably linear throughout the simulation and the analysis extracts the expected results: the amide I vibration between 1600 and 1800 $cm^{-1}$, the angular motion at around 500 $cm^{-1}$ and collective backbone motions between 800 and 1200 $cm^{-1}$. The angular motion has a large component at the bond stretching frequency of 1600 to 1800 $cm^{-1}$ showing that the angles are coupled to the bond stretches. However, the bond motion has no low frequency component.

The HHT can continue analysis to greater depth. In Figure 4.13, the HHT for the angle bending CA–C–N vibration between glycine and aspartate is shown, and in Figure 4.14, the spectrum is also presented for the C=O bond stretching vibration. The Hilbert spectra show energy flowing into and out from the high frequency vibrations at the same times, demonstrating the coupled nature of these vibrations. In Figure 4.15, the highest frequency IMFs of the glycine–aspartic acid backbone angle (CA–C–N) and of the glycine carbonyl length (C=O) are plotted,

and are clearly in phase. Subtracting the first IMF from the original signal and taking the Fourier transform of the result shows that this IMF is responsible for all the vibrational motion above approximately 1250 $cm^{-1}$ (not shown).

As previously discussed, the signals of most relevance to conformational change are the low frequency dihedral motions. These are generally nonstationary and Fourier analysis provides a poor description of conformational events.[93] By looking at the energy in low frequency regions as a function of time, Hilbert techniques allow recognition of events occurring simultaneously across several dihedrals. In Figure 4.16, the eight main-chain dihedral angles of YPGDV that are most important for conformational change (see Figure 4.11) are plotted as a function of time in a region where a spontaneous conformational transition occurred. The energy in the 0–10 $cm^{-1}$ region obtained by integrating the HHT spectrum is also shown, and the three highest energy peaks occur during 15–22 ps, corresponding to a rearrangement of dihedrals $\psi_2$, $\phi_3$, and $\psi_3$. Several dihedrals are seen moving at once, reducing the need for significant solvent reorganisation.

It is necessary to show that the Empirical Mode Decomposition method produces physically relevant IMFs. Results for the $\phi_3$ data of the spontaneous transition are shown in Figure 4.17. The low frequency IMFs follow physical motions from the signal and there is an obvious separation between the high frequency (IMFs 1 to 5) and low frequency motions (IMFs 6 to 10). The application of EMD as a signal smoothing technique is also clear.

As previously mentioned, it is a requirement of the EMD algorithm that frequencies of motion present in the system are separable. Meaningful separation is not always guaranteed for dihedral data due to the flexible nature of the simulation model, and so EMD results must always be interpreted with caution. However the physical relevance of IMFs has been shown and, as is shown in Figure 4.17, a single IMF often describes conformational events. Sometimes the frequency separation of IMFs is not complete and the physical relevance of a single IMF is not clear. A blind analysis of a single IMF is therefore not sensible, but by averaging results over several signals or by integrating over long time periods, a

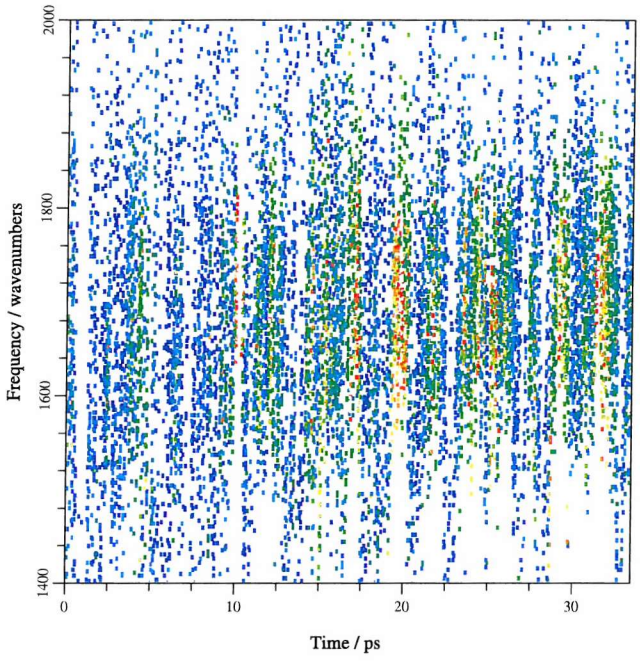



Figure 4.13: HHT of the bond angle trajectory CA–C–N.

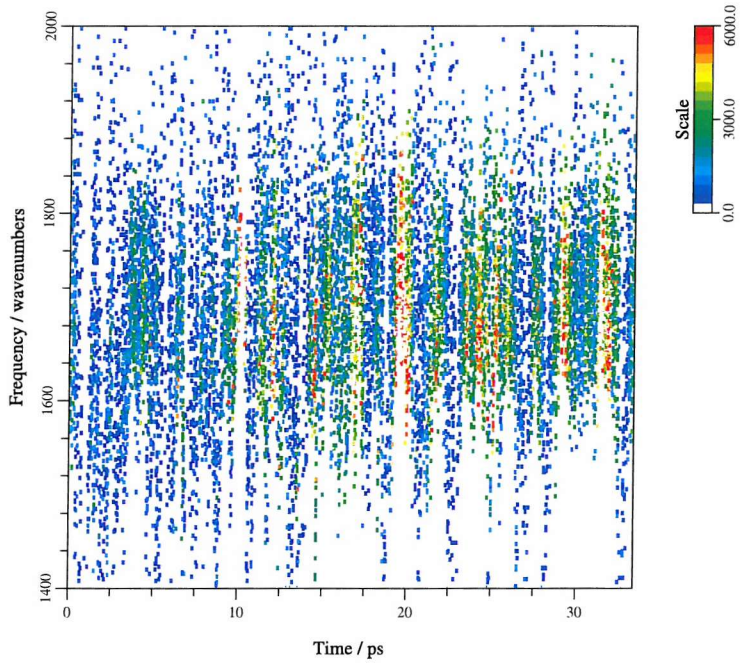



Figure 4.14: HHT of the bond length trajectory C=O. Arbitrary units are used for the scale axis.

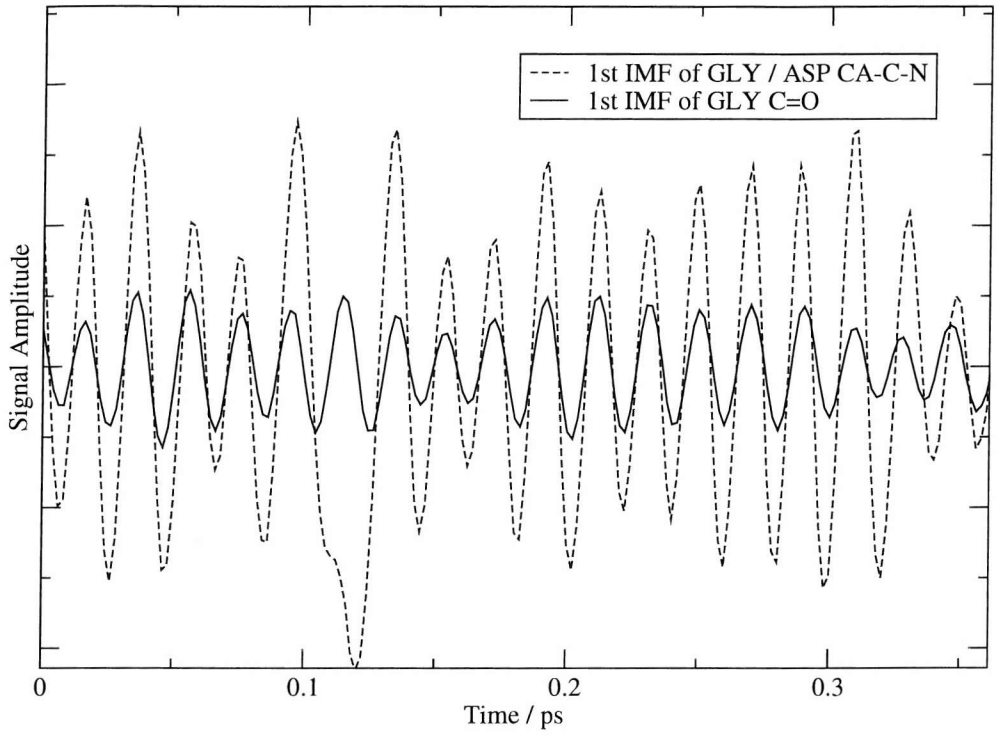



Figure 4.15: The first IMFs of the coupled angle, Gly–Asp CA–C–N, and bond motion, Gly C=O.

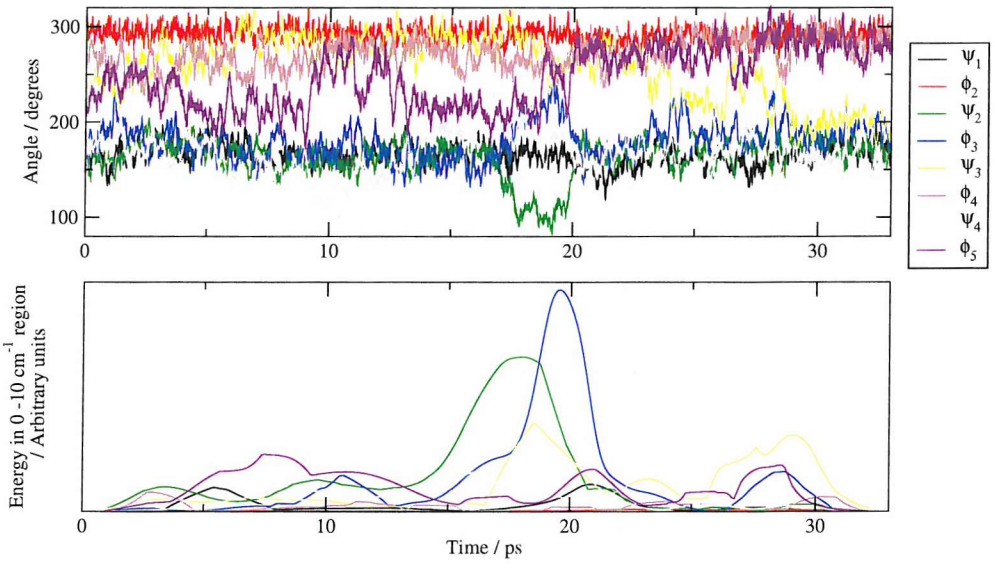



Figure 4.16: Top: The 8 dihedral angle trajectories around a spontaneous conformational transition. Bottom: The energy in the 0–10 $cm^{-1}$ region extracted using HHT.

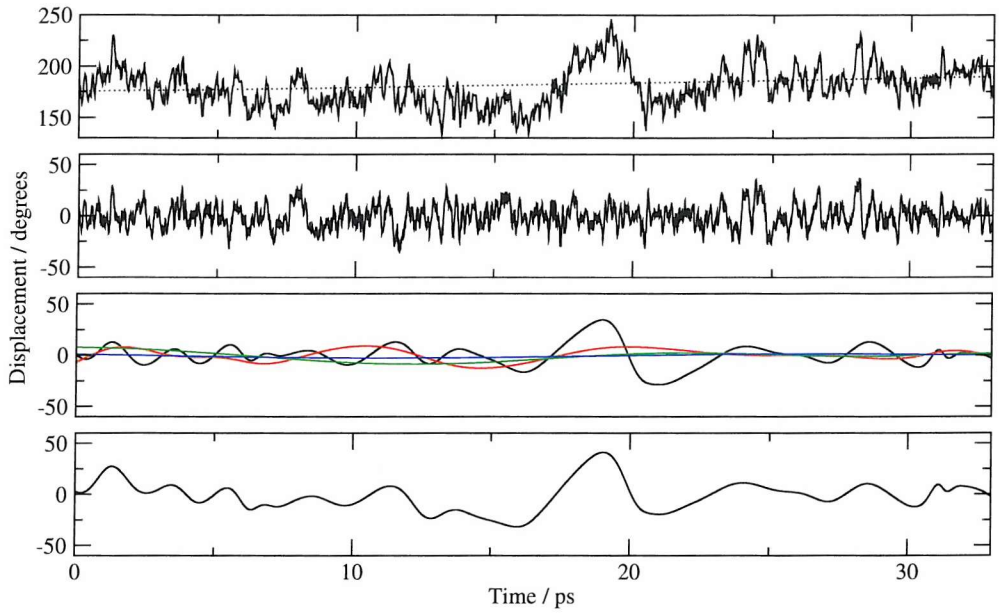



Figure 4.17: YPGDV $\phi_3$ data during a trajectory containing a spontaneous conformational transition. Top: Signal (solid line) and residual (dashed). Upper middle: Sum of high frequency IMFs (1 to 5). Lower middle: Low frequency IMFs (6 to 10). Bottom: Sum of low frequency IMFs (6 to 10).

more robust analysis may be obtained.

## 4.7 Limitations of the Hilbert-Huang transform in molecular dynamics

There are three significant limitations of the HHT method as applied to molecular dynamics simulations. First, frequency analysis must generally be performed on short simulations as the high sampling rates required for good frequency resolution lead to large output files. The shorter the simulation, and thus data set, the higher the low frequency limit will be. For example, storing coordinates for an RDFMD buffer of 1001 coefficients generated with a 2 fs timestep, results in a low frequency limit in the Hilbert transform of 16.7 $cm^{-1}$. Given that this limit lies in the frequency range over which amplification is likely to be required, useful analysis is limited.

The second limitation of the application of HHT to MD, is that conformational transitions can produce large residuals, with dihedral angle plots appearing similar to step functions. EMD produces physically unrealistic IMFs in an attempt to

recreate the unusual signal given by the data less the residual. A severe case of this effect is shown in Figure 4.18. Although the data could be windowed to reduce the size of the affected regions, low frequency resolution would be lost. This limitation is due to the nature of the data and suggests that the HHT must be used with care, constantly checking the physical relevance of IMFs.

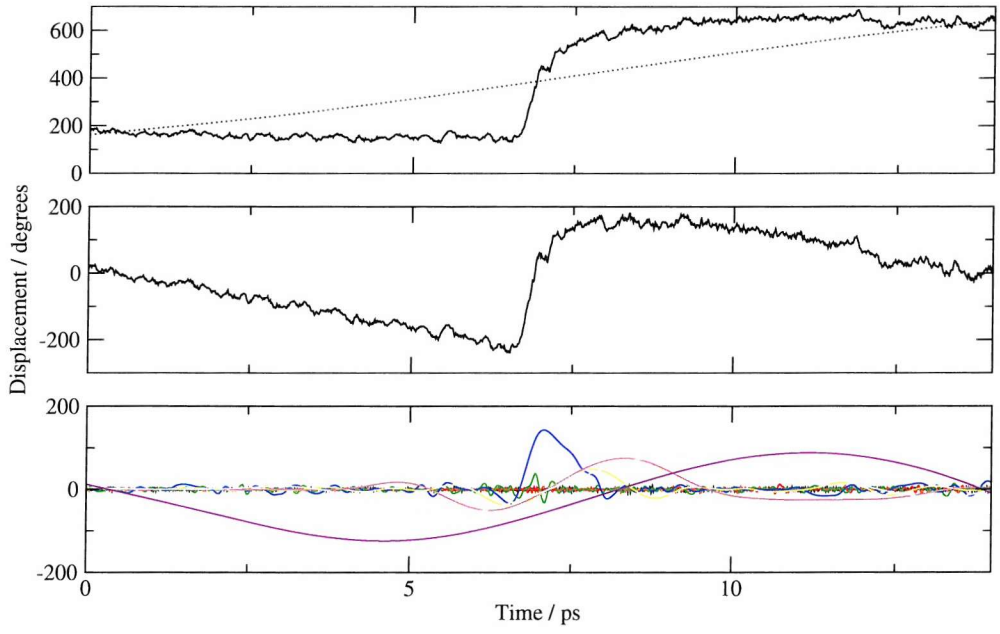



Figure 4.18: YPGDV $\phi_2$ data. Top: Signal and residual (dotted line). Middle: Signal after the residual is removed (equivalent to the sum of the IMFs). Bottom: IMFs generated by EMD.

The third limitation of HHT, is the EMD requirement of separable scales which is not necessarily met by signals from molecular dynamics simulations. Great care must therefore be taken when performing analyses by, for example, careful examination of the IMFs to confirm their physical relevance, repeated analysis over many different simulations, averaging over a number of simulation signals, or integration of the HHT spectra over frequency and time.

## 4.8 Conclusions

In this chapter examples of Hilbert-Huang techniques have been presented, yielding information over and above that available by a conventional Fourier analysis. The coupled nature of vibrational motion in molecular systems has

been clearly shown, as has the ability of EMD to seperate signals of different frequencies. The analysis of nonstationary conformational change events, that are unsuitable for analysis by Fourier techniques, has revealed motions associated with low frequency vibrations in the region 0–10 $cm^{-1}$. The decomposition of dihedral angles and the physical relevance of the IMFs obtained have been shown. The possibility of large residuals and incomplete separation of frequency scales indicate that HHT cannot be applied to molecular dynamics as a 'black box' and that, as with Fourier methods, care must be taken.

# Chapter 5

# Reversible Digitally Filtered Molecular Dynamics

## 5.1 Introduction

Reversible Digitally Filtered Molecular Dynamics (RDFMD) amplifies or suppresses motions of specific frequencies in a time-evolving MD simulation. This project focuses on the development of RDFMD as a non-equilibrium method for the promotion of conformational changes in proteins. The main limitation of RDFMD is the significant parameterisation required to tailor the method to give a desired response. There are therefore two goals for RDFMD addressed in this chapter; first, users of RDFMD require a validated protocol suitable for similar work to that presented in this thesis, and second, methods are required to choose RDFMD parameters for an application outside the scope of this project. It is hoped that by presenting a detailed analysis of parameter choices and their interdependence, that RDFMD will become a widely used tool in computational investigations of conformational change. The methodology and results described in this chapter are in press[103](Wiley *et al.*).

## 5.2 The RDFMD Method

A digital filter is a list of coefficients, $c_i$ that can be used to weight a discrete vector input, $\mathbf{x}_i$, and summed to give a vector output, $\mathbf{y}$ (Equation 5.1). The filter response is the filter's effect on the phase and amplitude of the input signal. The greater the number of coefficients, the closer the filter's response will be to that desired.

$$\mathbf{y} = \sum_{i=-m}^{m} c_i \mathbf{x}_i \tag{5.1}$$

The RDFMD filter sequence begins by filling a buffer of velocities, $\mathbf{v}$, using a microcanonical (NVE) ensemble MD simulation. This buffer has the same number of steps as the number of filter coefficients ($2m$ +1) in Equation 5.2. The external rotational and translational motion of the target system, $\mathbf{v}_{Ext}$, is removed (Equation 5.3) and a digital filter is then applied to the cartesian components of the internal velocities, $\mathbf{v}_{Int}$. Filters used for this work are designed using the *fircls* function in MATLAB[104] and have the property of yielding a frequency response of unity over the frequency range to be amplified and zero elsewhere.[6]

$$\mathbf{v}_{Filt} = \sum_{i=-m}^{m} c_i \mathbf{v}_{Int,i} \tag{5.2}$$

$$\mathbf{v}_{Int} = \mathbf{v} - \mathbf{v}_{Ext} \tag{5.3}$$

The filters used are typically designed to extract all components of velocities corresponding to motions most likely to induce conformational change. These filtered velocities, $\mathbf{v}_{Filt}$, are multiplied by an amplification factor, $A$, that can be changed to adjust the energy put into the system. The filtered and amplified velocities are then summed with the original velocity set for the central buffer point, $\mathbf{v}_0$, producing a new set of velocities, $\mathbf{v}'$, that are in phase with the coordinates at the centre of the buffer (Equation 5.4). An amplification factor of 2 would therefore produce a final set of velocities for which the targeted frequency components have been extracted, multiplied by 2 and then summed with the original velocities to give a three-fold increase in the kinetic energy of targeted motions (as they will

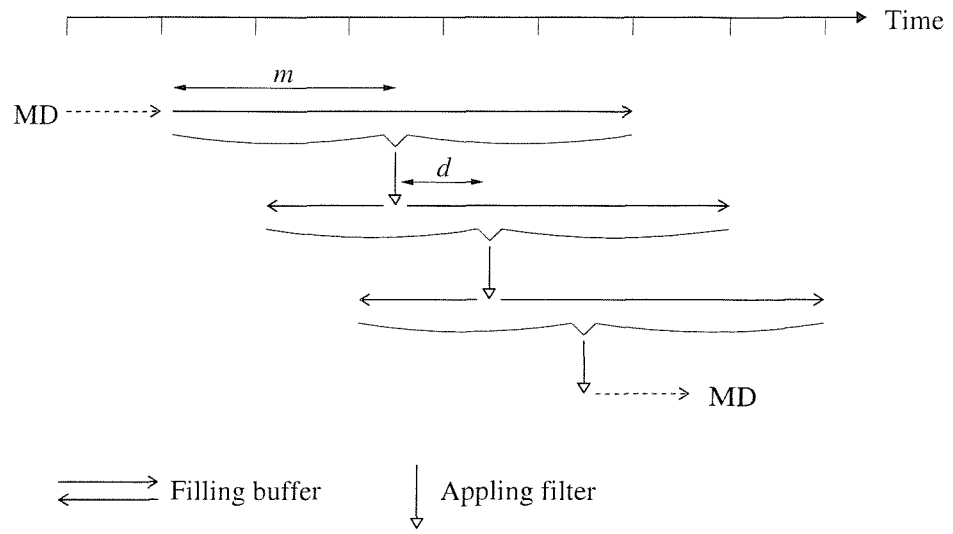



Figure 5.1: RDFMD sequence showing the delay parameter, $d$, and the filter length, $2m+1$.

also be included in the full set), with all other frequencies left unchanged.

$$\mathbf{v}' = \mathbf{v}_0 + A\mathbf{v}_{Filt} \tag{5.4}$$

From the new velocities, $\mathbf{v}'$, and the central coordinates of the buffer, conventional simulation continues both backward and forward in time so that a new buffer is filled. Another filter can then be applied, separated from the last by a specified time delay (the filter delay). The purpose of this delay is to allow the system to relax from the effects of the previous filter and to allow some progression over the potential energy surface. Filters are repeatedly applied in this manner until the kinetic energy in the system rises beyond some defined limit (the internal temperature cap) or until a certain number of filters have been applied (the filter cap).

RDFMD is typically run as the combination of periods during which the system velocities are modified by repeated applications of a digital filter, and of traditional MD in the canonical (NVT) or the isothermal-isobaric (NPT) ensemble (Figure 5.1). During MD for which a thermostat is applied, the temperature can be returned to that desired and new equilibrated velocities and coordinates generated, from which another set of filter applications can be performed.

## 5.3 Early work

Work in this section builds on that by Dr. Stephen Phillips, who introduced YPGDV as an RDFMD test case using the DLPROTEIN package.[105] The author located a problem with the parameters of the solvent model used, leading to an incorrect solvent density that will have significantly altered the conformational flexibility of the peptide. The author performed and analysed duplicate YPGDV simulations using the NAMD package,[100] with which the correct water density was produced. The results of these simulations have been published[6] (Phillips *et al.*). All simulations, figures and tables included in this chapter are the author's work.

### 5.3.1 Initial parameter set

The initial parameter set used for RDFMD on the YPGDV system was largely derived by Dr. Stephen Phillips. Parameters include a delay of 20 steps, corresponding to the half-life of energy dissipation after a filter application which amplified low frequency motions by a factor of 10. A temperature cap of 2000 K was arbitrarily chosen and a fixed amplification factor of 4 determined to be suitable by trials of several protocols. Sequences of up to 10 filter applications are separated by 10 000 steps of NPT MD simulation with a 2 fs timestep. Temperature and pressure were constrained using a Langevin thermostat damping parameter of 1 $ps^{-1}$ and a Nosé-Hoover Langevin piston barostat with a pressure target of 1 atm, a piston temperature of 300 K, a damping decay parameter of 200 fs and an oscillation period of 400 fs.

To determine a suitable frequency target, analysis of the NVE trajectory previously described in Chapter 4 was performed. Fourier transforms of the eight dihedral angle trajectories were integrated to give a cumulative amplitude of spectral components, indicating the significance of frequencies below a certain value to the variance of the vibrations occurring in the dihedral angles. These cumulative amplitudes have been normalised to lie between 0 and 1 and are shown in Figure 5.2. These confirm the presence of large-amplitude, low frequency

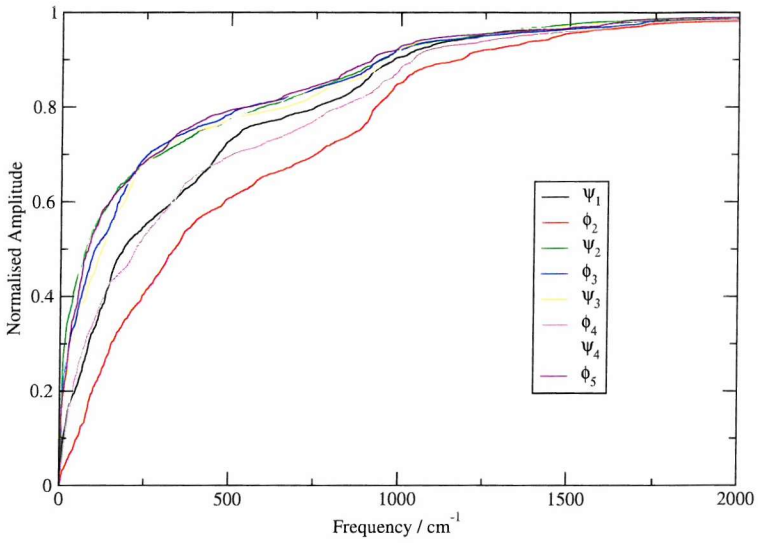



Figure 5.2: Cumulative sums of the amplitude spectra of the YPGDV backbone dihedrals. Amplitudes have been normalised to lie between 0 and 1.

motions.

To target low frequency motions, a frequency target of 0–25 cm$^{-1}$ was chosen and a filter designed using 1001 coefficients, using the MATLAB command:

*fircls*(1000,[0 0.002998 1],[1 0], [1.1 0.001],[0.9 -0.001],"text")

The first parameter gives the number of coefficients less one, in this case 1000. After this the normalised frequencies bounding the desired filter regions are given. To normalise the frequencies, the actual frequency is divided by the maximum frequency, $\nu_{MAX}$, that can be sampled given the timestep, $\delta t$, used: $\nu_{MAX} = \frac{1}{2\delta t c}$ (where $c$ is the speed of light, and the factor of 2 is included to satisfy the Nyquist criterion). The desired filter response is given in the second set of square brackets, indicating here that for the normalised frequency range 0–0.002998, a response of unity is required, and all higher frequencies should give a response of zero. The final parameters refer to accuracy limits required of the filter, and the nature of the output.

The desired filter response and the response generated by the 1001 coefficient filter are shown in Figure 5.3. The filter response is not equal to that desired, due to the limited number of coefficients. Only if the number of coefficients are infinite

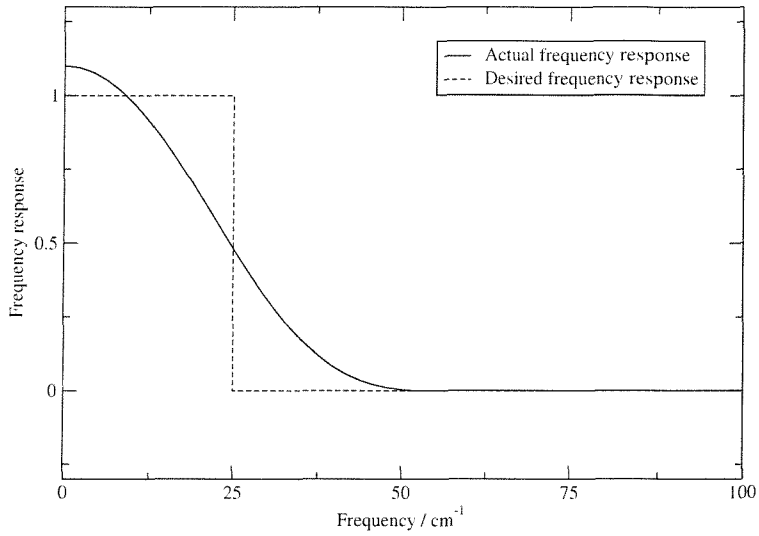



Figure 5.3: The desired and actual filter responses for a 0–25 $cm^{-1}$, 1001 coefficient filter.

will a truly exact response be generated. This filter is suitable for use however, as the frequency region for amplification need not be tightly bounded. When referring to the amplification of a 0–25 $cm^{-1}$, 1001 coefficient filter, it is important to recall that some amplification is occurring between 25–50 $cm^{-1}$.

The target region of 0–25 $cm^{-1}$ was chosen partly for efficiency reasons since the lower the frequency target, the larger the number of coefficients that are required to generate a frequency response of similar accuracy. Increasing the number of coefficients increases the computational expense of filling a velocity buffer and thus the region 0–25 $cm^{-1}$ could be considered a minimum that the method can efficiently target. The HHT results previously shown indicate energy in the 0–10 $cm^{-1}$ region, however to target this frequency region reliably requires filters of up to 3001 coefficients.

To explore the conformational space sampled by the initial protocol, six simulations were performed from the equilibrated YPGDV state (described in Chapter 4), using the equilibrated velocities and five sets of velocities that were randomised and scaled to give a temperature of 300 K. 10 000 steps of NPT MD simulation was performed before the first filter application. Four of the six simulations included ω *cis-trans* isomerisation across the four YPGDV ω

angles. ω angle motions shall be discussed further in Chapter 6, and at this stage isomerisation shall be assumed to be a negative event, suggesting exploration of regions of higher energy conformational space than desired.

The results of the two all-*trans* ω simulations are plotted in Figures 5.4 and 5.5. Each Figure should be considered as containing four quadrants- the minimum required to display eight data sets. Each quadrant therefore displays two dihedral angles and so one sampled conformer is represented by four points in the Figure. Conformers have been sampled every 200 steps (0.4 ps). Sampling of the ψ and ϕ angles in the two simulations shown is representative of the six performed.

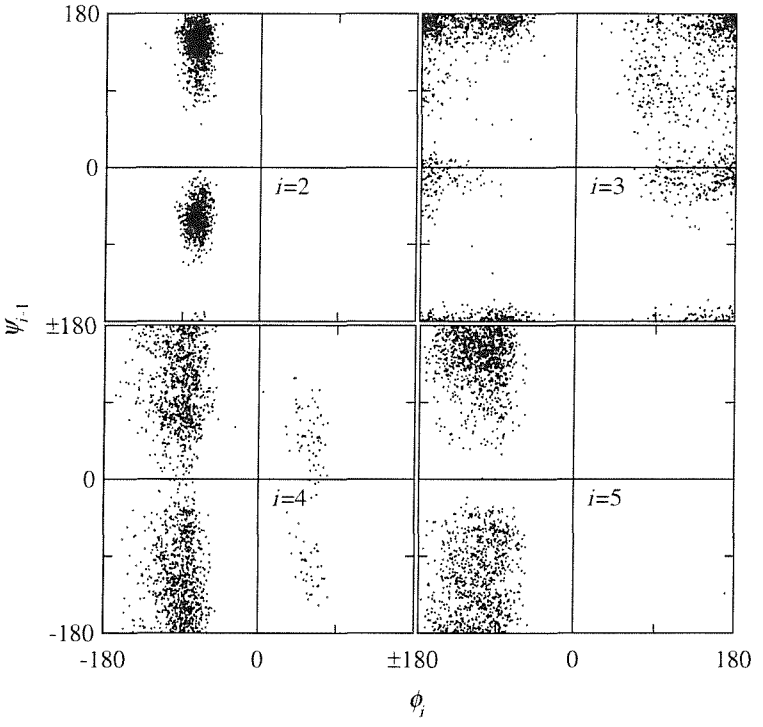



Figure 5.4: Backbone dihedral angle space sampled during 2 ns of RDFMD on YPGDV from randomised velocities. Parameters include a frequency target of 0–25 $cm^{-1}$, a 1001 coefficient filter, a filter delay of 20 steps and an internal temperature cap of 2000 K.

The two simulations required totals of 502502 and 505505 steps to apply the digital filters, adding approximately 50 % to the computational cost of the 1 000 000 steps (2 ns) of NPT. To show an example of the dihedral angle space that YPGDV would be expected to sample during MD, a 30 ns NPT MD simulation was performed and the results are shown in Figure 5.6. The sampling rate has been

adjusted so that the number of conformers displayed is similar to those shown in Figures 5.4 and 5.5. No $\omega$ transitions were expected or witnessed in this simulation.

The RDFMD simulations clearly sample similar phase space to that seen in the 30 ns of MD, and also see some conformers not otherwise sampled. One point of significance is that the two RDFMD simulations do not see all of the same conformers; in particular the first simulation sees a significant proportion of $\psi_1$ located in a conformation centred at approximately $\psi_1 = -60°$. Closer inspection of the first RDFMD simulation trajectory shows that only one transition is sampled in the $\psi_1$ angle, after which the simulation did not leave this conformation. RDFMD has therefore located a conformer that is separated from the starting state by an energy barrier that is significant at 300 K. There is no way of telling which, if either, should be the dominantly sampled state. This issue is addressed using parallel tempering results in Chapter 6. It is clear however, that the transition between these states is not adequately sampled by MD or the initial RDFMD protocol.

Another important issue is that the internal temperature cap of 2000 K is extremely high and cannot be realistically applied to simulations of larger proteins as these may be denatured by such high energies. Representative examples of results using the initial protocol and lower filter caps are shown in Figures 5.7 (with a 1250 K temperature cap) and 5.8 (with a 1500 K temperature cap). Conformational sampling is limited, and similar conformers are located to those seen with MD. A protocol is required that lowers the internal temperature needed to significantly enhance conformational sampling.

## 5.4 The systematic parameterisation of RDFMD

The initial protocol previously described requires a high internal temperature cap and cannot distinguish between the amplification of $\psi$ and $\phi$ angles with that of $\omega$ angles. An investigation into the parameterisation of RDFMD is required, to see whether refined parameters are capable of addressing the identified limitations. It is not necessary at this stage to apply RDFMD using significant portions of NPT simulation, and methods that are presented require little computation, analysing the

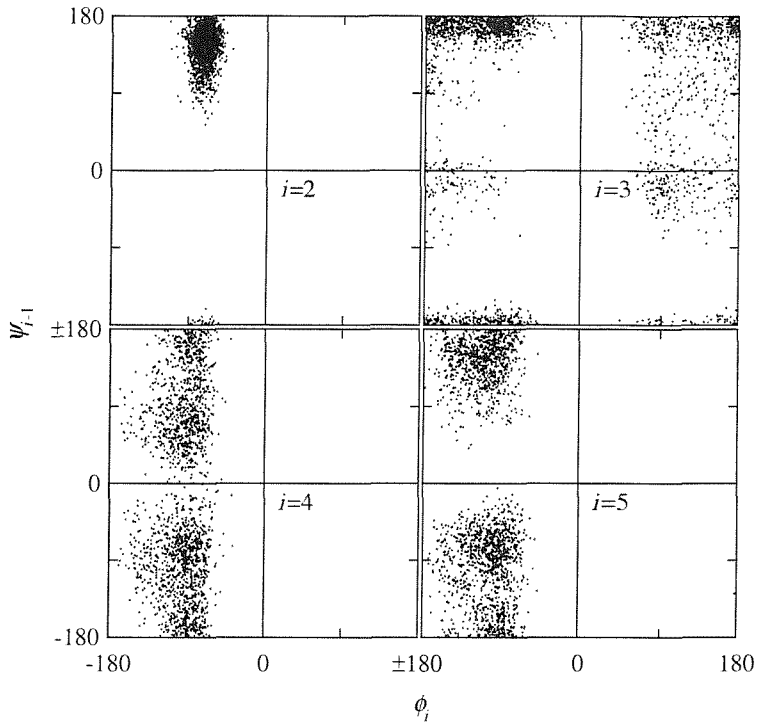



Figure 5.5: Backbone dihedral angle space sampled during 2 ns of RDFMD on YPGDV from randomised velocities. Parameters include a frequency target of 0–25 $cm^{-1}$, a 1001 coefficient filter, a filter delay of 20 steps and an internal temperature cap of 2000 K.

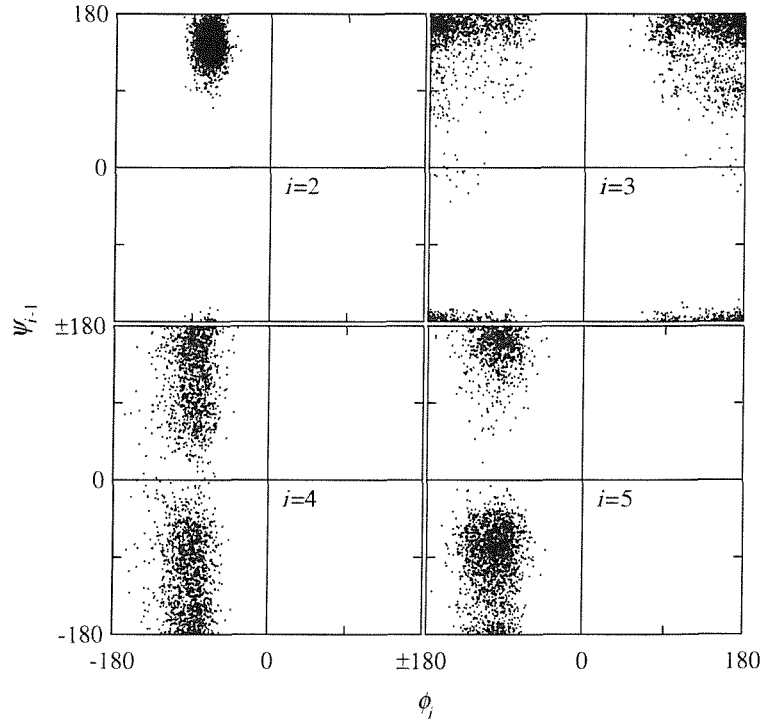



Figure 5.6: Backbone dihedral angle space sampled during a 30 ns NPT MD simulation of YPGDV.

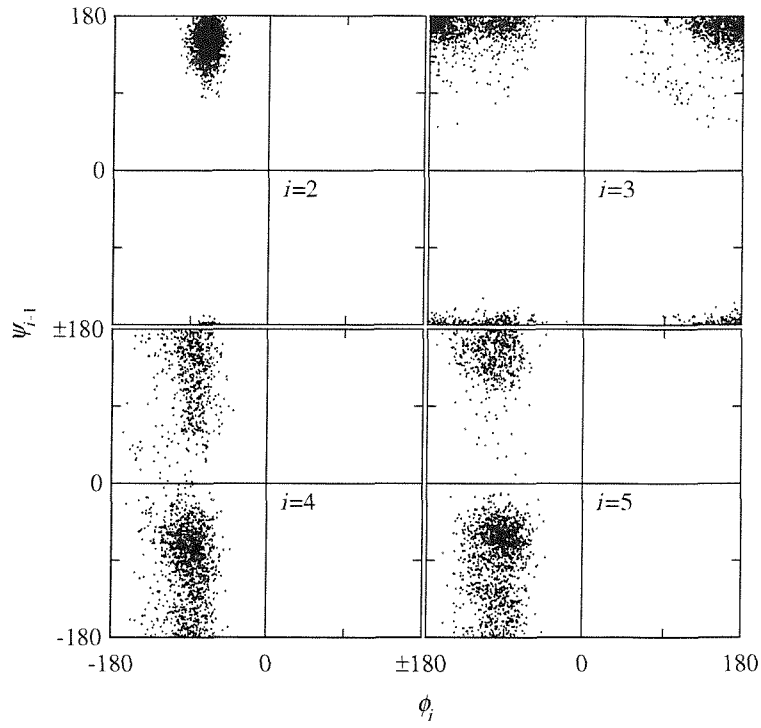



Figure 5.7: Backbone dihedral angle space sampled during 2 ns of RDFMD on YPGDV from randomised velocities. Parameters include a frequency target of 0–25 $cm^{-1}$, a 1001 coefficient filter, a filter delay of 20 steps and an internal temperature cap of 1250 K.

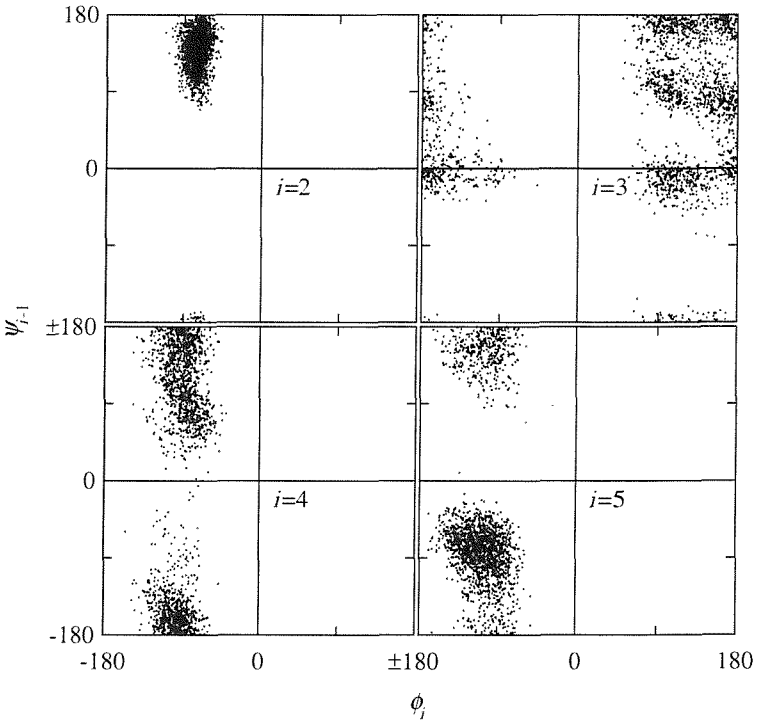



Figure 5.8: Backbone dihedral angle space sampled during 2 ns of RDFMD on YPGDV from randomised velocities. Parameters include a frequency target of 0–25 $cm^{-1}$, a 1001 coefficient filter, a filter delay of 20 steps and an internal temperature cap of 1500 K.

short NVE simulation previously discussed in Chapter 4 and the effects of RDFMD over the timescale of a velocity buffer. Parameters are therefore systematically varied from the initial protocol, and their interdependence discussed. The internal temperature cap parameter will be more fully discussed in Chapter 6. Unless otherwise stated, the RDFMD parameters are: a filter buffer length of 1001 steps, a target frequency of 0–25 $cm^{-1}$, an amplification factor of 4, a filter delay of 20 steps, an internal temperature cap of 2000 K and a filter cap of 10 filters. All atoms in the YPGDV peptide are targeted by the digital filter.

### 5.4.1 Frequency Target

The frequency target is the range of frequencies that are desired for amplification or suppression by RDFMD. This parameter is the most important and is required before all others can be optimised. It is suggested that some form of frequency analysis on MD simulations is performed to validate the frequency target, or alternatively the response of the system to different filters can be tested. If the range chosen is too broad, energy will be put into or removed from motions that are not intended for manipulation. If the range is too narrow however, the frequencies at which desirable motions occur may be missed.

A frequency target in the 0–25 $cm^{-1}$ regions was suggested on an amplitude argument previously discussed, and on the strength of HHT results showing low frequency motions accompanying conformational events.

To examine the frequency target further, the Empirical Mode Decomposition Method has been applied to the NVE 16 678 step simulation described in Chapter 4. Each backbone dihedral angle trajectory was split into three components: high frequency 'noise' from interactions with higher frequency degrees of freedom (angle and bond stretching motions), the dihedral angle motion that persists throughout the simulation, and the low frequency conformational motions resulting from intermittent large scale conformational changes in the system. Fortunately, the YPGDV system is inherently flexible and a conformational event was captured within the short NVE simulation. This was the brief formation of a β 3-turn

between 17.3 and 17.7 ps, as determined by the DSSP (Dictionary of Protein Secondary Structure[106]) algorithm using default options. This secondary structure change was accompanied, and followed, by significant rearrangement in $\psi_2$, $\phi_3$ and $\psi_3$, as shown in Figure 5.9.

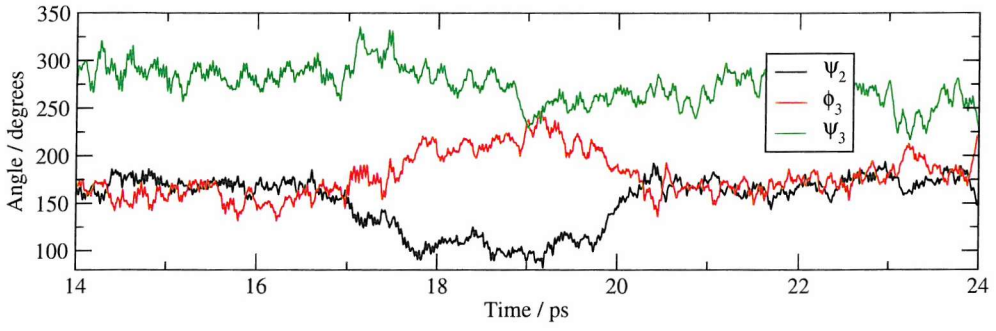



Figure 5.9: Relevant dihedral angles during the YPGDV conformational change event.

EMD performed on the $\psi_2$ signal produces 11 IMFs, the first two (those with the highest frequencies) describe motions over 200 $cm^{-1}$, as determined by Fourier transform (FT). The last five IMFs describe motions below 25 $cm^{-1}$. The physical relevance of the IMFs is best described when summed in frequency groups as shown in Figure 5.10. The Fourier transform of the summed signals are shown in Figure 5.11, showing the frequency ranges that incorporate the different signals. Although there is some frequency overlap between the signals, it cannot be determined whether this is due to the deficits of the EMD or the FT methods. Limitations of the application of EMD to MD analysis have been discussed in Chapter 4.

Results show that the vibrations below 25 $cm^{-1}$ are associated with the conformational change event itself. Therefore, the previous frequency target of 0–25 $cm^{-1}$ would only be able to amplify physically relevant signals (those of significant amplitude present in the system) if a large scale conformational event such as that shown is occurring. To target the dihedral angle motions present in the entire simulation, the frequency range 25–100 $cm^{-1}$ is clearly significant. Amplifying frequencies from 0 $cm^{-1}$ should be considered desirable in case motions are occurring at very low frequencies. It is clear that the inherent dihedral motions, i.e. vibrations within a potential energy well, occur predominately

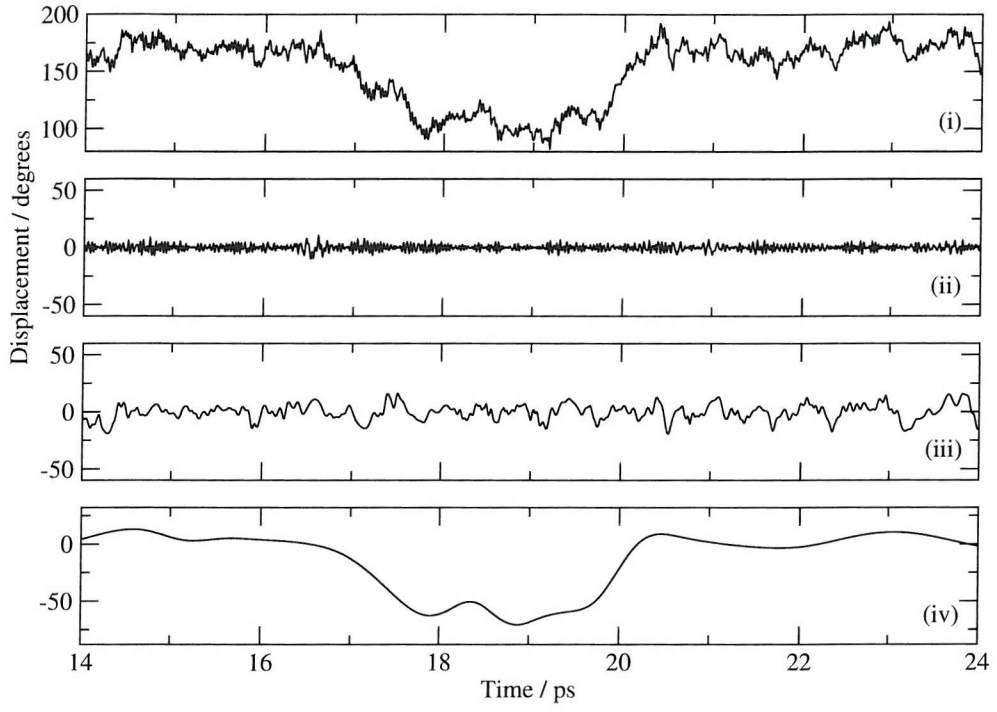



Figure 5.10: (i) $\psi_2$ signal during the conformational change event, (ii) Sum of IMFs 1 and 2 showing motions above 200 $cm^{-1}$, (iii) Sum of IMFs 3 to 6 showing motions predominantly between 25 and 250 $cm^{-1}$, (iv) Sum of IMFs 7 to 11 showing motions below 25 $cm^{-1}$.

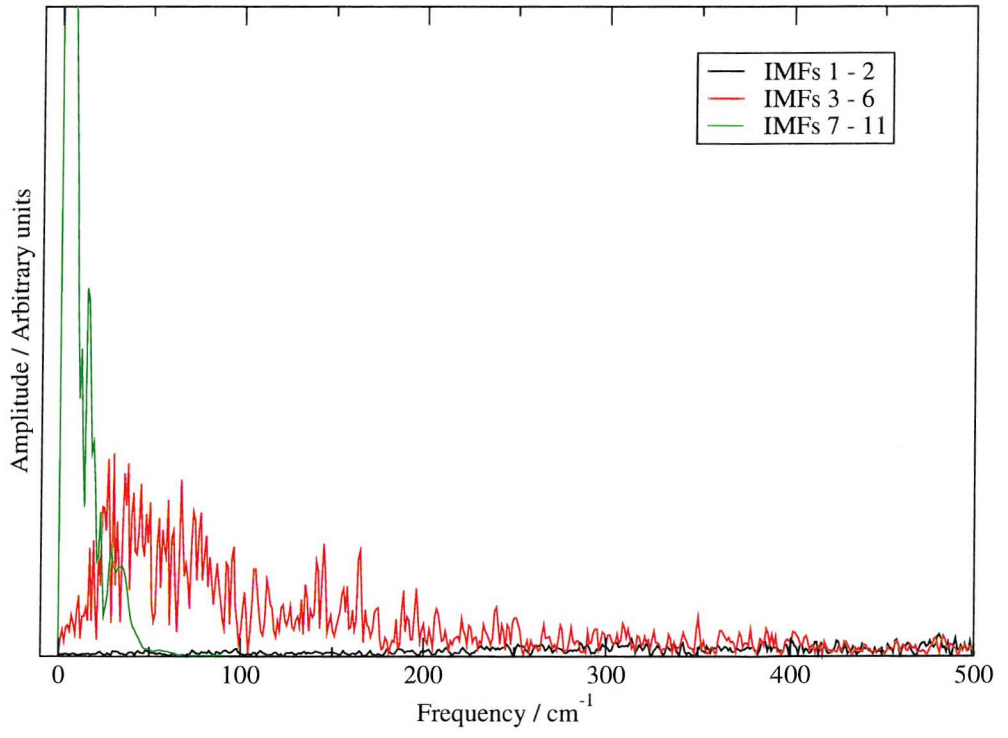



Figure 5.11: Fourier transforms of summed IMFs derived from the $\psi_2$ trajectory. A Hanning window was used.

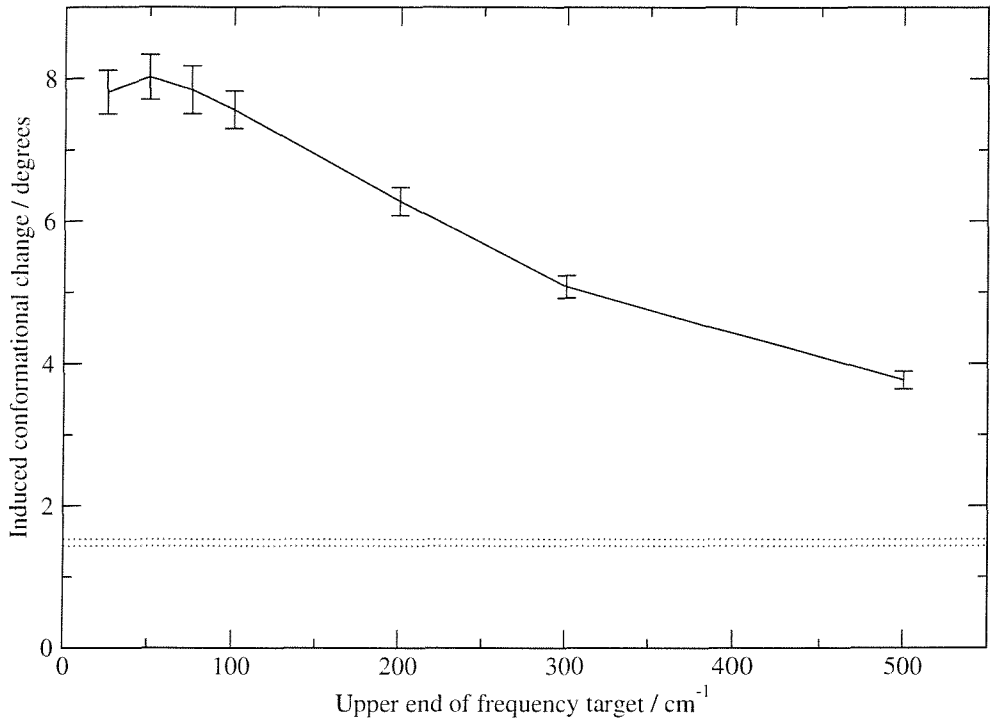
between 25 and 100 $cm^{-1}$, and that these frequencies should be targeted if large-scale conformational change involving escape from local minima is to be induced.

To measure the response of the system to different frequency targets, RDFMD simulations have been performed on YPGDV using filters designed to target frequencies below a specified cap. For each frequency target, fifty simulations of a single filter application to the equilibrated YPGDV system after a random reassignment of velocities at 300 K, and 1000 steps of NPT MD, were performed. Filters designed to target broader frequency ranges will see more degrees of freedom and thus the amount of energy put into the system cannot be controlled with a constant amplification factor as previously used (Equation 5.4). Instead the amplification factor is recalculated for each filter application so that the kinetic energy of the system is always increased by 25 kcal $mol^{-1}$. This is a comparable energy increase to that seen using the original parameter set.

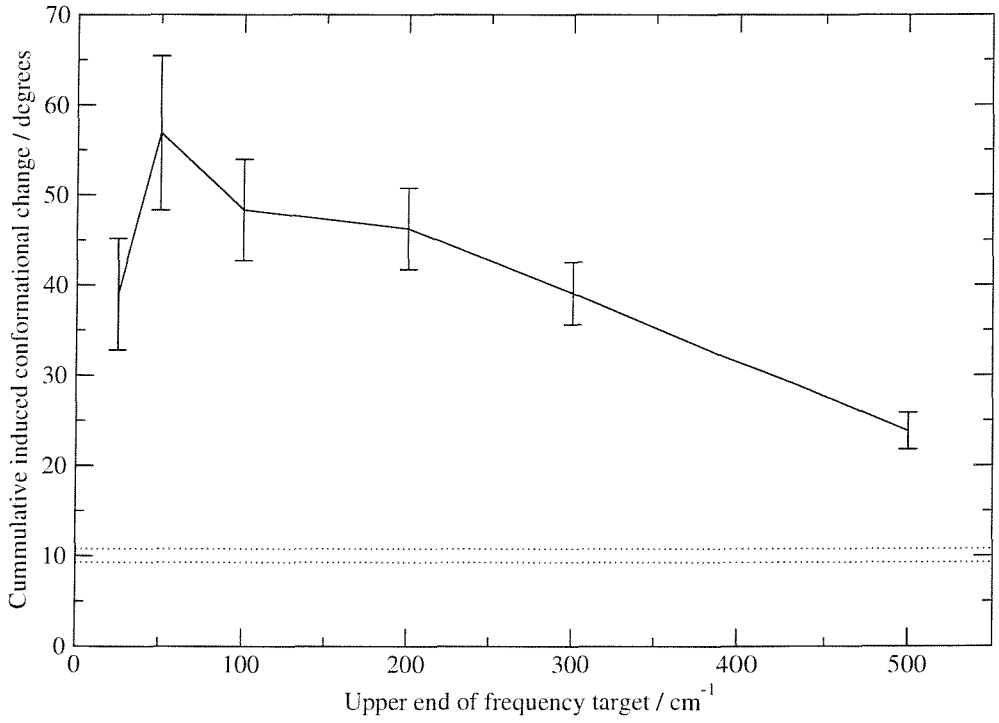
For analysis, the root mean squared deviation (RMSD) is calculated between the trajectories of the eight relevant dihedral angles before and after a filter application. This is done for 100 steps (200 fs) before and after each filter application so that only the effects of the filter application are measured. The sum of the RMSDs for each filter application is considered to be a measure of induced conformational change, and from the fifty simulations an average and standard error can be calculated. Averaged results are shown in Figure 5.12 (a), showing a drop off in the amount of induced conformational change after 100 $cm^{-1}$. For higher frequencies, less conformational change is induced and energy is being less efficiently placed in the dihedral angle motions. To demonstrate the importance of frequency specificity, a control simulation has been performed in which the filter used has a central coefficient of 1 and all other coefficients are zero. 25 kcal $mol^{-1}$ of kinetic energy is therefore put directly into the internal velocities, across all frequencies. The upper and lower error limits of the induced conformational change for the control simulation are shown in Figure 5.12 (a) by dashed lines. The amount of conformational change induced by this simple heating procedure is much less than that obtained using the targeted filters.

Repeated applications of filters have previously been shown to be more effective than use of a single filter and a greater amplification factor.[6] The different filter targets have therefore been tested for filter sequences, using a delay parameter of 20 steps and a filter cap of ten. Kinetic energy is increased by 25 kcal $mol^{-1}$ by each filter application, and the 10 equilibrated YPGDV states are used as simulation starting points. The induced conformational change is calculated as before for each filter buffer, and summed across all ten buffers for each run. The results are shown in Figure 5.12 (b). A non-frequency specific energy input has again been applied and the error limits are shown with dashed lines. As the upper limit on the frequency target is increased, the induced conformational change is reduced. The non-frequency specific energy input is once again far less efficient than any of the tested frequency targets.

There is a noticeable reduction in the conformational change induced by the 0–25 $cm^{-1}$ filter when repeated filter applications are used. It is believed that this is due to amplifying a frequency range that includes only low amplitude motions (such as 0–25 $cm^{-1}$ when a rare event is not occurring), rather than a range in which significant motions are apparent (such as 25–100 $cm^{-1}$ where dihedral motions persist throughout the simulation). This effect is best investigated by analysing an individual dihedral trajectory. Figure 5.13 (a) shows a breakdown of the $\psi_2$ dihedral angle from the first buffer of one of the simulations. Once again, the signal is separated by EMD into IMFs that can be grouped together as a high frequency component ($> 250$ $cm^{-1}$), a dominant motion of intermediate frequencies (30–150 $cm^{-1}$) and a low frequency motion ($< 60$ $cm^{-1}$). The Fourier spectra of the grouped IMFs are reported in Figure 5.13 (b). The intermediate frequency motion has significant energy around 70 $cm^{-1}$; outside the previously proposed frequency target for YPGDV. Figure 5.14 shows the effect of targeting either the low (0–25 $cm^{-1}$) or the low and intermediate (0–100 $cm^{-1}$) frequency regions identified here. Significant conformational motions are induced in both cases. However, amplifying the intermediate frequencies progressively targets the highest amplitude motion, yielding greater conformational change. In this analysis




(a) 50 applications of a single filter




(b) 10 filter sequences of 10 filters

Figure 5.12: Measure of induced conformational change with different frequency targets. Dashed lines indicate error bounds of conformational change induced by an input of non-frequency specific energy.

the frequency resolution is limited by the length of the buffer and a reduced resolution can be expected compared with the 16 678 step NVE simulation analysed previously.

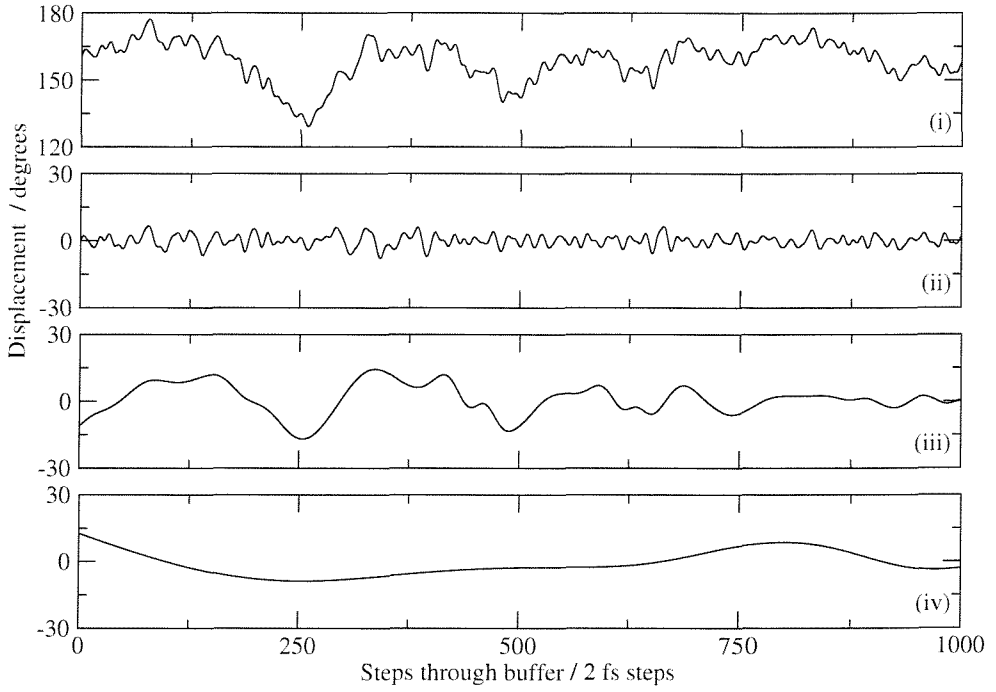
The filter target can therefore be selected by either measuring the system's response to a range of filters, or by predicting the desired frequency range from analysis of a sample trajectory. As has been shown, it is important to target frequencies relevant to the system, and the YPGDV results suggest no benefit in targeting above 100 $cm^{-1}$. The 0–100 $cm^{-1}$ range is suggested as optimal from both measurements of the system's response, and from analysis of frequencies related to inherent dihedral motions.

### 5.4.2 Buffer length

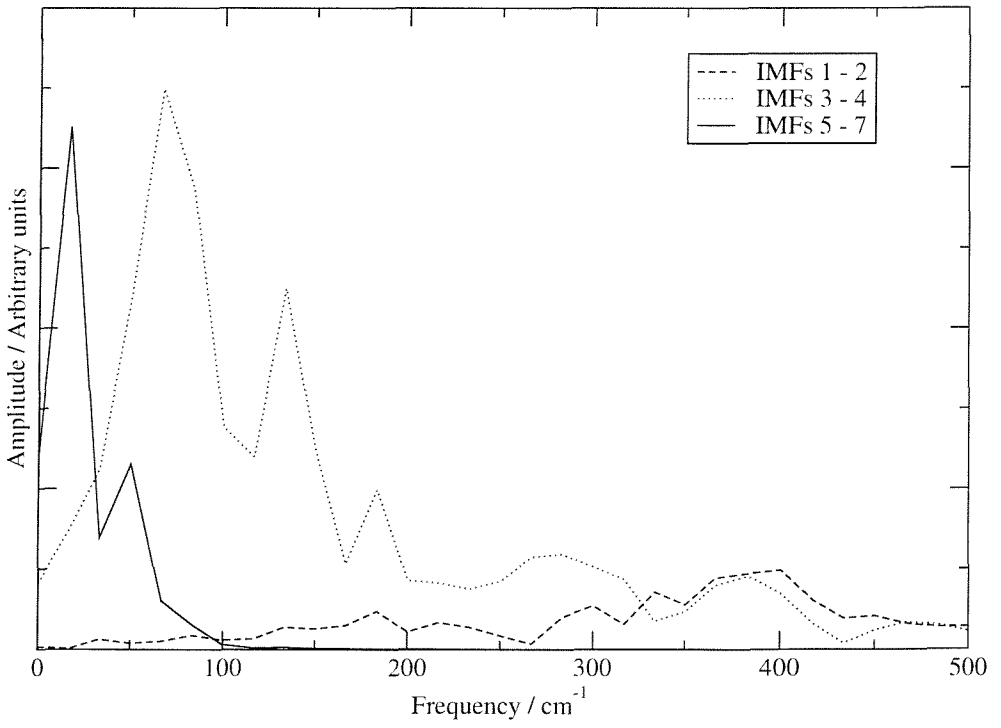
To apply a digital filter to a buffer of velocities, the buffer must contain at least the number of steps as there are coefficients in the filter. The larger the number of coefficients, the closer the frequency response of the filter will be to that for which it is designed. A shorter buffer requires reduced computational expense, but a filter of insufficient length will not produce a precise response and may result in undesired amplification or suppression of motions.

The previous filter target of 0–25 $cm^{-1}$ requires 1001 coefficients to achieve a reasonable filter response.[6] A higher cap on the frequency response suggested by the analysis presented here does not require as many coefficients to produce a sufficiently precise response. The responses of a number of 0–100 $cm^{-1}$ filters using different numbers of coefficients are shown in Figure 5.15.

RDFMD simulations using 0–100 $cm^{-1}$ filters with different numbers of coefficients have been performed. 50 starting points for single filter applications were produced from the equilibrated YPGDV system with randomised velocities and 1000 steps (2 ps) of NPT MD. The amplification of each filter is adjusted to increase the system's kinetic energy by 25 kcal $mol^{-1}$. The induced conformational change is calculated as before and the results are shown in Figure 5.16.

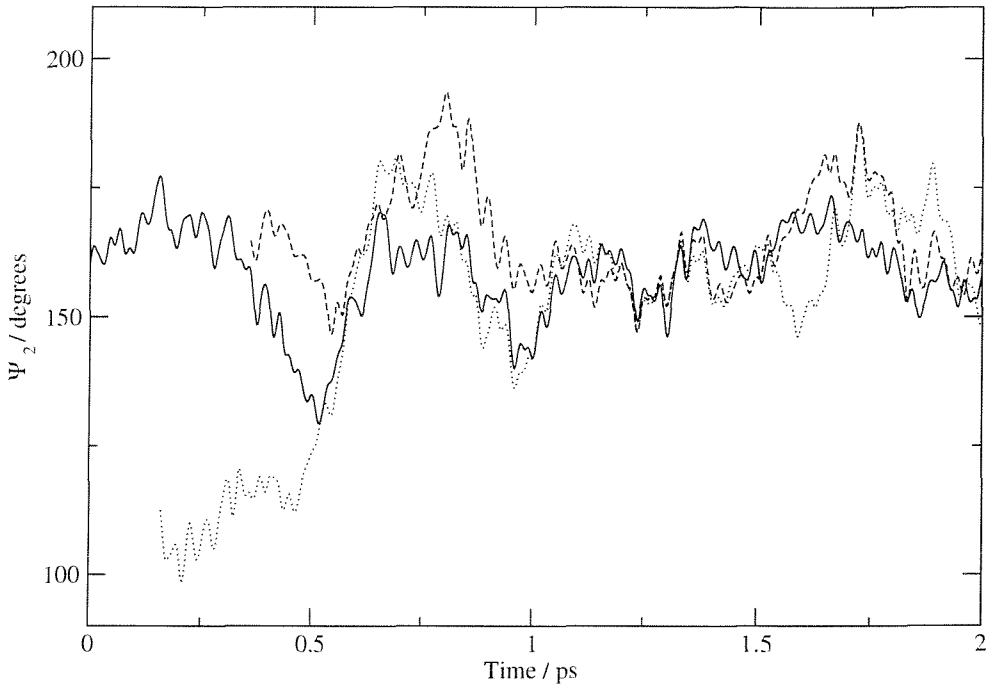



(a) (i) Signal, (ii) Sum of IMFs 1 and 2, (iii) Sum of IMFs 3 and 4, (iv) Sum of IMFs 5 to 7.

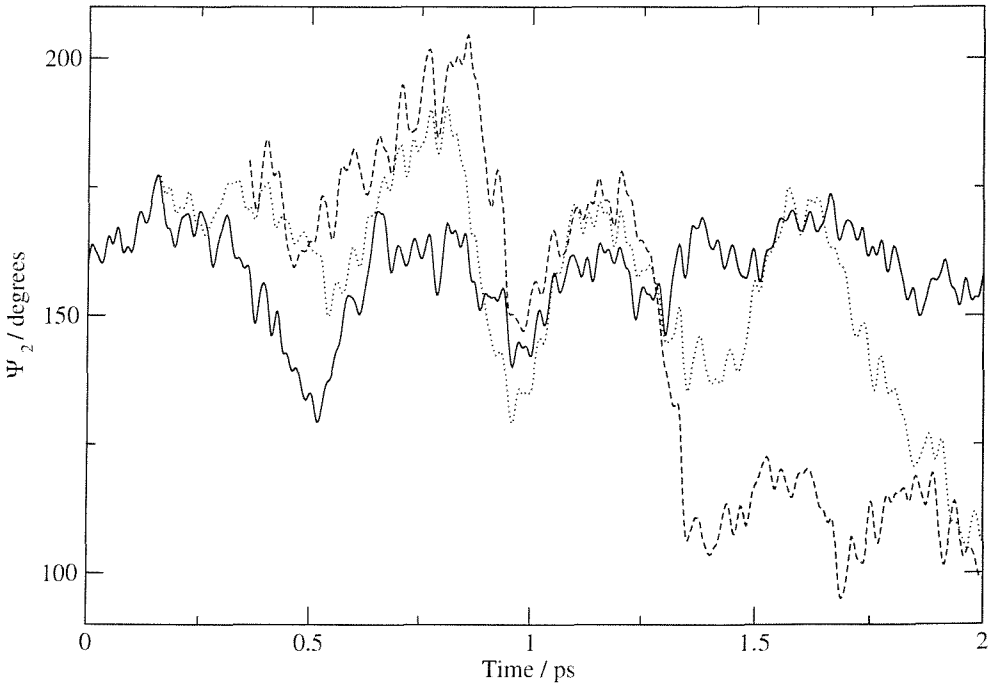



(b) Fourier Transforms of split signal using a Hanning window.

Figure 5.13: Decomposition of the $\psi_2$ trajectory from the first buffer of an RDFMD simulation.




(a) 0–25 $cm^{-1}$ filter applied.




(b) 0–100 $cm^{-1}$ filter applied.

Figure 5.14: $\psi_2$ trajectory after zero (solid), five (dotted) and ten (dashed) filter applications.

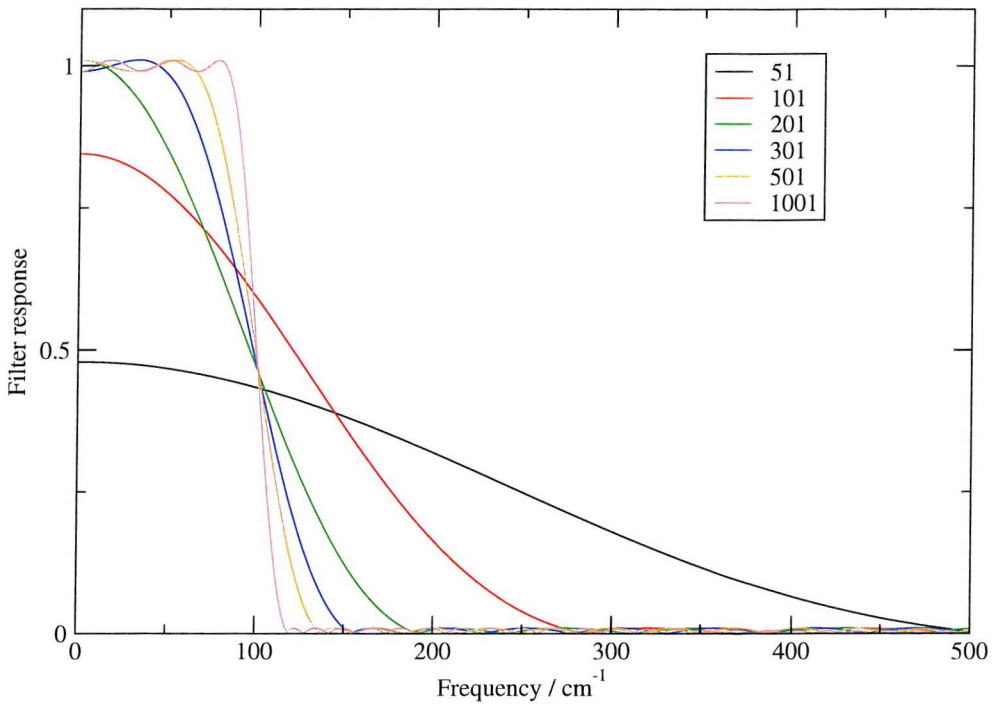



Figure 5.15: Filter responses of 0–100 $cm^{-1}$ filters using a different number of coefficients (shown in legend).

No benefit of using a greater number of coefficients than that required to produce an accurate filter response curve (200 to 300 coefficients) is seen. Fewer coefficients than this amplify higher than desired frequencies, for example, 100 coefficients targets 0–200 $cm^{-1}$ and induces a similar level of conformational change to a filter targeting this region with 1001 coefficients in Figure 5.12 (a).

### 5.4.3 Filter delay

A delay between filters allows energy put into the system by the previous filter to dissipate and thus prevents overheating, keeping the system away from the internal temperature cap. Ideally the delay should be as long as possible so that the energy build up is slow and the simulation advances over the potential energy surface. However the effects of each filter application can quickly dissipate and too long a filter delay will result in a series of essentially independent filter applications.

RDFMD simulations starting from each of the ten equilibrated YPGDV states were run using different filter delays. Other parameters are as specified

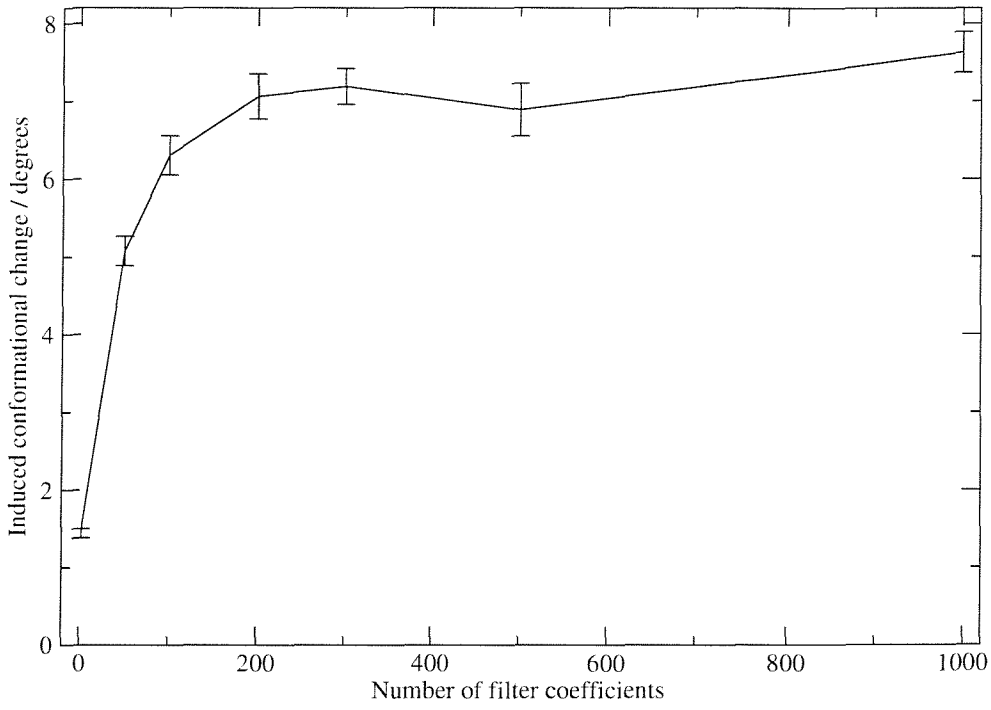
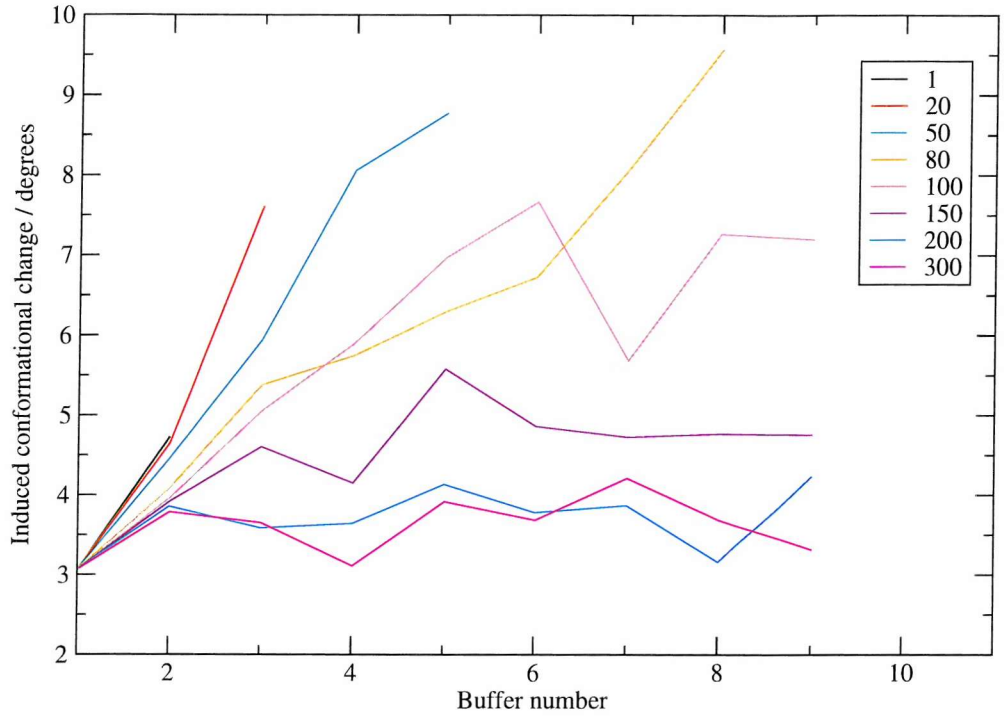


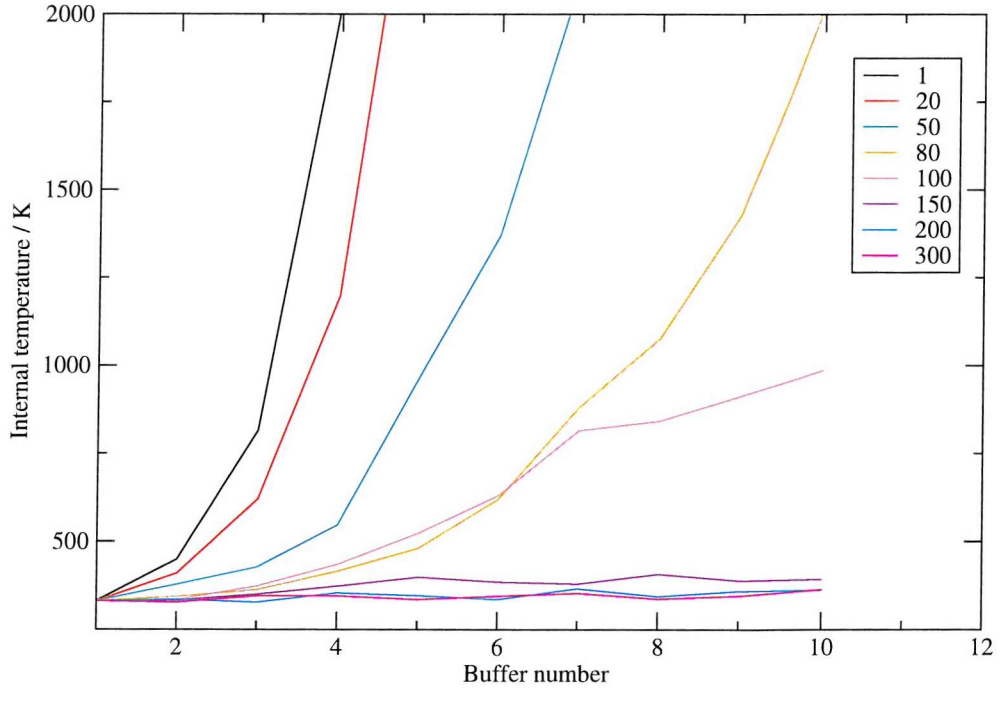

Figure 5.16: Results averaged from 50 RDFMD simulations using filters with different numbers of coefficients.

in the original protocol, including an amplification factor of 4 and a 0–25 $cm^{-1}$ filter. The induced conformational change produced by each filter application is shown in Figure 5.17 (a). The shorter the delay, the higher the initial induced conformational change, and the quicker the temperature cap is reached. Longer delays ($>$ 40 steps) do not quickly reach the temperature cap and more buffers are completed, each yielding progressive amplification of targeted motions. Very long delays ($>$ 150 steps) yield constantly low conformational changes as energy is dissipating between buffers and the filter applications become more independent and less progressive. The energy build up is more clearly seen in Figure 5.17 (b) in which the internal protein temperature after a filter application is plotted against the buffer number. Short delays quickly reach the 2000 K temperature cap and long delays see no progressive increase in internal temperature.

Using the Hilbert-Huang Transform, it is possible to see the energy build up in dihedral angle signals (the signal energy) for the targeted frequency range. Figure 5.18 shows the results for one of the RDFMD simulation starting points showing energy in motions occurring in the 0–50 $cm^{-1}$ region. A running average




(a) Conformational change induced in filter buffer.




(b) Internal protein temperature after filter application.

Figure 5.17: Effects of changing the filter delay parameter. Results averaged over 10 simulations.

over fifty steps has been performed on the data and the first and last 100 steps of each buffer removed as frequency information at the edges of HHT data can be unreliable.[93] A delay of 1 step shows the greatest increase in energy for each amplification performed, but only four buffers are completed before reaching the temperature cap. A delay of 50 steps completes seven buffers as the system is able to relax between filter applications. The total signal energy reaches a higher level than the simulation with a 1 step delay before reaching the temperature cap; an improvement that shows the importance of the filter delay. A long delay of 300 steps shows little amplification and each buffer does not necessarily reach higher levels of low frequency energy than the last, suggesting a non-progressive protocol.

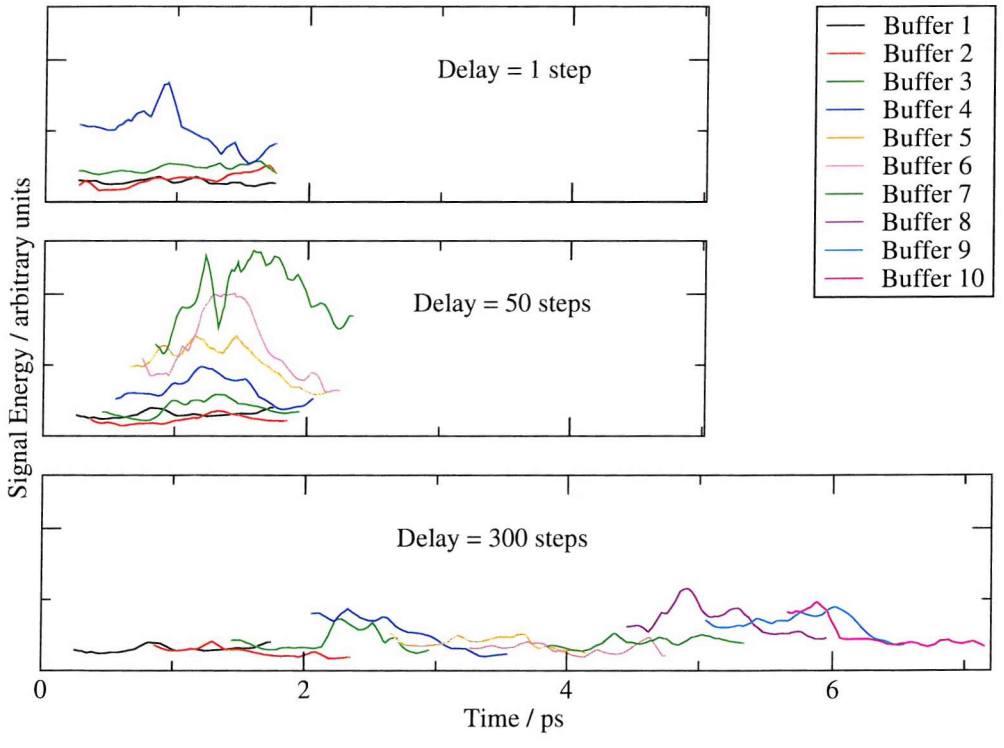



Figure 5.18: HHT of dihedral angle trajectories for different filter delay parameters. Buffers completed before reaching the internal temperature cap shown. y-axes are shown to the same scale for comparison.

For the original RDFMD parameter set, a filter delay between 50 and 100 steps is clearly most suitable, showing progressive amplification without overheating the system.

### 5.4.4 Amplification factor

A specified input of kinetic energy (as used to analyse the frequency target) or increase of temperature can be used to adjust the amplification of the velocities in RDFMD. A fixed amplification factor was originally implemented that adjusts the level of kinetic energy put into the system according to the amount that was already there. If there is a small amplitude, low frequency motion present in the system, the filter will increase the kinetic energy by less than if the motion is of larger amplitude.

Regardless of the method of energy insertion, sufficient energy must be put into the system so the effect of one filter application has not dissipated before the next. Equally, too great an amplification of velocities could overheat the system; a protocol that would risk denaturing larger protein systems.

A range of RDFMD simulations using different amplification factors from each of the ten equilibrated YPGDV states have been performed. The induced conformational change is measured as previously described. The amount of kinetic energy in the low frequency range is calculated from the velocities extracted by the applied 0–25 $cm^{-1}$ filter ($\mathbf{v}_{Filt}$ in Equation 5.2). A linear correlation is found between low frequency energy and induced conformational change, even when results are averaged over several simulations (typically with a correlation coefficient $> 0.8$). Lines of best fit are plotted in Figure 5.19 (a) to the edges of the data set. The amount of conformational change induced when a specific quantity of energy is measured in the filter region increases as the amplification factor is raised. An amplification factor of 1 is inefficient, and later buffers do not explore regions of significantly increased low frequency energy. RDFMD is stopped by reaching the maximum filter cap of ten. An amplification factor of 2 yields greatly increased energy in the filter region, and is clearly the most suitable compliment to the rest of the parameter set. At, and above, an amplification factor of 3, the internal temperatures reach the cap of 2000 K, and simulations are stopped. Although significant conformational motion is induced by high amplification factors, there are minimal increases in low frequency energy over a small number of buffers

before the temperature cap is reached. The protocol is therefore neither gentle nor progressive, as is desired.

The energy input required for gentle but progressive low frequency energy amplification is heavily dependent on the rest of the parameter set. For example if the delay parameter is increased to 50 steps (a suitable choice as seen from previous analysis), an amplification factor of 2 shows slow energy build up and factors of 3 to 4 are more suitable, as shown in Figure 5.19 (b). The improvement using the longer delay is clear, with 50 % more conformational motion induced when comparing the best result to that of the simulations using a 20 step filter delay.

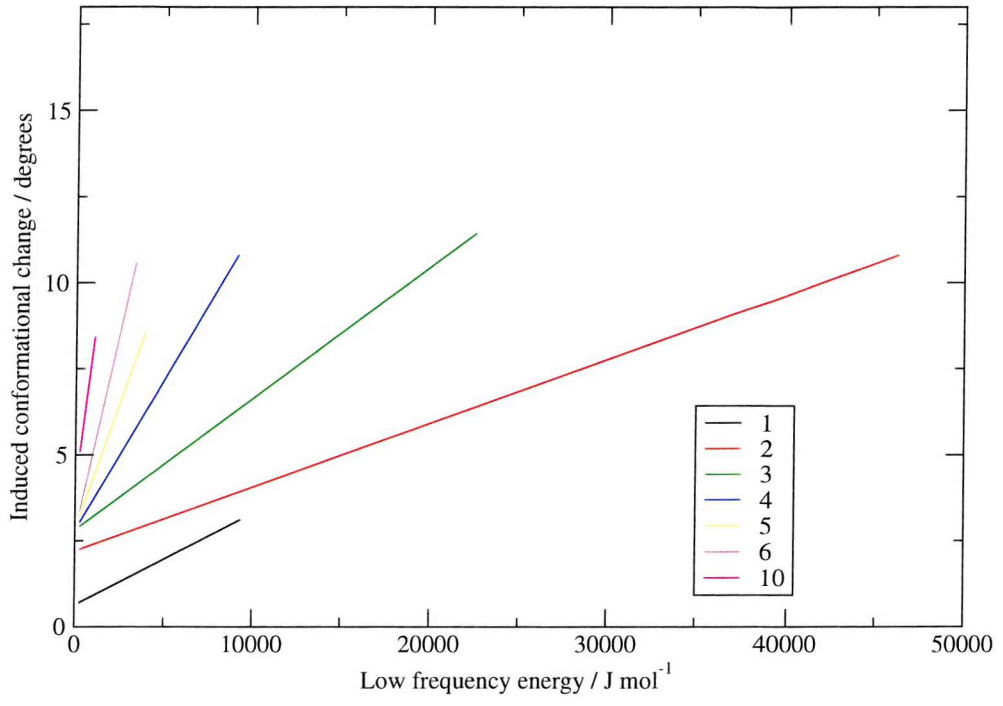
Once a frequency target and filter delay have been chosen for application to a system, analysis similar to that shown here will assist in tailoring the method of amplification.

### 5.4.5 Interdependence of parameters

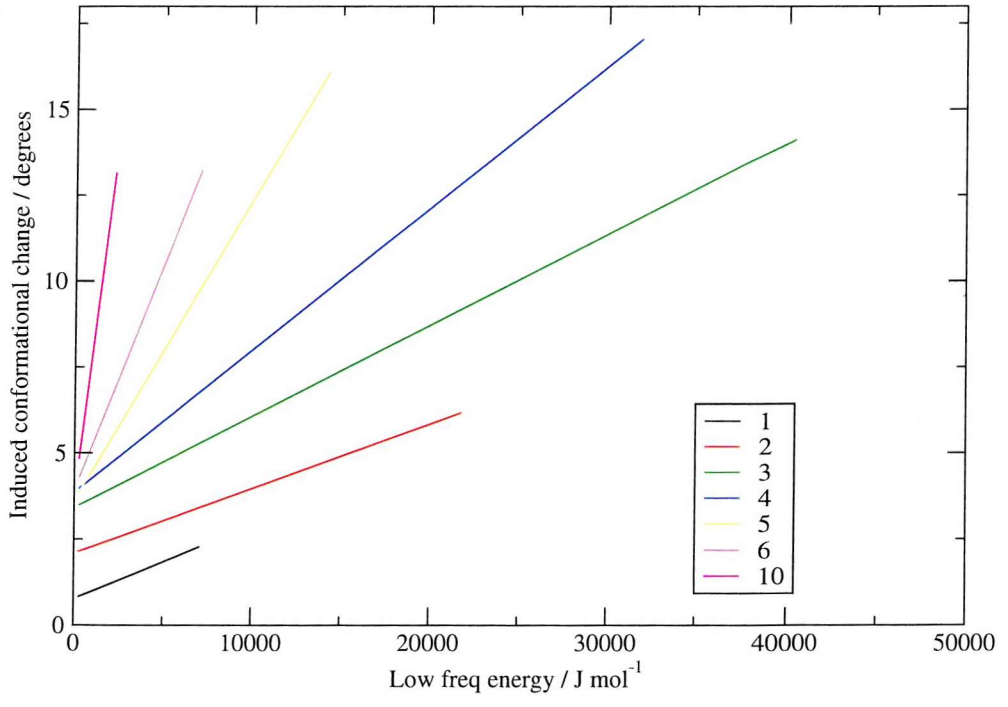
As previously discussed, the parameters used for RDFMD are heavily interdependent. Once a frequency target has been chosen however, optimising the other parameters discussed here can be done systematically with a small number of trial simulations.

Results presented so far suggest use of a 0–100 $cm^{-1}$ filter, a delay between 50 to 100 steps (0.1 to 0.2 ps) and amplification factors less of than 4. An accurate frequency response for a 0–100 $cm^{-1}$ digital filter is produced using 301 coefficients (Figure 5.15). Averaged results for delay parameters and amplification factors in the suggested regions are presented in Figures 5.20 (a) and (b) as previously described. An amplification factor of 2 is shown to be the best choice for a filter delay of 50 or 100 steps. It is worth noting that for both parameter sets the levels of induced conformational change and of low frequency energy are much higher than with previous protocols.

The internal temperature cap has not been discussed during the optimisation of the RDFMD protocol. A duplication of the simulations reported in Figure 5.20

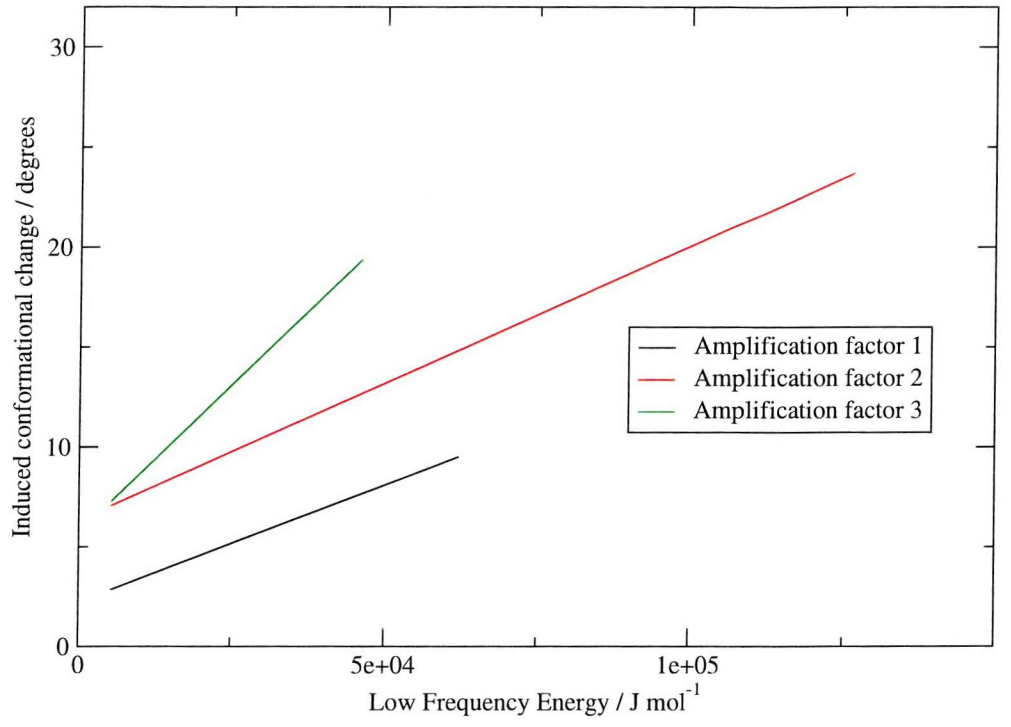



(a) Filter delay of 20 steps.

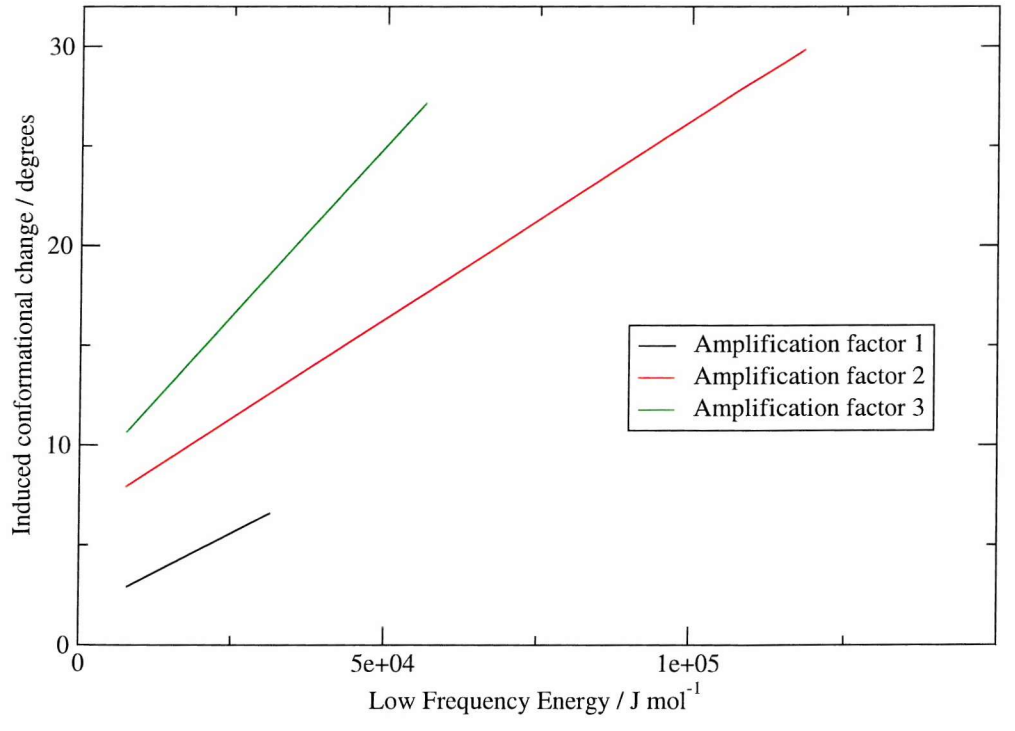



(b) Filter delay of 50 steps.

Figure 5.19: Effects of differing amplification factor (shown in legend).




(a) Filter delay of 50 steps.




(b) Filter delay of 100 steps.

Figure 5.20: Results for a range of filter delay parameters and amplification factors. Data are averaged over 10 simulations and presented as previously described.

have therefore been produced for an internal temperature cap of 1000 K, and the results are shown in Figure 5.21. The conclusions are unaltered; an amplification factor of 2 gives the best increases in induced conformational change and low frequency energy for a filter delay of 50 or 100 steps. It is presumed therefore that the derived protocol is the most suitable regardless of the internal temperature cap.

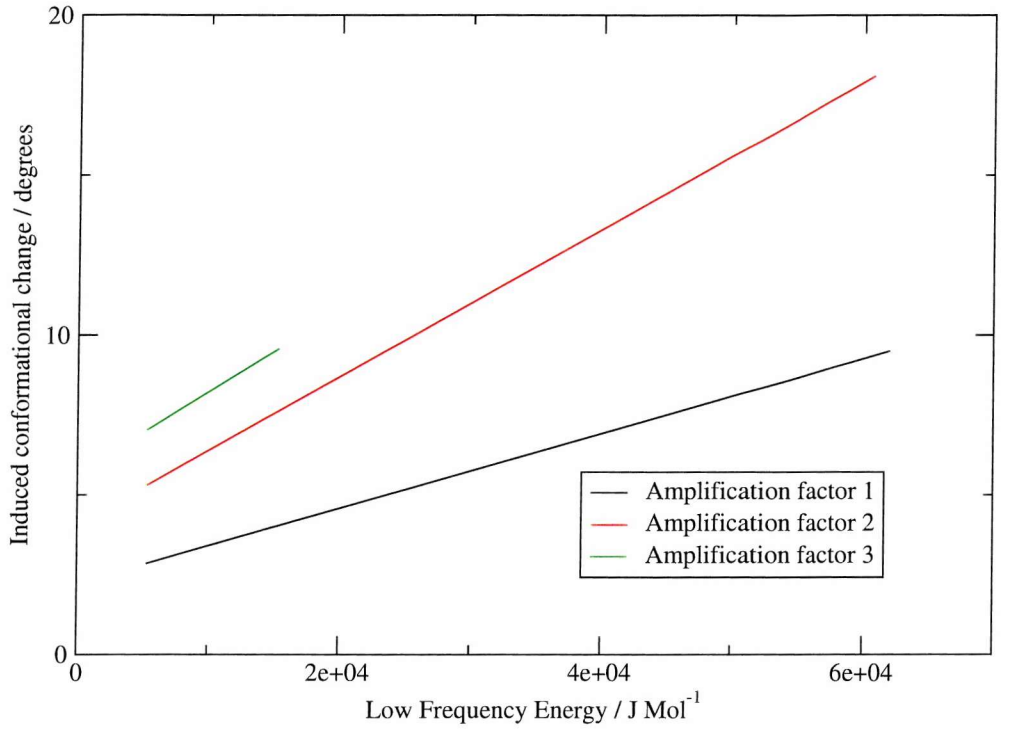
## 5.5 Applications of derived protocols to YPGDV

The parameter sets suggested by the analysis presented here have been applied to the equilibrated YPGDV system, interspersed with NPT MD simulation as previously described. The 10 000 steps (20 ps) of NPT simulation used between filter sequences has been reduced to 2 000 steps (4 ps) as this was found to be sufficient to re-equilibrate the temperature to the desired 300 K. 100 filter sequences are applied, totalling 400 ps of NPT simulation. All simulations have been duplicated, using randomly assigned velocities to replace the equilibrated set. The new simulations require a total of approximately 150 000 steps; 10 % of that used by the initial protocol.

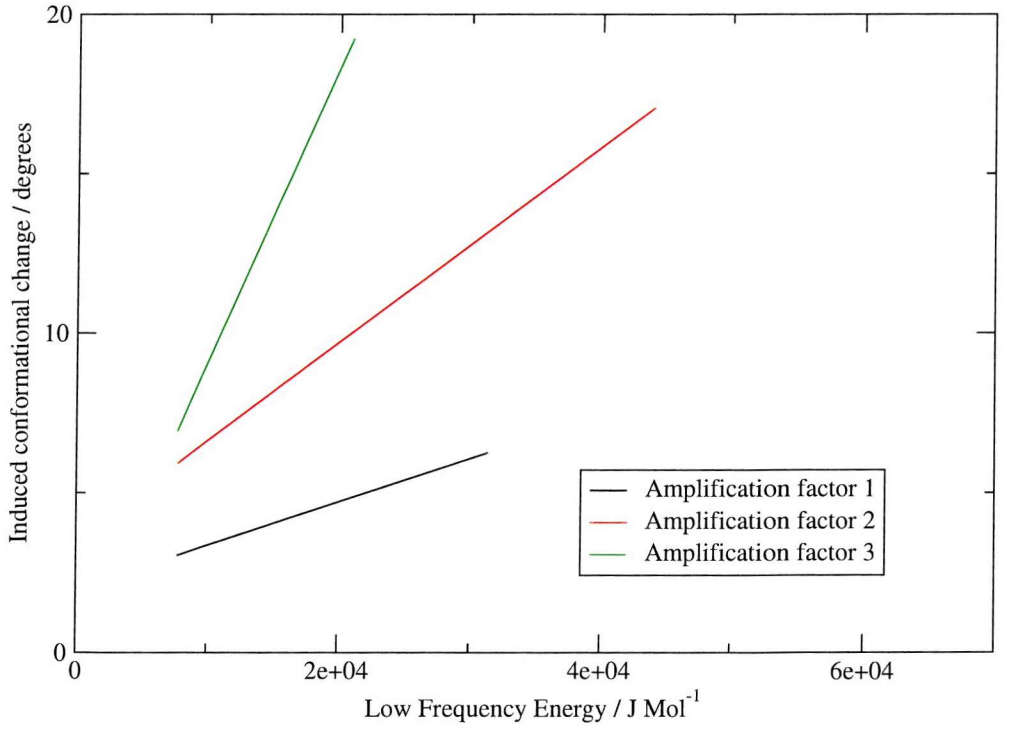
The protocols suggested for maximum efficiency of inducing dihedral angle motions are: a frequency target of 0–100 $cm^{-1}$, a 201 coefficient filter, an amplification factor of 2, and a filter delay of between 50 and 100 steps. Figures 5.22 and 5.23 show representative results from 400 ps of NPT simulation applied between filters using delays of 50 and 100 steps respectively, an internal temperature cap of 2000 K and a maximum of ten filter applications in each sequence.

Similar sampling is seen for the two filter delay parameters, as suggested by the levels of conformational change the parameters are shown to produce. Sampling is greatly increased to that produced by 30 ns of NPT simulation (shown in Figure 5.6). All simulations performed using the new protocol and an internal temperature cap of 2000 K report all protein $\omega$ angles 'flipping' from *trans* to *cis* states.

To explain the significant (and undesired) amplification of $\omega$ angle motions, cumulative amplitude plots of $\psi$ and $\phi$ dihedral angles and the $\omega$ angles can be




(a) Filter delay of 50 steps.




(b) Filter delay of 100 steps.

Figure 5.21: Results for a range of filter delay parameters and amplification factors. Data are averaged over 10 simulations and presented as previously described. An internal temperature cap of 1000 K has been used.

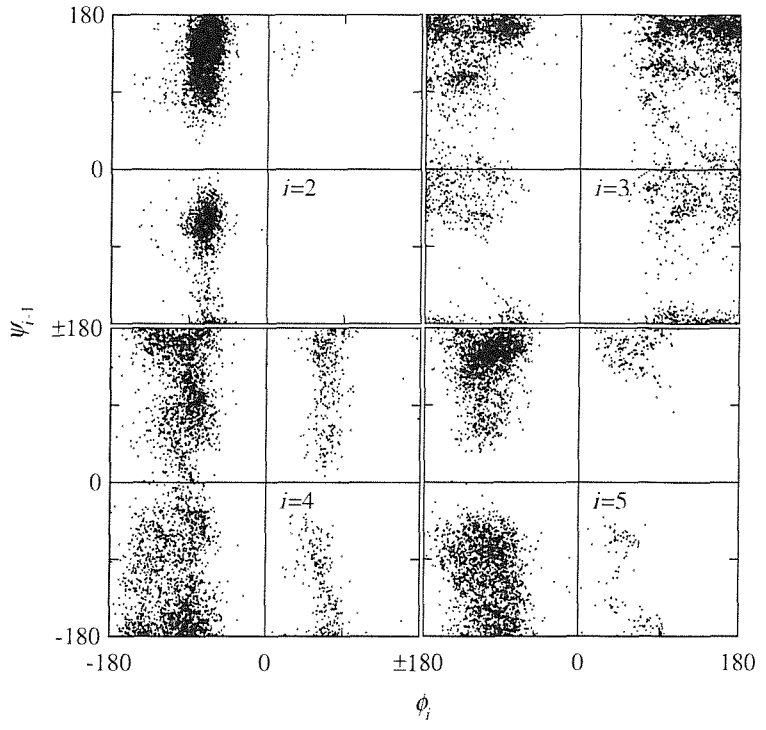



Figure 5.22: Backbone dihedral angle space sampled during 400 ps of RDFMD on YPGDV using a 0–100 $cm^{-1}$, 201 coefficient filter, a filter delay of 50 steps, an amplification factor of 2 and an internal temperature cap of 2000 K.

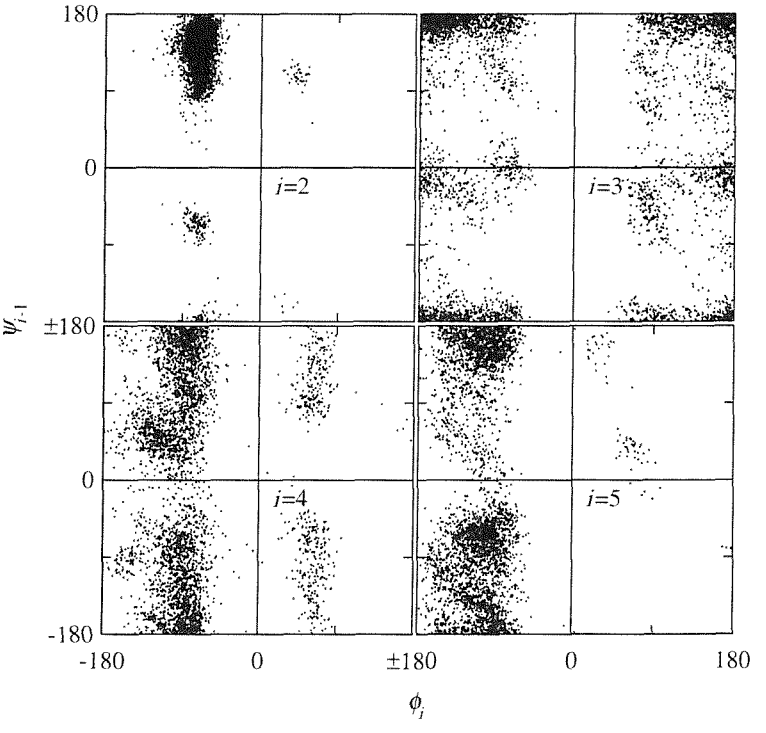



Figure 5.23: Backbone dihedral angle space sampled during 400 ps of RDFMD on YPGDV using a 0–100 $cm^{-1}$, 201 coefficient filter, a filter delay of 100 steps, an amplification factor of 2 and an internal temperature cap of 2000 K.

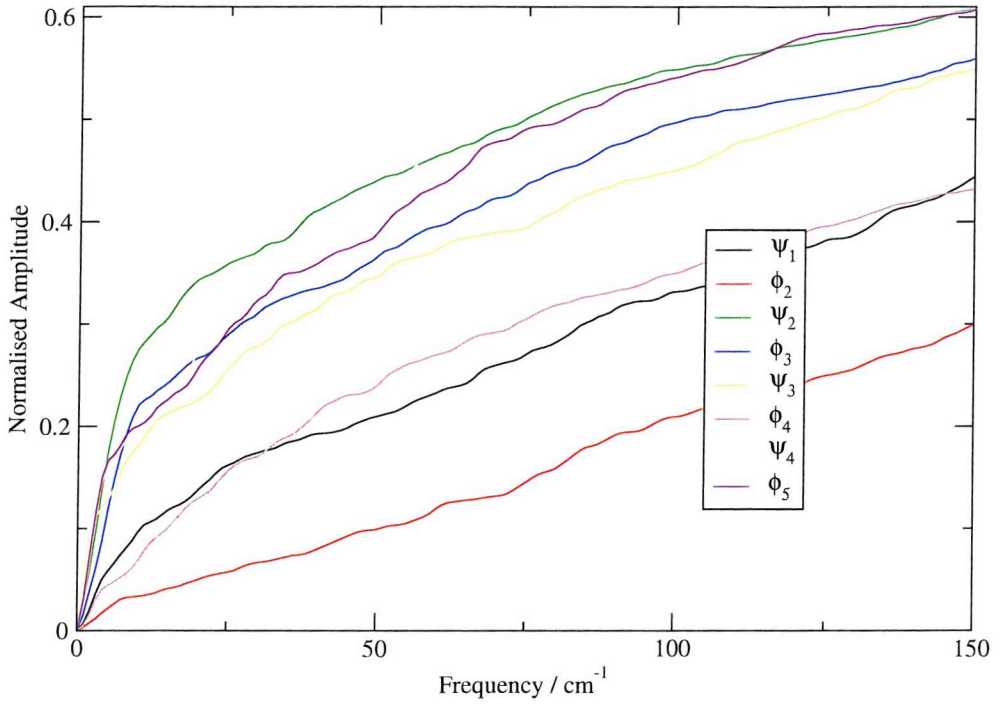
analysed. Figure 5.24 shows these results; plot (a) shows a zoomed in section of Figure 5.2 and plot (b) shows the equivalent results generated for the ω angles. Some motion in the ω angles must be attributed to low frequencies, however less than that of the more mobile dihedral angles. The only exception to this is the $\phi_2$ angle which is highly constrained by the proline ring. Exclusively targeting relevant ψ and ϕ angle motions without increasing ω angle motions is therefore not possible using present RDFMD methodology. However, the ψ and ϕ angles should be amplified to a greater extent than the ω angles, due to the increased components of motions with frequencies below 100 $cm^{-1}$.

A lower internal temperature cap should therefore promote only ψ and ϕ angle motions and not induce *cis-trans* isomerisation of the ω angles. A range of RDFMD simulations have been performed using temperature caps between 500 K and 1500 K to test this hypothesis, and the results are presented in Table 5.1. All simulations were duplicated by replacing the equilibrated YPGDV state velocities with randomly assigned velocities scaled to reproduce the desired 300 K temperature. Results between the original and duplicate trajectories were identical except for the simulation with a filter delay of 50 steps and an internal temperature cap of 1000 K. Here a single transition in $\omega_1$ was seen in the original simulation and no ω flips were seen in the duplicate trajectory.

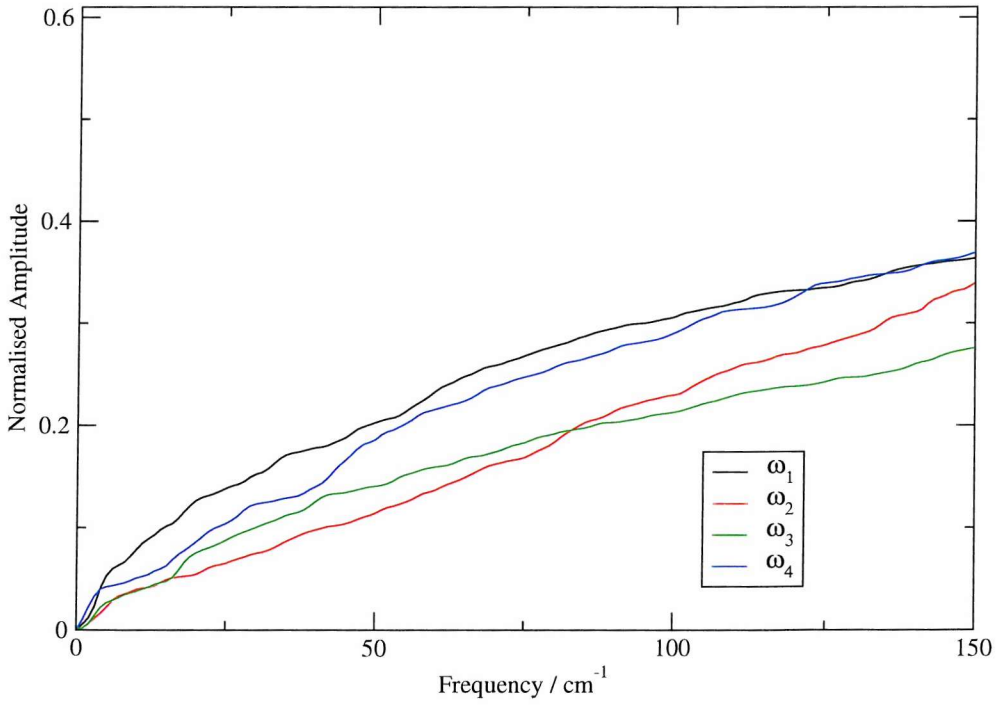
| Internal Temperature cap / K | Flipped ω angles delay = 50 steps | Flipped ω angles delay = 100 steps |
|---|---|---|
| 500 | - | - |
| 600 | - | - |
| 700 | - | - |
| 800 | - | - |
| 900 | - | - |
| 1000 | - / $\omega_1$ | - |
| 1500 | $\omega_1$ | $\omega_1,\omega_2,\omega_3$ |
| 2000 | $\omega_1,\omega_2,\omega_3,\omega_4$ | $\omega_1,\omega_2,\omega_3,\omega_4$ |

Table 5.1: *cis-trans* isomerisation of ω angles using different temperature caps.

Below 1000 K, no *cis-trans* isomerisation of ω angles is seen. Representative ψ/ϕ angle plots for results at 900 K are shown in Figures 5.25 and 5.26 using filter




(a) ψ and φ angles.




(b) ω angles.

Figure 5.24: Cumulative sums of the amplitude spectra of the YPGDV backbone angles

delays of 50 and 100 steps respectively. Interestingly, for all RDFMD simulations using the new protocol and temperatures below 1000 K, no significant motion in the $\psi_1$ angle is seen. This result can easily be justified by inspection of Figure 5.24. By 100 $cm^{-1}$, $\psi_1$, $\phi_2$ and $\phi_4$ all have similar, or lower normalised amplitudes to those shown by the $\omega$ angles. $\phi_2$, as, previously noted, is constrained by the proline ring, and closer analysis of the $\phi_4$ trajectories show that transitions are rarely sampled by RDFMD. Since $\phi_4$ has the highest normalised amplitude over the targeted frequency region, it is the most likely to experience a transition. The cumulative amplitude plots therefore seem to be a reliable indicator of the qualitative internal temperature caps required to promote different motions.

## 5.6 Conclusion

A range of methods have been presented for the systematic analysis of RDFMD parameters and their interdependence. Fourier and Hilbert-Huang techniques have been applied, showing the uses of each method in a frequency-based investigation such as this. The goal of parameterisation has been to maximise motion in $\psi$ and $\phi$ dihedral angles whilst limiting the energy put into the system. The protocols should be applicable to any similar protein, or part of a protein, for which this is the goal. The analysis methods developed here are however intended as a guide for the application of RDFMD to any system, and for any purpose. For example, the measure of success could be applied to a different conformational goal, such as the stretching of a hydrogen bond that RDFMD is being used to break, or the translation along an essential mode that governs domain motions.

The first parameter to be chosen is the frequency target. For this, frequency analysis can be performed on signals extracted from a sample trajectory, using Fourier or Hilbert based methods. Alternatively, the response of the system to filters amplifying different frequency regions can be used. Both methods have been presented here, suggesting filters targeting 0–100 $cm^{-1}$ for promotion of dihedral motions. Empirical Mode Decomposition has been used to separate dihedral signals into high frequency noise from coupled degrees of freedom, a signal in

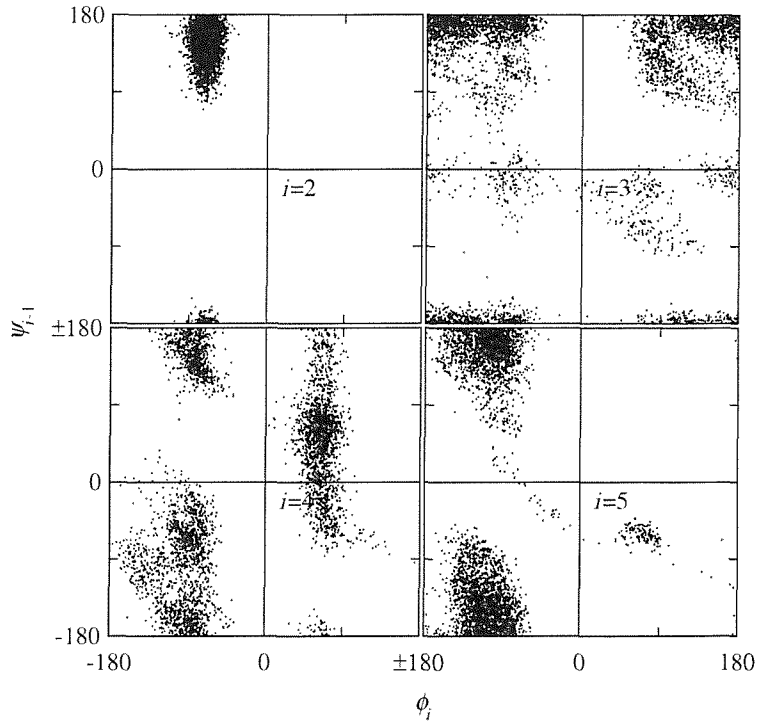



Figure 5.25: Backbone dihedral angle space sampled during 400 ps of RDFMD on YPGDV using a 0–100 $cm^{-1}$, 201 coefficient filter, a filter delay of 50 steps, an amplification factor of 2 and an internal temperature cap of 900 K.

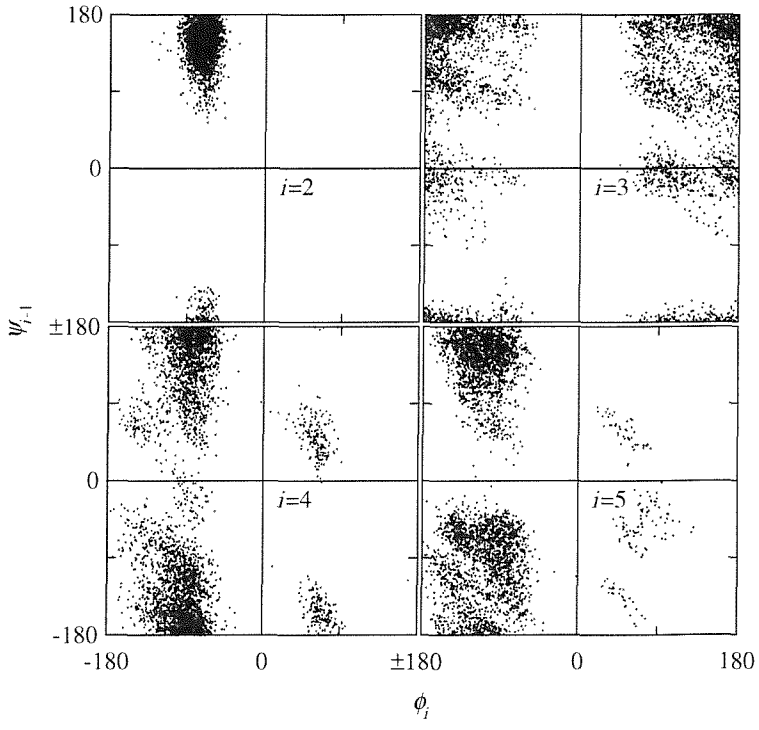



Figure 5.26: Backbone dihedral angle space sampled during 400 ps of RDFMD on YPGDV using a 0–100 $cm^{-1}$, 201 coefficient filter, a filter delay of 100 steps, an amplification factor of 2 and an internal temperature cap of 900 K.

the region 25–100 $cm^{-1}$ that persists throughout simulation, and a low frequency component that only has significant amplitude during rare conformational events. Use of a higher upper limit on the frequency target allows the length of the buffer to be dramatically lowered, thus reducing the computational expense of each filter application. This is particularly important when applying RDFMD to larger systems.

The number of steps between filter applications (the filter delay parameter) has been analysed in detail, including the use of the Hilbert-Huang Transform. Short delays show rapid energy increases that achieve significant conformational change, but quickly overheat the system. Few buffers are completed, limiting the total conformational change achieved. Long delays show no progressive increase in low frequency energy or induced conformational change, the effect of each filter application having dissipated before the next. Intermediate delays of 50 to 100 steps show a compromise between slow energy increases and the rate at which conformational change is induced.

There are several methods of tailoring the amount of energy put into the system. A fixed amplification factor inputs energy according to how much is already present in the target region; this prevents energy being placed into motions of low amplitude that have little relevance to the system. Dynamically adjusting the amplification factor so that a set amount of energy is put into the system is also of use when comparing the system's response to filters that target different frequency ranges. An increase in either kinetic energy (as presented in frequency target section) or an adjustment dependent on the system temperature can be specified.

The resulting parameter set includes a frequency target of 0–100 $cm^{-1}$, a filter with 201 coefficients, a filter delay of 50 or 100 steps and an amplification factor of 2. The maximum of 10 filter caps is not a required parameter as the internal temperature cap is always reached within 10 buffers. Significant conformational sampling using this protocol has been shown and the internal temperature cap tailored to avoid $\omega$ transitions. There is a considerable saving of computational

expense compared to the protocol initially prepared in collaboration at an early stage of the project. The derived protocol requires only 10 % of the number of steps and produces significantly increased $\psi$ and $\phi$ motion sampling.

Cumulative amplitude plots were shown to give an indication of which motions would be most affected by the frequency target chosen. The identification of overlapping frequencies of motion for the $\omega$ angles with some $\psi$ and $\phi$ angles is a major limitation of the RDFMD method to its use maximising dihedral angle motion. The method is only capable of exclusively targeting motions with frequencies that are separable from others present in the system.

The conformer distributions generated by the derived RDFMD protocols will be compared to those found by parallel tempering in Chapter 6. The validity of the parameter set for applications involving large protein systems will also be explored in Chapters 7, 8 and 9.

# Chapter 6

# Parallel Tempering

## 6.1 Introduction

Conventional MD simulation is often of limited use for conformational sampling of protein systems, due to the large number of local minima that can exist on the potential energy surface. The timescales required to leave some minima may be beyond those accessible by simulation using current algorithms. There are several approaches to overcoming this problem, such as driving the simulation over energy barriers, for example, with reversible digitally filtered molecular dynamics (RDFMD). Methods that include a driving force are inherently non-equilibrium, due to the oversampling of transition states, and will not generate the correct distribution of conformers.

An alternative approach to the many-minima problem is to provide assistance over the barriers separating phase space in a statistically rigorous manner, thus maintaining the equilibrium. This is the approach of generalised ensemble methods such as the multicanonical algorithm[81,82] (MUCA), simulated tempering[79] (ST) and the replica-exchange method[82,84] (REM), discussed in Chapter 3. MUCA and ST require weighting factors that are iteratively determined from trial simulations, and the quality of results depends on the convergence of these. The replica-exchange method has no such requirement, and is a highly flexible method that is widely adopted in literature.[5,82,84–86,107] A series of independent replicas are run,

varying across a physical property that is designed to overcome conformational barriers. At least one replica is performed at the desired condition for which a distribution of conformers is required. A Monte Carlo test is used to move replicas between physical states, thus maintaining the statistical rigour of the method.

REM has been performed on biological molecules using several different physical properties as the ensemble coordinate, of which temperature is perhaps the simplest to implement. No manipulation of the protein force field is required, and it is one of the most widely used replica-exchange methods. It is termed parallel tempering (PT), and has been used for a range of applications from protein folding[84,90,107] to the testing of implicit solvent models.[40]

Parallel tempering is severely limited by the computational expense of the algorithm, which increases with system size (see Appendix C.1) and the height of energy barriers to be overcome (see Appendix C.4). Whilst it cannot therefore be considered as a replacement for all traditional MD, the popularity of PT and the availability of suitable codes has increased significantly in recent years. MD and PT simulations reported in literature are consistently increasing in length, fuelled by the falling costs of computer hardware.

## 6.2 Aim of this chapter

In this chapter, parallel tempering will be applied to the RDFMD test-system, YPGDV. This system is known to exhibit quasi-ergodicity; that is, the accessible phase space of the protein has been partitioned into subsets that conventional MD at biological temperatures cannot move between. Whilst YPGDV is a flexible system that is known to form a β-turn in solution, the proline ω angle will always remain in its initial state during a short (i.e. tens of nanoseconds) MD simulation due to the high energy barrier inhibiting transitions. Experimentally, YPGDV is known to exist in solution with a significant proportion of both *trans* and *cis*-proline ω angles. NMR data indicates a 23 % population of *cis*-proline at 278 K,[98] however there is no associated error given with this data, and the NMR was performed in a pH 4.1 (± 0.1) environment. Since aspartic acid has

a solution pKa of 3.65, and simulation will be performed in neutral conditions (for comparison to previous work), some deviation from the experimental results should be expected.

The analysis of an ω dihedral transition, and the resulting *cis-trans* distribution, is beyond the optimisation applied to current force-fields, since only the *trans*-ω is typically sampled. Comparison to experimental results is therefore of limited validity, but the success of the PT algorithm in yielding a consistent dihedral distribution can be thoroughly tested.

The all-*trans* conformations of the protein should be well described by the force field used, and NMR data suggests an approximate 50 % population of β-turn conformers. The same limitations of differing experimental and simulation conditions apply.

Once the conformational distribution is determined by an exhaustive PT simulation, sampling using MD and RDFMD can be analysed. The RDFMD protocol described in Chapter 5 targets maximum dihedral angle motion that should search phase space efficiently, but the generated conformer distribution is likely to have little relevance to that sampled by MD and PT. However, the conformers generated by RDFMD using different internal temperature caps can be compared to those seen with PT, suggesting a suitable value from this parameter.

The PT algorithm will be applied to larger protein systems in later chapters, using the methods and parameters presented here.

## 6.3 The biological importance of *cis*-peptide bonds

For convenience, the ω angle shall be assigned to the residue that contains the nitrogen atom of the peptide bond, and the four ω angles in YPGDV shall be numbered from the N-terminus. The labels, *cis*, and *trans*, shall be used to refer to the conformation of the peptide bond.

Isomerisation of the peptide bond is rarely discussed in computational

chemistry due to the difficulties that established algorithms encounter for such a problem. This lack of interest is partially justified since the stability of the *trans*-ω conformation means that by far the majority of amino acids will never be seen at room temperature in a *cis* state. This is due to the partial double bond character of the peptide bond which stabilises the planar, *trans* configuration. Recently however, a review of structures from the Brookhaven Protein Data Base by Jabs *et al.*, found 43 non-proline *cis*-peptide bonds from 571 proteins (with a total of 153 209 peptide bonds) analysed.[108] A stringent $0° \pm 45°$ was used to define a *cis* conformer. These were most commonly found in high resolution X-ray structures, and the review authors propose algorithms that successfully identify overlooked *cis* peptide bonds in lower resolution structures.

One interesting non-proline *cis*-peptide bond located in a high resolution X-ray structure, is that of between two glycine residues (95 and 96) in E-coli Dihydrofolate Reductase. The formation of this *cis*-peptide bond has been shown to be linked to the rate determining step of protein folding.[109]

Proline is an imino acid, without the partial double-bond stabilisation of the *trans* conformer. Instead, a rigid ring system is connecting to the backbone atoms twice, with no amide hydrogen. Although the *trans* conformer is favoured for steric reasons, the work of Weiss *et al.* found a 5.21 % occurrence of *cis*-proline in protein systems.[110] *Ab initio* calculations on the proline dipeptide indicate activation energy barriers of at least 17.9 kcal $mol^{-1}$ for *trans* to *cis* isomerisation,[111] placing the reaction beyond the timescale of conventional, room temperature, MD simulations.

Several algorithms have been applied to the proline isomerisation problem. Analytical rebridging Monte Carlo[112] combines Monte Carlo moves that can switch ω angles between 0 and 180°, with high temperature parallel tempering. This use of discrete ω states does not sample intermediate configurations, and thus the high barriers that prevent isomerisation are ignored. Simulations must be run using an implicit solvent as the method does not take solvent reorganisation into account. Parallel tempering was used to improve sampling of the isomerisation, and top

temperatures reached $10^7$ K. An AMBER force field was used, and general results for several cyclic peptides were comparable to those found experimentally. This method, although rigorous and quick to converge, is highly constrained, allowing no motion in the proline ring, amino acid side chains, solvent, or deviations from the values chosen for discrete ω angle states. These limitations prohibit the use of the method beyond that of approximating ω angle distributions. This is a worthy target, but is only applicable to the validation of force fields.

Another computational technique to have tackled the isomerisation of proline residues is the reversible scaling of dihedral angle barriers.[113] By reducing the parameters associated with dihedral angle motions, increased sampling of backbone motions is experienced, including isomerisation of the proline ω bond. The AMBER force field was used, along with a Generalised-Born implicit solvent model. Explicit solvent was tested with the method, however ω angles were constrained in a *trans* conformation, due to the reduced rate of isomerisation given by test simulations that were not reported in detail. The different kinetics seen using implicit and explicit solvent suggest some artificial increase in sampling using an implicit solvent, as previously discussed in Chapter 2. Final conformers are reported from simulations with progressively scaled dihedral angle barriers, some containing *cis*-ω angles. Since the isomerisation takes place early in the simulation when energy barriers are low, no information is given about the conformational distribution expected using the unaltered force field.

There is therefore much work to be done in the investigation of ω angle isomerisation using computational methods.

## 6.4 Parallel Tempering Methodology

A replica in parallel tempering requires two labels in order to identify it, one to specify the temperature (subscript), and one to track the coordinate set (superscript in square brackets). A replica 'state' can therefore be described as in Equation 6.1.

$$X = x_m^{[i]} = (\mathbf{r}^{[i]}, \mathbf{v}^{[i]})_m \tag{6.1}$$

In the canonical ensemble the probability of a state existing at a given temperature, $W(x)$, is weighted by the Boltzmann factor, as shown in Equation 6.2. $\beta$ is the inverse temperature, $\frac{1}{k_B T}$, and $Q$ is the partition function; a normalising value equal to the sum of all Boltzmann factors.

$$W(x) = \frac{e^{-\beta E(\mathbf{r},\mathbf{v})}}{Q} \tag{6.2}$$

In this manner, the probability of a general ensemble state existing can be expressed as the product of the Boltzmann factors for each replica (Equation 6.3).

$$W_{REM}(X) = \frac{e^{-\beta_1 E_1} e^{-\beta_2 E_2} \ldots e^{-\beta_S E_S}}{Q_{REM}} \tag{6.3}$$

A Monte Carlo test can then be derived to swap replicas, based upon their potential energies, $E_P$, and temperature. The test is shown in Equation 6.4. Since the difference between temperatures affects the probability of acceptance exponentially, only neighbouring replicas are tested. To move replicas in temperature space, velocities are rescaled by the ratio of the two temperatures as shown in Equation 6.6.

$$w(X \rightarrow X') \equiv w(x_m^{[i]} | x_n^{[j]}) = max\{1, e^{-\Delta}\} \tag{6.4}$$

where

$$\Delta = (\beta_n - \beta_m)(E_P(\mathbf{r}^{[i]}) - E_P(\mathbf{r}^{[j]})) \tag{6.5}$$

$$\begin{cases} X = x_m^{[i]} \rightarrow X' = x'^{[i]}_n \\ Y = x_n^{[j]} \rightarrow Y' = x'^{[j]}_m \end{cases} \quad \begin{cases} \mathbf{v}^{[i]\prime} \equiv \sqrt{\frac{T_n}{T_m}} \mathbf{v}^{[i]} \\ \mathbf{v}^{[j]\prime} \equiv \sqrt{\frac{T_m}{T_n}} \mathbf{v}^{[j]} \end{cases} \tag{6.6}$$

In order for the exchange process to converge to equilibrium, it is necessary that the detailed balance condition (that forward and backward transitions are equally likely) holds on the transition probability $w(X \rightarrow X')$. The derivation of

the PT test and the proof that the temperature rescaling in Equation 6.6 satisfies the detailed balance condition is shown in Appendix C.2.

Parallel tempering is performed from replicas that have been equilibrated to a chosen temperature. Each replica is run for a set period of time and pairwise swaps attempted using the given test. By repeating the exchange process, testing against alternate replica neighbours, movement in temperature space is achieved according to the location of a replica on the potential energy surface. If a replica approaches an energy barrier, its potential energy will increase, and it is likely to swap to a higher temperature. If, at this higher temperature, the replica is able to overcome the barrier and reaches a region of lower potential energy, it is then likely to swap back to a lower temperature. Thus random walk in conformational space is induced whilst maintaining statistical ensembles at each temperature.

Figure 6.1 shows an example trajectory for a simple parallel tempering simulation consisting of three blocks of MD, separated by PT tests. Replica A passes two tests, moving upward in temperature each time. The replicas at the maximum and minimum temperatures in the ensemble are tested half as often as others, as the test is applied to alternate neighbours.

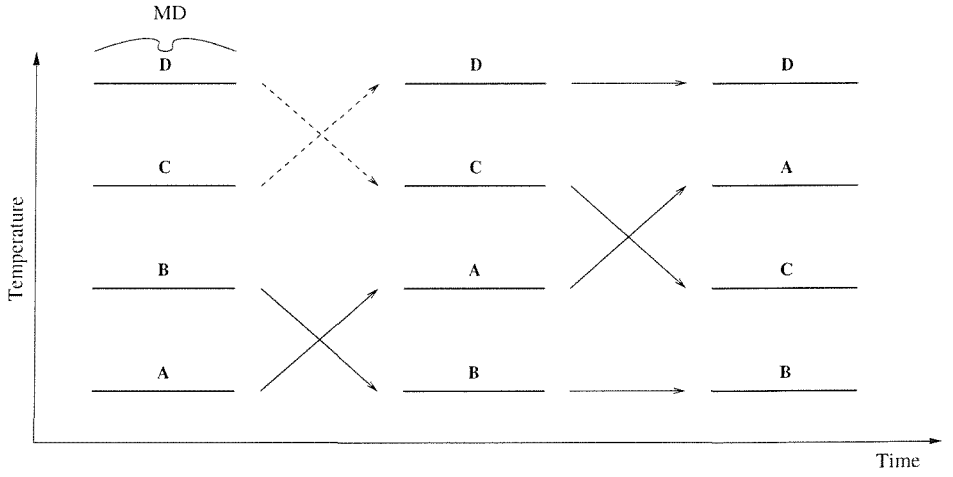



Figure 6.1: An example parallel tempering sequence. Three blocks of MD are shown separated by attempted tests denoted by arrows. A dotted arrow indicates a failed test, whereas a solid arrow line indicates a successful PT move.

The potential energy surface is unaffected by the level of thermal energy available to a system, but presently used force fields are optimised to give correct

results at a single temperature (typically 298 K). Increasing kinetic energy has the desired effect of allowing greater motion across the potential energy surface. It is important to be clear that motion is not occurring across a different surface for the new temperature. The high energy information in a potential energy surface is not optimised in the same fashion as at the minima and is therefore likely to be less accurate. For example, different force fields are suspected to yield different transition temperatures for β hairpin folding.[87] This does not limit the validity of results taken from the temperature at which the force field was optimised. Various force fields have been tested with PT and conformer population distributions can be obtained at 298 K that are in agreement with experimental data.[88]

### 6.4.1 Thermostat parameterisation

The convergence of a PT simulation depends on the mobility of the replicas in temperature space. Thus it is most efficient to attempt replica swaps frequently. However, a replica is not equilibrated at its new temperature immediately after its velocities have been rescaled. Initially the kinetic energy will drop, as the potential energy adjusts to the new level of thermal energy available. The rate at which temperature will be equilibrated depends upon the strength of the thermostat used. Too strong a thermostat and the system will exhibit oscillations in the potential energy about a suitable value.[111] Too weak a thermostat and PT swaps will have to be performed infrequently, reducing the efficiency of the method. This topic is not often discussed in literature applications of PT, in which the time between swaps is generally the only parameter reported.

Figure 6.2 shows the results of applying several damping parameters with a Langevin Thermostat.[101] The dashed line indicates an average potential energy generated over 50 k steps. The results correspond to a simulation from the equilibrated YPGDV system (described in Chapter 4) with the velocities rescaled from 300 K to 320 K at the start of the simulation. A damping parameter of 1 $ps^{-1}$ has been used for NPT MD simulations in this project, but the time taken to equilibrate the potential energy after a 20 K increase is approximately 5 ps (beyond the scale

shown). A 10 $ps^{-1}$ damping parameter raises the potential energy rapidly, however significant oscillations occur. A damping parameter of 5 $ps^{-1}$ shows a compromise between rapid equilibration, and the magnitude of potential energy oscillations. After 1 ps the potential energy reaches the equilibrated value and fluctuations remain approximately constant. PT simulations presented in the thesis therefore use a 5 $ps^{-1}$ damping parameter is therefore used, and tests are attempted every 1 ps.

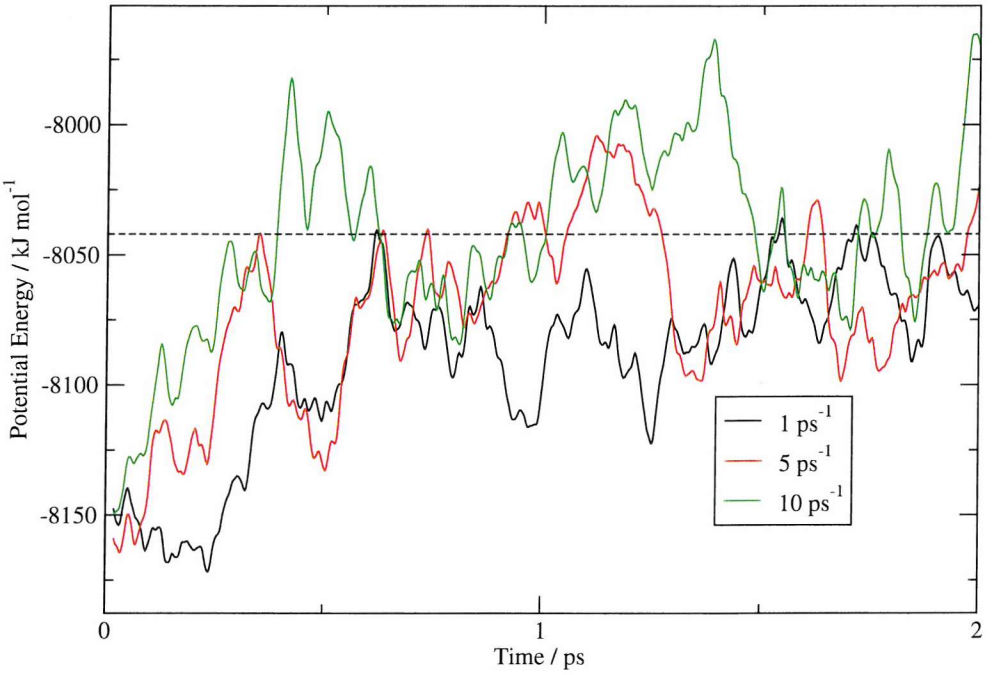



Figure 6.2: Analysis of Langevin thermostat equilibration using a range of damping parameters. The dashed line indicates the target potential energy mean, as determined over 50 k steps of simulation.

### 6.4.2 Temperature distribution

Another factor affecting the efficiency of the PT algorithm is the temperature distribution of replicas. Replicas with temperatures closely spaced will exchange frequently due to the high overlaps of the potential energy distributions, but many replicas will be required to span a given temperature interval. If spaced far apart, fewer replicas are required, but the acceptance probability will be greatly reduced. Also, the energy fluctuations of a system increase as temperature is increased (see Appendix C.1), and thus if replicas are evenly spaced, the acceptance probability

as a function of temperature (the probability profile) will not be constant. For maximum efficiency, a uniform probability profile is desired, and a method is required to produce a temperature distribution for a given acceptance probability.

Short simulations run at intervals across the temperature range of interest can give the mean and variance of potential energy, $E_P$, as a function of temperature. From these, potential energy distributions can be approximated for any temperature. The total probability of accepting a parallel tempering move is then the product of the probability, $W$, of finding the first replica, $x_m^{[i]}$, at a certain energy, $E_s$, the probability of finding the second replica, $x_n^{[j]}$, at a certain energy, $E_t$, and the probability of accepting a PT move given the specific energies and temperatures of the two replicas, $max\{1, e^{-\Delta_{n,m}^{[s],[t]}}\}$. This is summed across all combinations of energy bins as shown in Equation 6.7.

$$W\left(x_m^{[i]}|x_n^{[j]}\right) = \sum_{s=1}^{N}\sum_{t=1}^{N} max\{1, e^{-\Delta_{n,m}^{[s],[t]}}\} W\left(E_P(x_m^{[i]}) = E_s\right) W\left(E_P(x_n^{[j]}) = E_t\right) \quad (6.7)$$

$$\Delta_{n,m}^{[s],[t]} = (\beta_n - \beta_m)(E_s - E_t)$$

Once the temperatures of two replicas are determined, the interval between them can be extrapolated using an exponential distribution that should generate a uniform probability profile (see Appendix C.3). This approach has not been used in this project as differents timesteps are used for different temperatures. When simulating at high temperatures, energy flucations are increased, and thus the timestep must be reduced to maintain energy conservation. Reducing the timestep does not change the average energy of simulation, but will reduce its variance. Therefore the calculation shown in Equation 6.7 is iteratively performed, increasing the temperature by 0.1 K intervals and recalculating the probability of acceptance. Once this is greater than a specified value, the new temperature is accepted and the calculation repeated to find the next. Since this calculation is dependent on functions that describe the mean and variance of potential energy as a function of temperature, fitted across the entire temperature range, the effects of

changing the timestep are taken into account.

### 6.4.3 Acceptance probability target

The choice of acceptance probability is a compromise between the expense of the method and the rate of convergence. A high acceptance rate will require more replicas to span the same temperature interval, as they must be spaced closer together. A low acceptance rate requires fewer replicas, however they will be less mobile and more simulation may be required for convergence.

To ensure sufficient replica mobility, a 0.4 probability of acceptance has been initially chosen. Further investigation is required into the effects of this parameter.

### 6.4.4 Simulation size

Simulation size is of great importance to the parallel tempering algorithm, since the number of replicas required to span a specified temperature gap scales in the order of $\sqrt{D}$, where $D$ is the number of degrees of freedom in a system. Large systems become significantly more expensive because of this, and many PT protein simulations use implicit solvent models. This dramatically reduces the number of degrees of freedom, but the limitations of results are linked to the quality of the solvent model used (as discussed in Chapter 2) and consequently explicit solvent is used throughout this project.

## 6.5 Parallel tempering simulations of YPGDV

All PT simulations reported in this thesis are run using NVT simulation with the NAMD package. A particle mesh Ewald treatment of electrostatics is used, with SHAKE applied to all bonds containing a hydrogen atom. A switching function is applied to Lennard-Jones interactions between 8 Å and the 12 Å cutoff. A Langevin thermostat damping parameter of 5 $ps^{-1}$ is used for all PT simulations, with tests attempted every 1 ps with alternate neighbours. The YPGDV PT targets a 0.4 probability of accepting PT moves, and replicas are equilibrated using 100 ps

of NVT at the desired temperature. Sampling rates of all conformer distributions are adjusted for comparison purposes to include 4000 conformations on the $\psi$, $\phi$ plots. Other results, such as the distribution comparison functions and secondary structure use a sampling rate of 0.1 ps. The CHARMM force field is used for all YPGDV simulations and unspecified simulation parameters are as described in Chapter 4.

### 6.5.1 Early simulations

The first YPGDV PT simulation was setup to test the method of generating a temperature distribution and used 30 replicas from 284.4 to 510.7 K. 510.7 K was determined as the maximum temperature for use with a 2 fs timestep, by NVE simulations that showed a significant increase of energy fluctuations above 500 K. 200 ps NVT simulations were performed from 250 to 550 K to generate the parameters for the temperature distribution calculation. 2 ns of parallel tempering was performed, giving a total of 60 ns simulation time. Conformers generated at 300.0 and 510.7 K are shown in Figures 6.3 and 6.4 respectively.

Early RDFMD simulations showed significantly more dihedral angle sampling than that reported in Figure 6.4. Particularly, a conformer seen with $\psi_1$ between 0 and -180° is not sampled by the PT simulation. It is a requirement of parallel tempering that the maximum temperature is sufficient to overcome all energy barriers in the system, otherwise simulation may only sample a subset of accessible phase space for the desired temperatures. The maximum temperature is clearly insufficient and the simulation can be discarded.

A second PT YPGDV simulation was performed reaching a maximum temperature of 906.7 K. 900 K was determined to be sufficient to see isomerisation in the proline $\omega$ angle using thermal NVT simulations at 100 K intervals from 500 K to 1500 K. A timestep of 1 fs is used between 500 and 1000 K and a timestep of 0.5 fs is used above this. The temperature distribution calculation was parameterised by 200 ps NVT simulations at 50 K intervals from 250 to 900 K. 52 replicas are required and 2 ns of PT was performed, giving a total of 104 ns of simulation. The

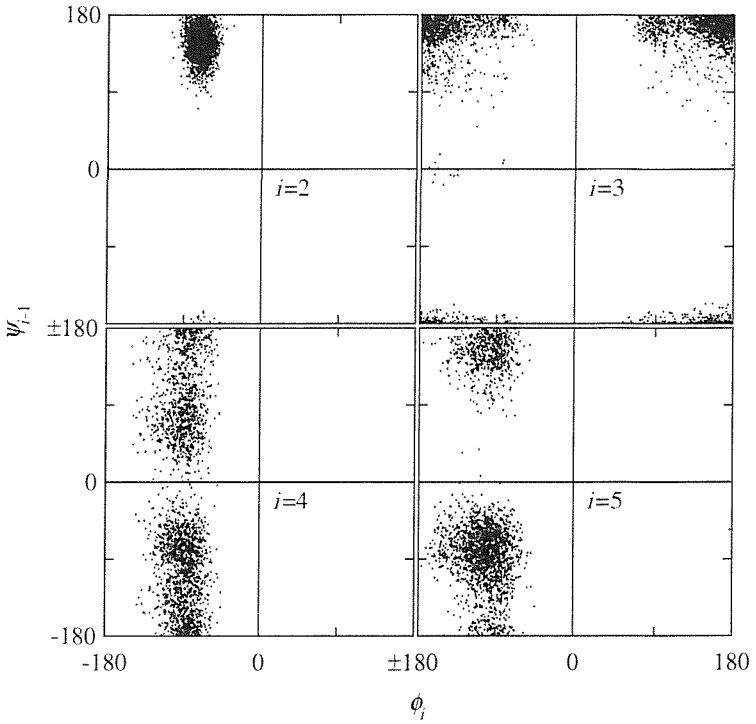



Figure 6.3: Backbone dihedral angle space sampled by 300.0 K replicas during 2 ns of REM on YPGDV. A maximum temperature of 510.7 K was used.

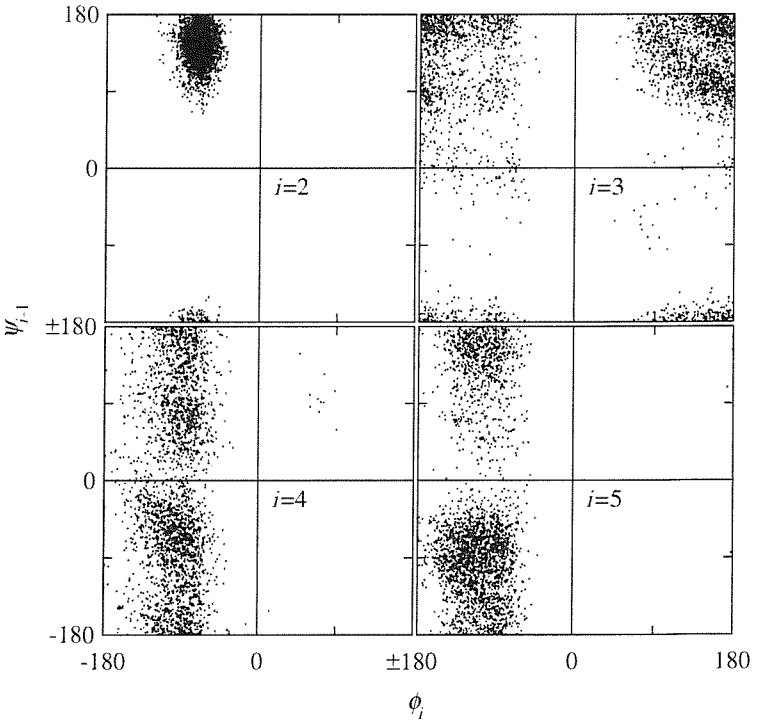



Figure 6.4: Backbone dihedral angle space sampled by the highest temperature replicas (510.7 K) during 2 ns of REM on YPGDV.

dihedral angle sampling of replicas 300.0 K and 906.7 K are shown in Figures 6.5 and 6.6 respectively.

At the highest temperature, conformers visited by early RDFMD simulations are seen, and the proline ω angle ($\omega_1$) undergoes isomerisation. However, the ω angle between glycine and aspartic acid ($\omega_3$) also undergoes infrequent transitions. Without long simulations at high temperatures, it cannot be determine whether the other ω angles will flip in a longer PT simulation. Since these transitions are poorly sampled and *cis*-conformers are likely to have high potential energies, several replicas could become effectively 'stuck' at the highest temperatures. For example, three transitions to *cis*-$\omega_3$ are seen, each occurring at the highest temperature. No transitions back to a *trans* conformation are seen, and at the end of the simulation, the three *cis*-$\omega_3$ conformers occupy three of the four highest temperature replicas. This reduces the mobility of replicas in the high temperature region and since there are few replicas run at high enough temperatures to sample the desired $\omega_1$ transitions, the PT simulation is unlikely to converge.

A final YPGDV PT simulation was set up to reach a maximum temperature of 1237.5 K. 1200 K was known to be sufficient to see all ω transitions from the NVT simulations at 100 K intervals up to 1500 K previously mentioned. 61 replicas are required, and after seeing several transitions in each ω angle over 2 ns of PT simulation, the temperature cap of 1237.5 K was deemed sufficient.

## 6.6 Parallel Tempering results on YPGDV

The 61-replica PT simulation of YPGDV was run for 20 ns, giving a total of 1.22 $\mu$s simulation time. Each replica is identified using an alphabetical label, and the temperature distribution is shown in Figure 6.7. An exponential curve has been fitted to the distribution, shown with a solid line. The temperature distribution is clearly exponential, as predicted in Appendix C.3. The probability profile is presented in Figure 6.8, and a dashed line indicates the transition from use of a 2 fs to a 1 fs timestep. The presence of a glitch in the probability profile at this point indicates that by fitting a temperature distribution across all temperatures,

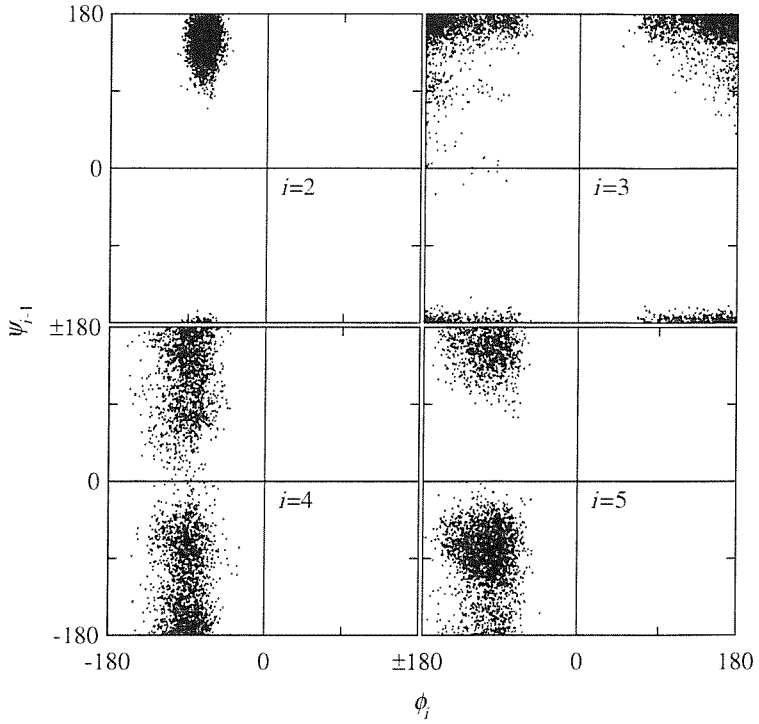



Figure 6.5: Backbone dihedral angle space sampled by 300.0 K replicas during 2 ns of REM on YPGDV. A maximum temperature of 906.7 K was used.

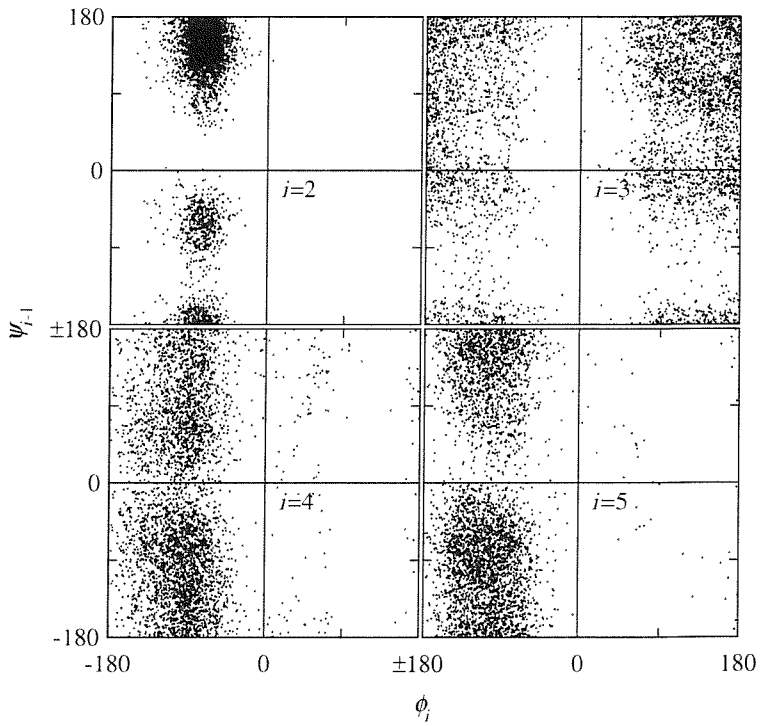



Figure 6.6: Backbone dihedral angle space sampled by the highest temperature replicas (906.7 K) during 2 ns of REM on YPGDV

the effects of changing the timestep have not been fully modelled. However, the probability acceptance target of 0.4 is well reproduced across the entire temperature range. Since alternate neighbours are tested, the end replicas can only be tested half as often, and thus have a probability of acceptance of half the targeted value.

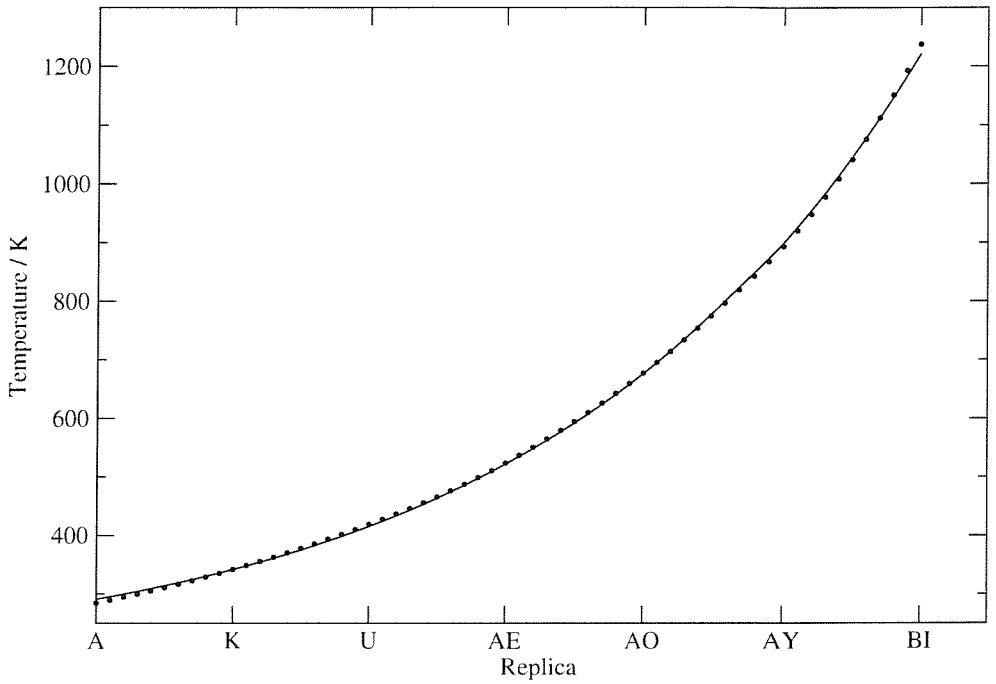



Figure 6.7: Temperature distribution used for YPGDV parallel tempering. The black line represents an exponential curve fitted to the data.

The mobility of replicas in temperature space can be shown by plotting the temperatures a coordinate set visits throughout the PT trajectory. The temperatures visited by four replicas initially spaced evenly through the general ensemble are shown in Figure 6.9. Each replica visits both the top and bottom temperatures during simulation.

The dihedral angle space sampled by *trans* conformers during the PT trajectory is shown in Figures 6.10 and 6.11 for temperatures 300.0 and 1237.5 K respectively. These figures are included for comparison to all-*trans* RDFMD and MD simulations previously reported. Results at intermediate temperatures are included in Appendix D. As the temperature is decreased, the clouds of dots representing conformers become more defined, and gradually separate as they drop below

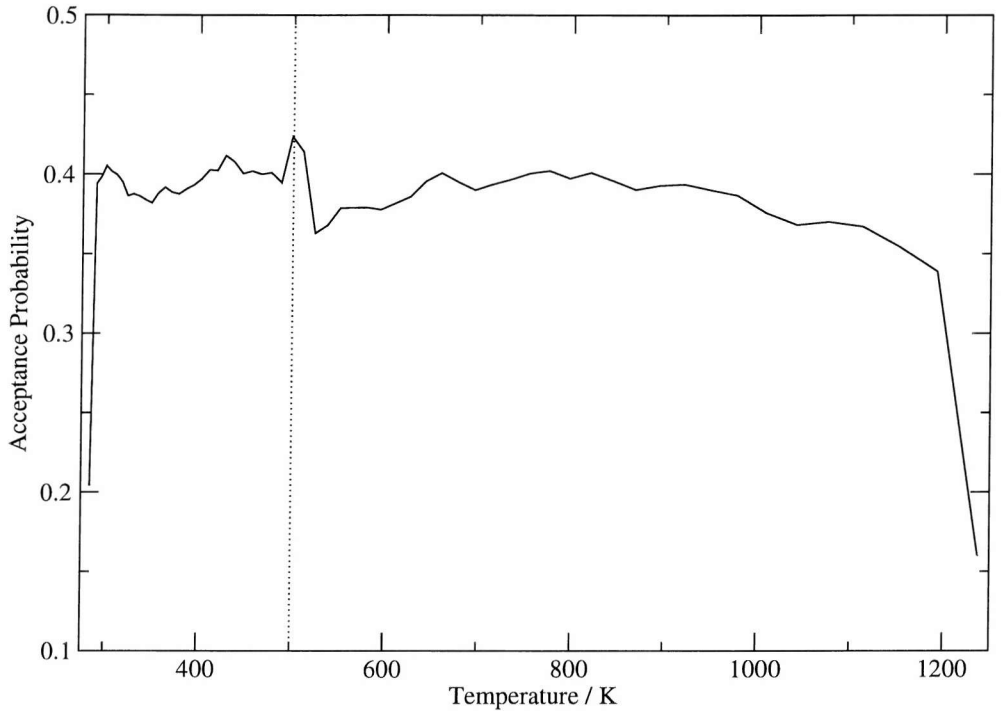



Figure 6.8: Probability profile from 20 ns of YPGDV parallel tempering.

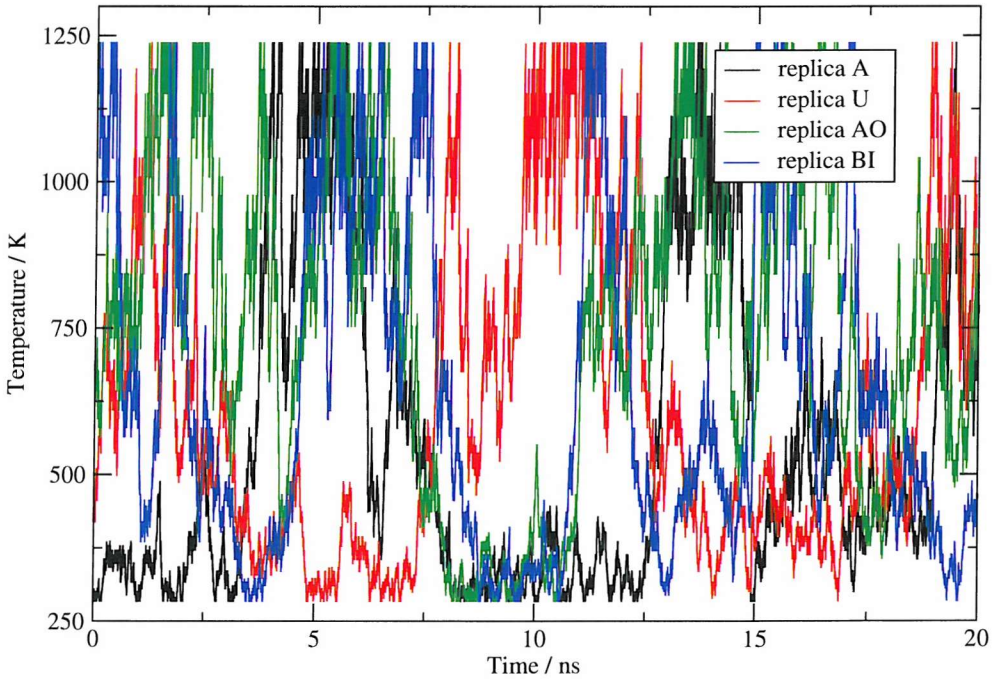



Figure 6.9: Temperature mobility of replicas. Each replica reaches top and bottom temperatures. Four evenly spaced replicas shown.

temperatures at which transitions occur.

The population of *cis* and *trans*-ω angles begins with all but one conformer in an all-*trans* state. This one transition was sampled during the 100 ps equilibration

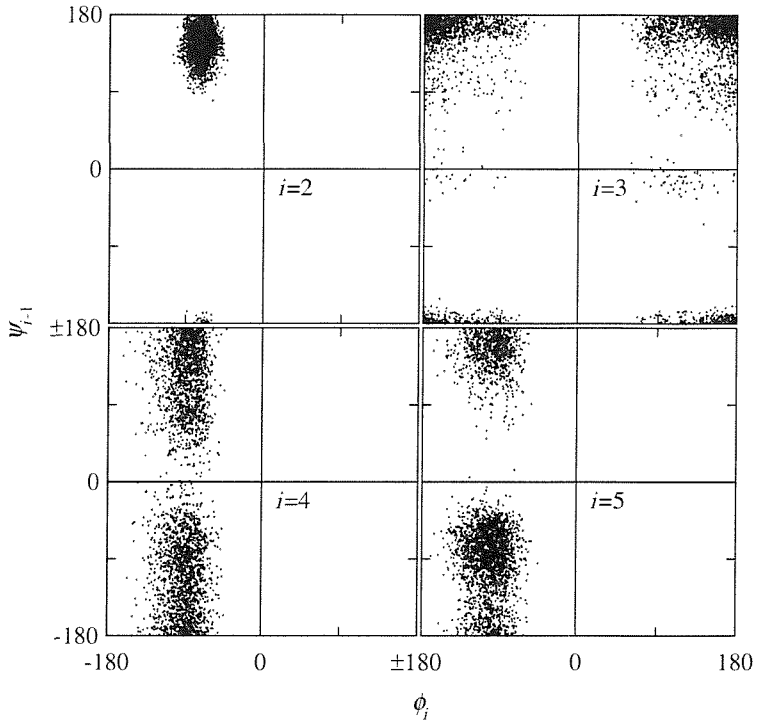



Figure 6.10: *trans*-ω backbone dihedral angle space sampled by replicas at 300.0 K. A maximum temperature of 1237.5 K was used.

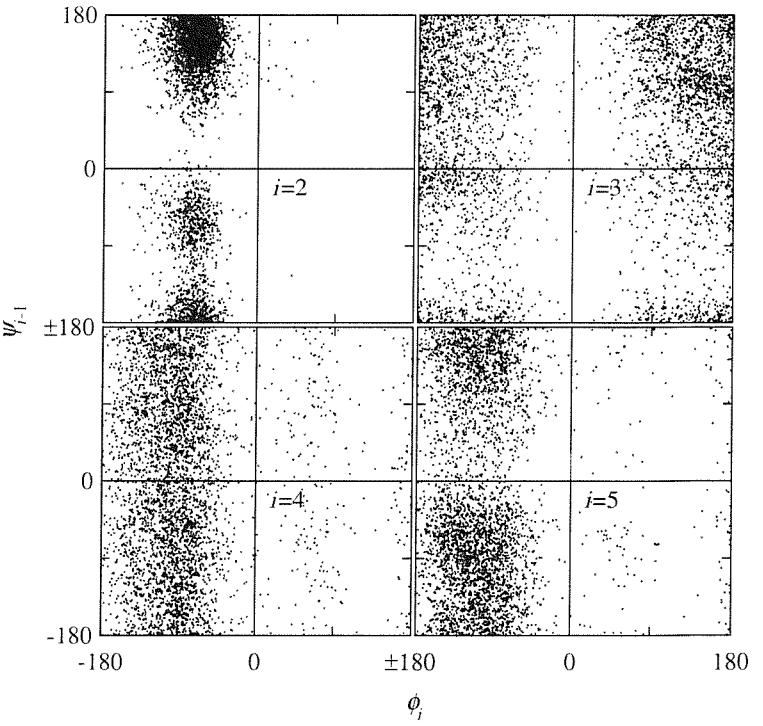



Figure 6.11: *trans*-ω backbone dihedral angle space sampled by the highest temperature replicas (1237.5 K) during 20 ns of REM on YPGDV

stage. As time proceeds, *cis*-ω conformers are generated at high temperatures, and, where suitable, the PT algorithm allows them to drop in temperature. Thus the ω angle distributions at high temperatures rely on the rate of transitions, but distributions at lower temperatures are generated purely by replica swapping. It can be expected, therefore, that the distributions at the highest temperatures equilibrate most rapidly, and this is indeed the case. A system disturbed from equilibrium will exponentially decay back towards its equilibrium proportion (see Appendix E). The population of *cis-trans* ω states at 1237.5 K has therefore been approximated by an average across the replicas that visit the temperature in 2 ns blocks, and an exponential decay curve has been fitted with a correlation coefficient above 0.9. The result is shown in Figure 6.12, and the half-life of decay is 1.6 ns. With such a short half-life the population of *trans*-$\omega_1$ in the later half of the simulation is close to the equilibrium proportion of 0.43 predicted by the exponential fit. A standard error can be assigned by taking the average *trans* population over the last five 2 ns blocks, which gives a mean proportion of 0.42 ± 0.03 *trans* conformers.

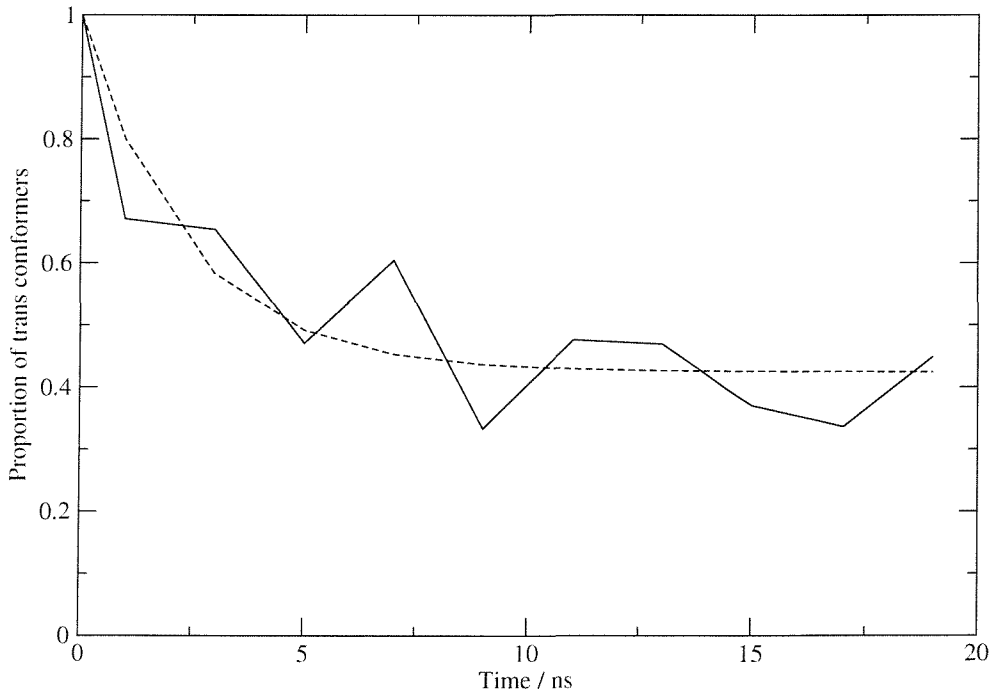



Figure 6.12: The proportion of *trans*-proline conformers of replicas visiting 1237.5 K. The dashed line represents an exponential curve fitted to the data.

Checking for the convergence of $\omega_1$ populations at lower temperature where transitions cannot occur is more difficult. Figure 6.13 plots the $\omega_1$ distributions,

again averaged in 2 ns blocks, for a range of temperatures. It would appear that all temperatures are converging to the same $\omega_1$ population as that seen in 1237.5 K although it is not clear whether the lower temperature populations are stable. Further simulation is required to generate a reliable estimate of the conformer distribution, although it is clear that the proportion of *cis*-proline conformers is significantly higher than the 23 % suggested by NMR.[98] It is not possible to determine whether this is due to an inaccuracy of the force field used, or to the different conditions under which the NMR was performed.

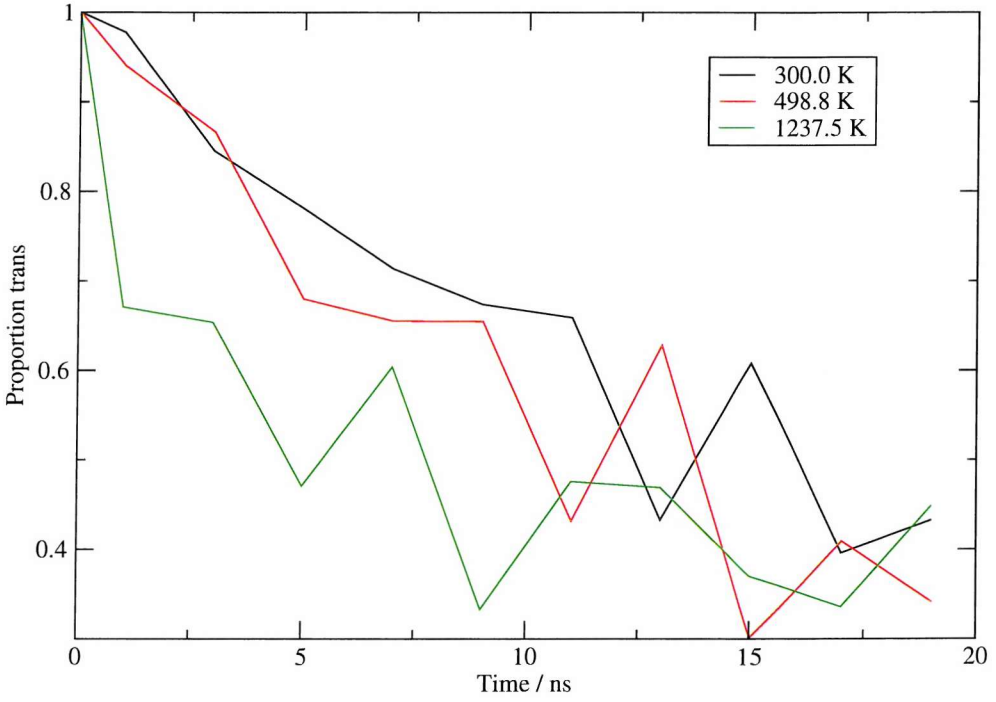



Figure 6.13: The proportion of *trans*-proline conformers throughout 20 ns of PT.

The populations of the other three ω angles in YPGDV are shown in Figure 6.14, averaged over the last 5 ns of PT simulation. No non-proline *cis*-conformers have reached 300.0 K although significant populations exist at higher temperatures. Owing to the higher energy barriers for the isomerisation of non-proline ω angles, fewer transitions are sampled, and it is unclear whether the data has converged.

Figure 6.15 shows the number of ω transitions sampled at different temperatures. A transition has been defined by the crossing of ± 90° by at least 10°. The total number of transitions sampled are: 2048, 256, 384 and 232 for $\omega_1$, $\omega_2$, $\omega_3$ and $\omega_4$ respectively, giving just under three thousand isomerisation events. To the

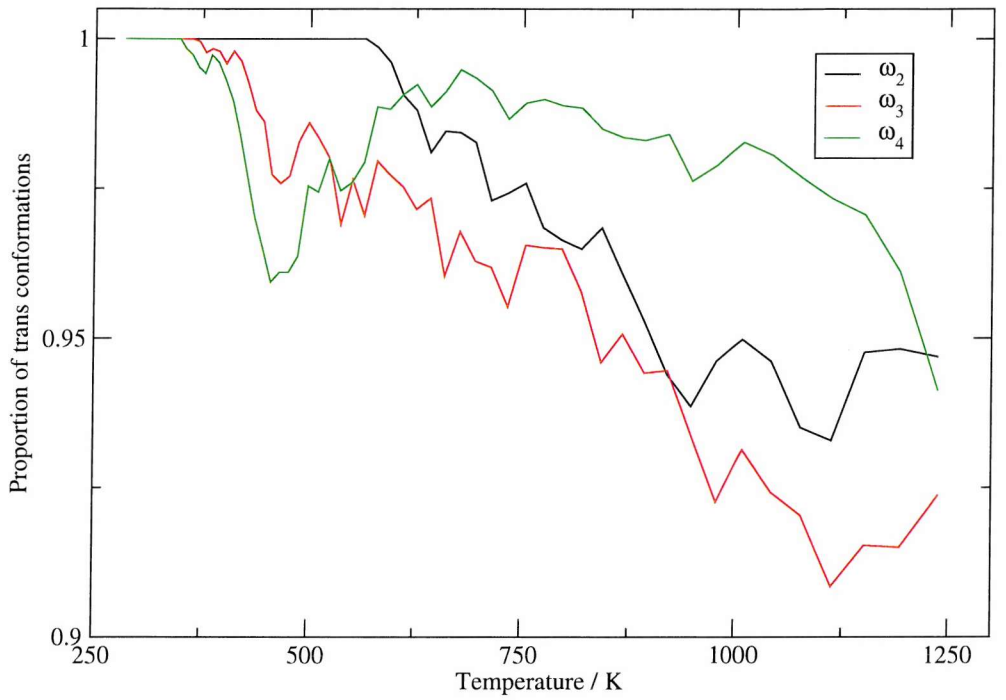



Figure 6.14: ω angle populations averaged over the last 5 ns of the 20 ns PT simulation.

author's knowledge, there are no published results from simulation studies similar to those shown in Figure 6.15.

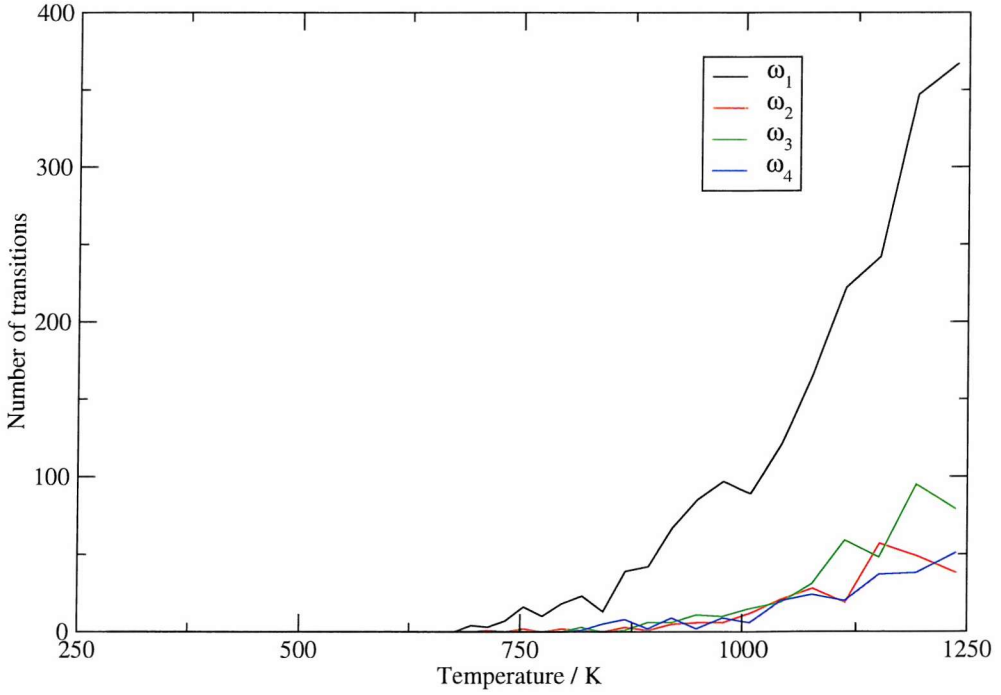



Figure 6.15: The number of isomerisations seen throughout 20 ns of PT.

### 6.6.1 Analysis of conformers generated at 300.0 K in the PT simulation

A secondary structure analysis of the all-*trans* conformers from the 20 ns PT trajectory has been performed and the results are shown in Figure 6.16. The DSSP (Dictionary of Secondary Structure Prediction[106]) was used, with the default hydrogen bond energy cutoff of -0.5 kcal $mol^{-1}$. The bend conformer (3-β turn) is clearly dominant, accounting for 38 % of all sampling. This value is below the 50 % estimate offered by experimental data,[98] although no error is included for this figure, and the accompanying text indicates that it should be loosely interpreted.

For comparison a secondary structure analysis of the 30 ns NPT MD simulation is shown in Figure 6.17. Further details of this simulation are provided in Chapter 5. Again the bend conformer is dominant, however the population sampled is lower, at 32 %- moving away from the experimental data. The PT replicas see an extra secondary structure element, that of a 3-turn between proline and valine, and the corresponding hydrogen bond formed between the proline carbonyl oxygen and the valine amide hydrogen. The PT replicas also show increased sampling of the 4-turn between tyrosine and valine. One point of interest is the regularity of sampling across the PT trajectory; for example, the hydrogen bond between the aspartic acid amide hydrogen and the valine carbonyl occurs frequently and intermittently, but is seen in dense 'clumps' in the MD trajectory. This effect is due to the parallel nature of the PT algorithm, with many trajectories contributing to the conformer distribution.

The secondary structure analysis is insufficient to show significant difference between the PT and MD conformational distributions, and an alternative method of viewing the data is required.

## 6.7 Comparing conformational distributions

The most detailed analysis of generated conformers presented so far has been with plots showing the $\phi$ and $\psi$ backbone angles. These are of limited use when com-

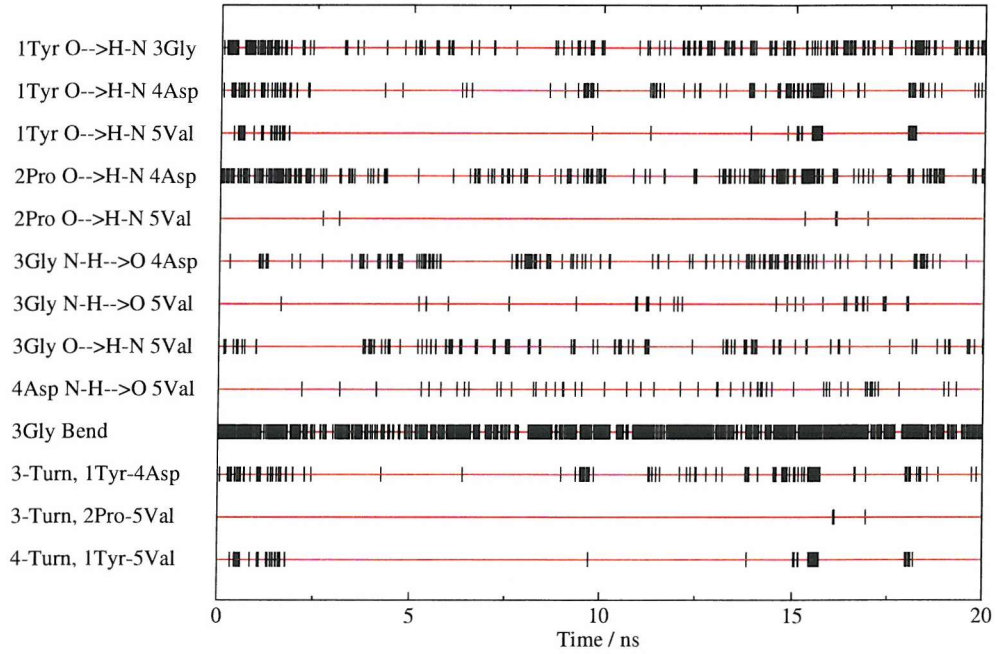



Figure 6.16: Secondary structure analysis of 300.0 K replicas over 20 ns PT.

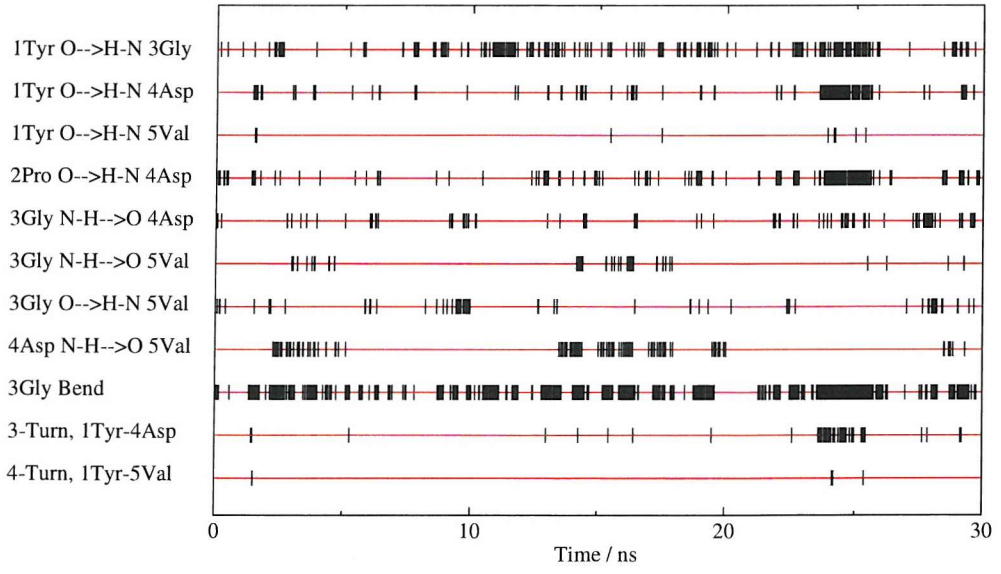



Figure 6.17: Secondary structure analysis of 30 ns NPT MD simulation.

paring the distributions of YPGDV conformers generated by different methods. Although eight dimensions of data are represented on the plots, information about the interdependence of these is lost for all but the two angles placed together in a quadrant. Two functions are therefore defined that address this issue, allowing eight dimensions of data to be compared. All analysis in this section is based upon *trans*-peptide conformers.

### 6.7.1 8D Distribution Similarity Function

The 8-dimensional distribution similarity function, $S(X,Y)$, splits two distributions, $X$ and $Y$ into bins, and reports a scalar value that is the proportion of overlapping distributions. Equation 6.8 shows the summation required for this calculation, where $X(i)$ indicates the proportion of population $X$ in the $i^{th}$ bin. The simplest use of the function is to separate each dihedral angle into a number of bins of equal width. 120° bins prove suitable, giving a total number of possible conformers, $N$, from the eight YPGDV backbone dihedral angles of, $\left(\frac{360}{120}\right)^8 = 6561$. If separated into narrower bins, the 'noise' of conformers existing on a boundary leads to increased errors. It is worth noting that $S(X,Y)$ equals $S(Y,X)$ in every case.

$$S(X,Y) = \sum_{i=1}^{N} min\{X(i),Y(i)\} \tag{6.8}$$

Figure 6.18 shows an application of the similarity function, in which different lengths of the NPT MD simulation previously mentioned are compared to the all-*trans* conformational distribution given at each parallel tempering temperature. Clearly there is a peak at 300 K, indicating that, unsurprisingly, the MD distribution is most similar to that generated at an identical temperature. There is however little difference indicated between MD simulations of 10, 20 and 30 ns.

A more satisfactory approach to the division of each dihedral angle into bins of equal width, is to separate each angle according to the minima of a frequency histogram of the data. Thus bins are created with more physical meaning, and the noise associated with having clusters centered close to a bin boundary is reduced.

Fortunately the YPGDV system has well-defined dihedral histograms that can be easily separated into conformers by eye. The results of splitting each dihedral angle using histograms from a range of PT temperatures are shown in Table 6.1. An alternative clustering method was attempted using an algorithm developed by Wu *et al.*,[97] but the method was highly dependent on chosen parameters. Heavily populated conformers were often insufficiently separated, and sparsely populated conformers were generally ignored.

| Dihedral Angle, $x$ | Conformers identified |
|---|---|
| $\psi_1$ | $(-122 < x < 27)$ $(27 < x,\ x < -122)$ |
| $\phi_2$ | $(-180 < x < 0)$ $(0 < x < 180)$ |
| $\psi_2$ | $(-65 < x < 45)$ $(-120 < x < -65)$ $(45 < x < 125)$ $(x > 125,\ x < -120)$ |
| $\phi_3$ | $(-125 < x < 0)$ $(0 < x < 125)$ $(x > 125,\ x < -125)$ |
| $\psi_3$ | $(-133 < x < 0)$ $(0 < x < 117)$ $(x > 117,\ x < -133)$ |
| $\phi_4$ | $(-180 < x < 0)$ $(0 < x < 180)$ |
| $\psi_4$ | $(-138 < x < 35)$ $(x < 138,\ x > 35)$ |
| $\phi_5$ | $(-180 < x < -30)$ $(-30 < x < 180)$ |

Table 6.1: Conformers identified from the analysis of YPGDV dihedral angle histograms of PT trajectories.

Using the results of Table 6.1, the similarity function analysis of different lengths of MD simulation can be regenerated and are shown in Figure 6.19. A progressive improvement of the conformational distribution given by MD is now seen for longer simulations. The distribution of conformers generated by 30 ns of MD simulation overlaps that of the PT simulation by over 95 %. Although the division of conformers may appear non-rigorous, moving the bin boundaries by 5° in either direction, changes the results of Figure 6.19 for the 30 ns trajectory by less than 5 % for each PT temperature.

### 6.7.2 Phase Space Sampling Comparison Function

The phase space sampling comparison function, $B(X,Y)$, compares the occupied bins of two distributions and produces the proportion of bins seen by $X$ that are also seen by $Y$ (Equation 6.9). $B(X,Y)$ is not necessarily equal to $B(Y,X)$. In this

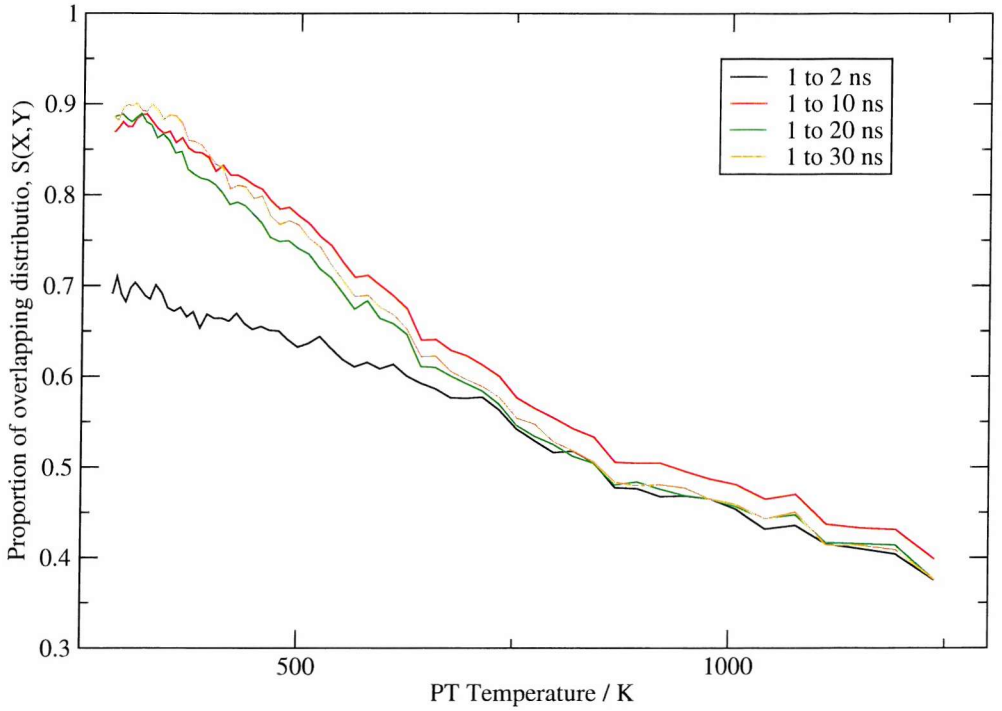



Figure 6.18: The 8D distribution similarity function of MD conformations against those generated by PT. 120° bin width used.

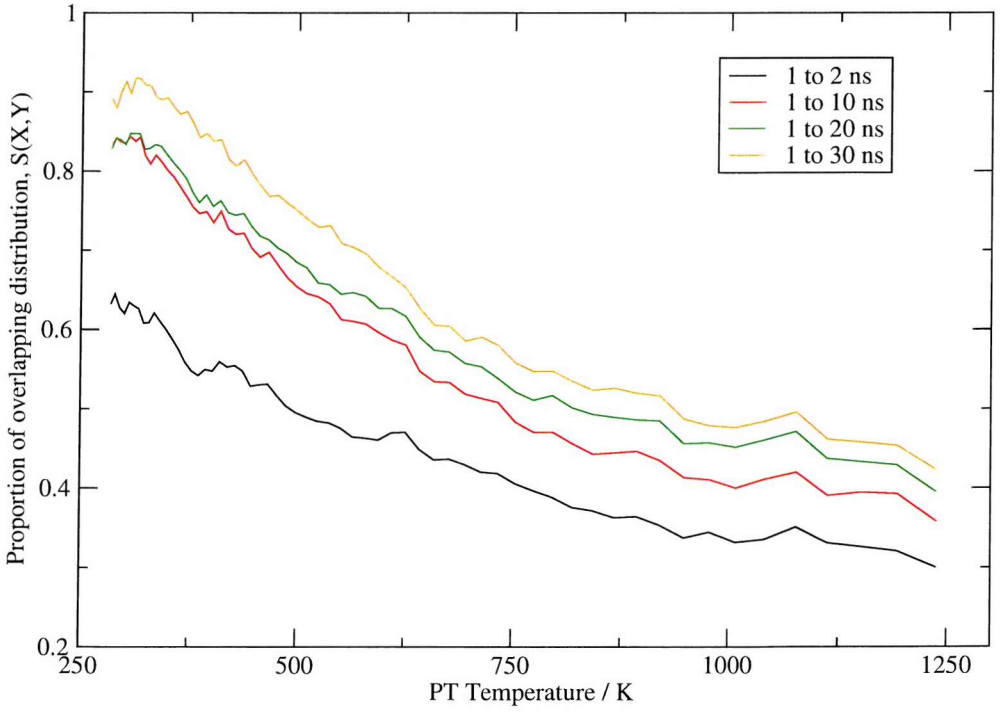



Figure 6.19: The 8D distribution similarity function of MD conformations against those generated by PT. Bins are separated as shown in Table 6.1.

manner the exploration of phase space of one method can be compared to that of another.

$$b(X(i),Y(i)) = \begin{cases} 1 & \text{if } X(i) > 0 \text{ and } Y(i) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$B(X,Y) = \frac{1}{\text{Total number of bins occupied in X}} \sum_{i=1}^{N} b(X(i),Y(i)) \qquad (6.9)$$

Results for the comparison of PT conformational distributions are shown, using the bin with boundaries given in Table 6.1, in Figure 6.20. The dashed lines indicate the proportion of phase space sampled by PT that is also sampled by MD, and the solid line gives the reverse. Sampling by PT covers almost all phase space seen in the MD trajectories. Importantly, a significant proportion ($> 10$ %) of phase space sampled by PT is not seen by 30 ns of MD simulation. Extending the MD simulation from 10 to 30 ns appears not to sample significantly increased phase space, although the distribution of conformers has previously been shown to converge towards that generated by PT. This gives a suggestion of the long timescale that could be required for traditional MD to sample the phase space seen by the PT simulation, even ignoring isomerisation of the peptide bonds.

## 6.8 Analysis of RDFMD conformer distributions

As previously mentioned, the RDFMD results are not expected to produce conformational distributions that compare with those produced by room temperature MD simulation. However, comparison of conformational sampling by RDFMD, to that seen by parallel tempering at a range of temperatures, should give an indication of the validity of generated states.

Results generated using the RDFMD protocol prior to optimisation (referred to as the initial protocol) as described in Chapter 5 is present first. Six simulations were performed with a 1001 coefficient, 0–25 $cm^{-1}$ filter, an internal temperature cap of 2000 K, a filter delay of 20 steps, a maximum of ten filters applied in

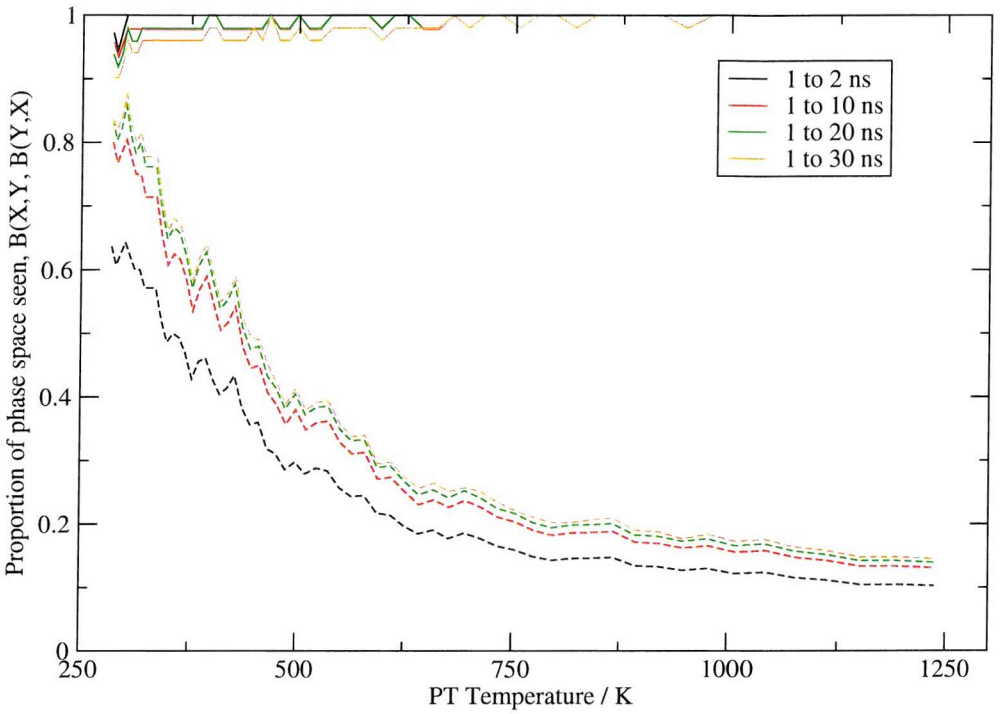



Figure 6.20: The phase space sampling comparison function of conformations from 300 K MD simulation against those generated by PT. Bins are separated as shown in Table 6.1. The dashed lines indicate the proportion of phase space sampled by PT that is also sampled by MD, and the solid line gives the reverse.

one sequence and 10 ps of NPT MD simulation separating filter sequences. A total of 2 ns of NPT simulation was produced, from which conformers are taken for analysis. Four of the six simulations saw infrequent ω transitions, and are discarded. Further details of these simulations can be found in section 5.3 of Chapter 5. The similarity function and phase space sampling function results are shown in Figures 6.21 and 6.22 respectively. The 30 ns NPT MD simulation is included for comparison and bins are bounded by the results shown in Table 6.1.

There is clearly no consistency between the distributions sampled, and many of the conformers sampled by one of the RDFMD simulations are not seen by PT at the maximum temperature (Figure 6.22). The poor sampling is particularly surprising considering the relatively long 10 ps NPT MD simulations run between filter sequences. Results suggest that the RDFMD protocol is generating high energy conformers in which the MD simulation becomes trapped.

Next, results generated using the RDFMD protocol optimised to maximise dihedral angle change are presented. A range of internal temperature caps have

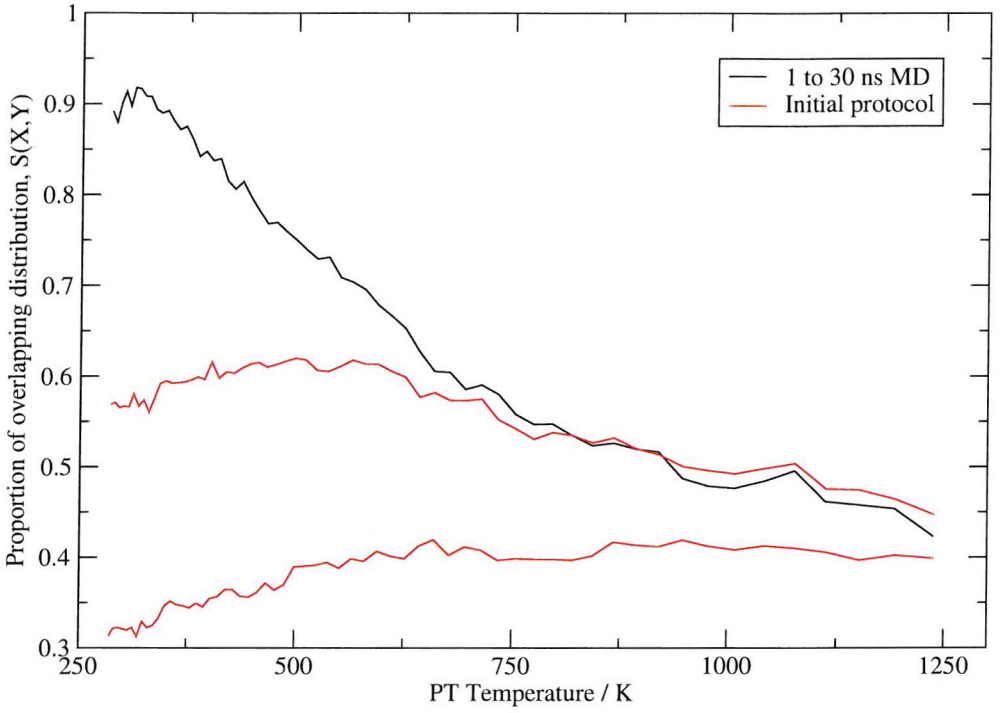



Figure 6.21: The 8D distribution similarity function of RDFMD (using the initial protocol) and MD conformations against those generated by PT. Two simulations using the initial RDFMD protocol are presented. Bins are separated as shown in Table 6.1.

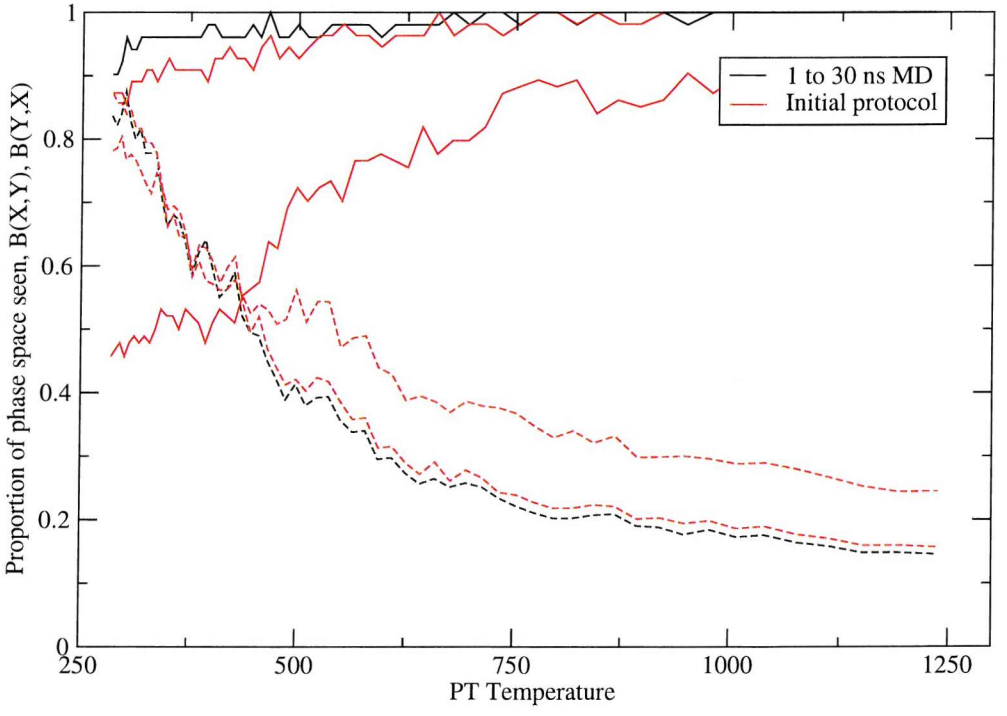



Figure 6.22: The phase space sampling comparison function of RDFMD (using the initial protocol) and MD conformations against those generated by PT. Two simulations using the initial RDFMD protocol are presented. Bins are separated as shown in Table 6.1. The dashed lines indicate the proportion of phase space sampled by PT that is also sampled by RDFMD, and the solid line gives the reverse.

been used and two simulations for each cap have been performed. Simulations shown sample only all-*trans* conformers. A 201 coefficient, 0–200 $cm^{-1}$ filter is used, along with a filter delay of 100 steps, no maximum number of filters per filter sequence and 4 ps of NPT MD simulation between filter sequences. 100 filter sequences are performed, producing a total of 400 ps of NPT MD simulation, from which conformers are sampled. The optimisation procedure suggested a filter delay of either 50 or 100 steps (or a value intermediate to these), and simulations have been duplicated using a filter delay of 50 steps. These results are similar to those shown here and so are not included.

The similarity function analysis of the conformers generated by the optimised RDFMD protocol are shown in Figure 6.23 and the phase space sampling function analysis is shown in Figure 6.24. Results from 30 ns of NPT MD simulation are included for comparison. Increased conformational sampling produced by the RDFMD protocol with an internal temperature cap of 900 K has already been shown in Chapter 5. The conformational distributions generated are clearly far more consistent than those produced with the initial protocol. Deviations of 100 K on the internal temperature cap have little difference on the distributions generated (shown in Figure 6.23), but it would appear that the higher caps generally show an increased proportion of overlap with distributions of higher temperatures. Phase space sampling is clearly increased for higher internal temperature caps, however at 900 K and above, a significant proportion of phase space sampled is not seen by the PT simulation, even at its maximum temperature. Clearly caution must be used with high internal temperature caps.

## 6.9 Further work

The NMR data available for the YPGDV system suggests that the population of *cis*-proline conformers has been over estimated by PT at room temperature. As previously discussed, the NMR data was obtained under acidic conditions for which the aspartic acid residue should be protonated. It is unclear how simulating at neutral pH (with an unprotonated aspartic acid) may have affected results.

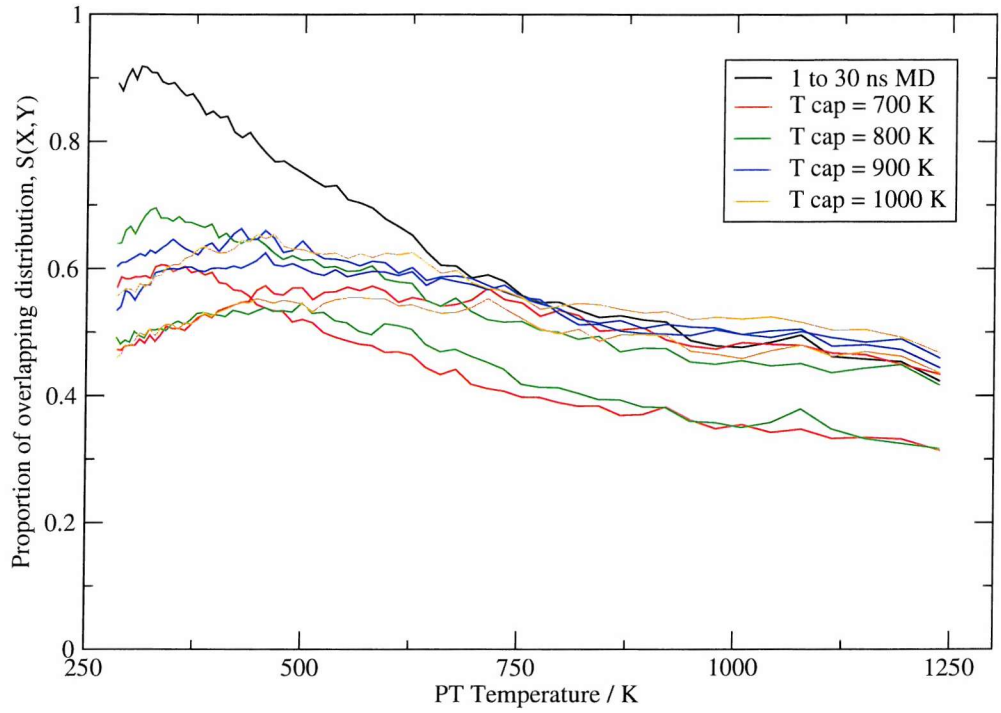



Figure 6.23: The 8D distribution similarity function of RDFMD (using the optimised protocol with a filter delay of 100 steps) and MD conformations against those generated by PT. The internal temperature cap used for RDFMD is shown in the legend. Bins are separated as shown in Table 6.1.

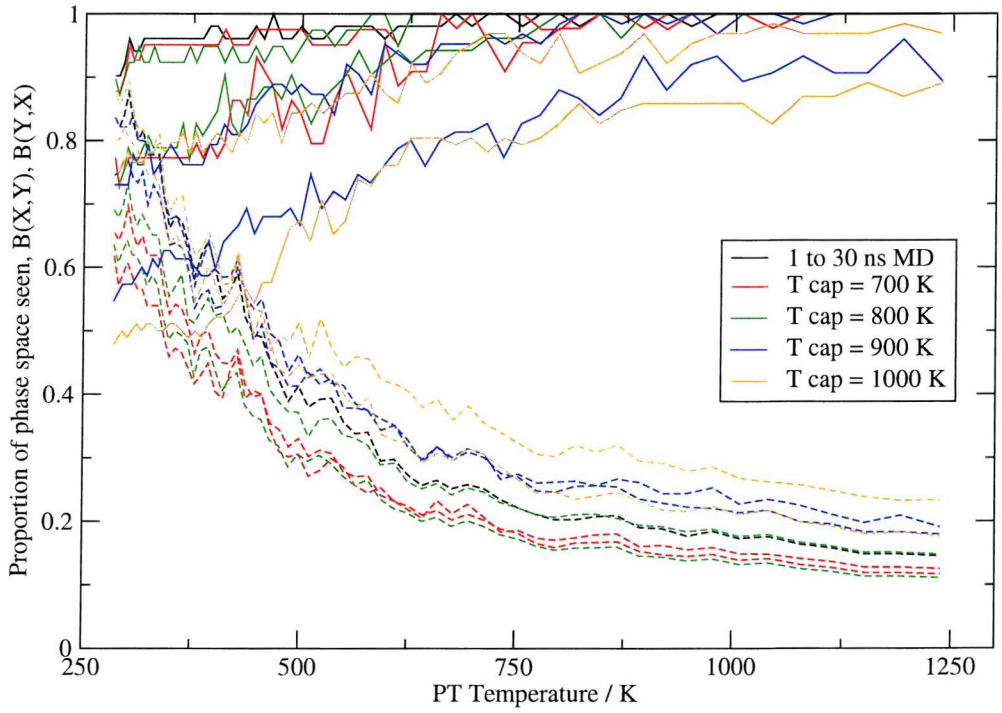



Figure 6.24: The phase space sampling comparison function of RDFMD (using the protocol with a filter delay of 100 steps) and MD conformations against those generated by PT. The internal temperature cap used for RDFMD is shown in the legend. Bins are separated as shown in Table 6.1. The dashed lines indicate the proportion of phase space sampled by PT that is also sampled by RDFMD, and the solid line gives the reverse.

The CHARMM force field has been used for YPGDV simulations, and a parallel tempering investigation has also been prepared using the AMBER force field. The chosen test-case system is the cyclic pentapeptide, cFFAiLP (Phe-Phe-Aib-Leu-Pro), for which information is available from both experimental[114] and simulation studies.[115] A dominant *cis*-proline conformer is expected. The protein contains the proline-phenylalanine-phenylalanine sequence responsible for the cytoprotective ability of antamanide and cyclolinopeptides.[116]

cFFAiLP is a published target of the analytical rebridging Monte Carlo method[115] which used the AMBER force field and an implicit solvent model designed to represent bulk acetonitrile. For this work, an explicit six-point solvent model is used, previously shown to reproduce bulk acetonitrile properties correctly.[117] The protein is solvated in 283 solvent molecules and no neutralising ions are required. Simulations have been performed with the NAMD package using similar parameters to those previously described for YPGDV. The force field version used with the analytical rebridging Monte Carlo method is not specified and thus the most recent prior to the publication, AMBER99, has been selected.

α-methyl-alanine (Aib) is not one of the essential amino acids commonly found in proteins, and is not included with the AMBER force field. To produce a parameter set for the residue, the parameters associated with alanine have been used. The methyl sidechain was duplicated, and the α-carbon's charge was adjusted from 0.0337 to 0.1137 to maintain a neutral residue.

A parallel tempering simulation of cFFAiLP has been set up to a maximum temperature of 1500 K using 41 replicas. The temperature distribution was determined as previously described, targeting an acceptance probability of 0.2. It is hoped that the results will show whether replica mobility is significantly affected by the lower acceptance probability in comparison to 0.4 targeted for YPGDV. Starting cFFAiLP conformations have been produced in a similar manner to that used for YPGDV, beginning with all-*trans* peptide bonds.

Only 2.5 ns of PT simulation has so far been completed on the cFFAiLP system

(a total of just over 100 ns of MD). This is too short a time to expect convergence of *cis-trans* populations. However, the algorithm is producing expected results, with the proportion of *cis*-proline residues steadily increasing at 300 K, presently making up 32 % of conformers (compared with 58 % *cis*-proline reported by the converged analytical rebridging Monte Carlo simulation). The PT simulation is being continued and will be analysed again once 20 ns of PT is complete.

## 6.10 Conclusions

In this chapter the methodology required to perform a parallel tempering simulation has been fully described, including parameterisation of the thermostat used and a technique to chose a temperature distribution that optimises the efficiency of the algorithm. A 20 ns PT simulation of YPGDV has been reported, involving a total of 1.220 $\mu$s of MD simulation.

A secondary structure analysis of the all-*trans* YPGDV conformers generated by PT at 300.0 K shows a dominant $\beta$ turn configuration. The population of this conformer is in closer agreement to experimental data than a 30 ns MD NPT simulation. Isomerisation of the proline peptide bond has been shown to equilibrate quickly at high temperatures, and at lower temperatures, a strong argument that the system is at, or near to, an equilibrium distribution has been made. Non-proline *cis* conformers are shown not to reach room temperature although this may occur during longer simulations.

The challenging target of this investigation- to tackle the quasi-ergodicity of the YPGDV system- has therefore been successfully achieved, sampling nearly 3000 $\omega$ isomerisation events. A higher proportion of *cis*-proline is obtained in comparison to NMR data, although this could be due to the differences between the NMR and simulation conditions. Further work has been started on the cFFAiLP system, for which results from an alternative computational method of sampling proline isomerisation, are available. This work uses the AMBER force field, and the success of the algorithm should be separable from the force field used. After only 2.5 ns of PT simulation, a significant population of *cis*-proline conformers

has already reached 300.0 K.

Two functions have been defined that compare distributions for any number of dimensions, the distribution similarity function, and the phase space sampling comparison function. Use of a conformer set defined using histograms of dihedral angles has been shown to improve the separation of different distributions when compared to a numerical division of each dihedral angle. The PT conformer distributions have then been compared to results from MD, showing that substantially extending the timescale of MD simulation offers only a limited improvement of phase space sampling.

Conformers obtained from RDFMD using different protocols are compared to those generated using PT. The optimised RDFMD protocol described in Chapter 5 produces more consistent results than the initial protocol used early on in this project. A suitable internal temperature cap of 900 K is suggested for the flexible YPGDV system, at which point conformations are obtained that are not sampled by the extensive PT study.

# Chapter 7

# T4 Lysozyme

## 7.1 Introduction

Lysozyme is a member of the glycosidases or glycohydrolases- enzymes that catalyse the transfer of water to a glycosyl group by accelerating the normally slow cleavage of a glycosidic C-O bond.[118] T4 lysozyme is a relatively small protein with 164 residues arranged into two domains; the N-terminal domain (residues 13 to 75) and the C-terminal domain (residues 76 to 164 and 1 to 12).[119] The domains are separated by a deep cavity that is bounded by an α-helix that runs almost the entire length of the protein, and by strong polar interactions between residue side chains. The structure of the enzyme was first determined in 1974 using X-ray crystallography,[119] and one important observation was the inaccessibility of residues known to be important for catalytic activity. The publication concludes: "It seems certain that in order to allow the substrate to enter, the enzyme would have to undergo a fairly substantial conformational change." The nature of this change has been the subject of a number of experimental[52,119–131] and theoretical studies.[45,60,132–136]

In this chapter an investigation of the conformational behaviour of T4 lysozyme is performed, beginning with a summary of previously reported studies. The results of long molecular dynamics and simulations at raised temperatures are presented, and the computational methods developed in this thesis are applied.

### 7.1.1 Experimental studies of T4 lysozyme

The small size of T4 lysozyme has meant it has been a popular system for experimental methods seeking to analyse an enzyme system.[120] It is also a common target for studies of protein mutants, investigating, for example, the effect on the stability of the molecule by changing the hydrophobicity of key residues,[121] or by reducing the overall charge at neutral pH from +9 to +1.[122] Residue substitutions have also shown the catalytic importance of aspartic acid 20, glutamic acids 22, 105 and 141, tryptophan 138 and asparagine 140.[119] These residues all lie around the edge of the cavity, as shown in Figure 7.1, and if any one of these is removed, the efficiency of catalysis is heavily reduced.

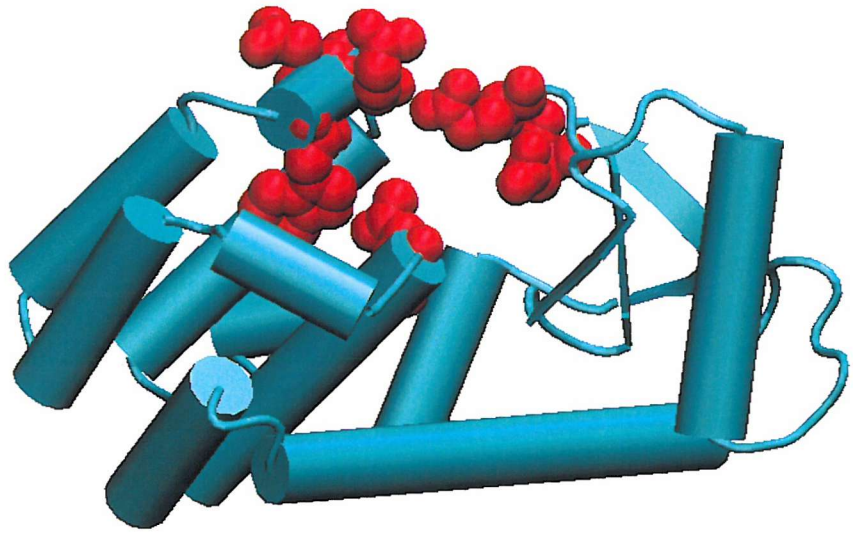


Figure 7.1: Cartoon representation of pdb structure 3LZM. The van der Waals radii of catalytic residues are shown in red.

A significant set of studies has involved the analysis of histidine 30 and aspartic acid 71. In the unfolded protein, the histidine pKa is 6.3,[123] suggesting a majority of unprotonated states at neutral pH. However, experimental work has shown that in the folded T4 lysozyme, the pKa of histidine is shifted close to 9.1.[124] Such a large shift in pKa is rarely seen, and has been attributed to the formation of a salt bridge with aspartic acid 71. The pKa of the aspartic acid is also shifted from a solution

value of 3.5–4.0, to a value less than 1.[124] Mutant studies that substituted either of the residues that form the salt bridge show a destabilisation of the folded protein of around 3 kcal $mol^{-1}$.[124,125] The salt bridge is located in an important region of T4 lysozyme, at the base of the active site. One side is solvent accessible and the other side is placed against a rather rigid network of amino acid residues that begin to form the hydrophobic core of the molecule. The importance of this salt bridge will be returned to in the discussion of theoretical studies performed on this protein.

A review of 200 Protein Databank[137] structures of T4 lysozyme and its mutants showed 25 crystal forms that differ in the rotational state of the two domains, and the accessibility of the protein cavity.[52] These are described in terms of a hinge bending angle, which differs by over 50 degrees across the 25 crystal forms. Although wild-type T4 lysozyme (hereafter referred to as WT) only crystallises in one of the more closed conformations, the study concludes that a hinge bending motion is intrinsic to the T4 lysozyme system, and the conformation of the crystal structures differ due to crystal packing forces. This conclusion seems reasonable considering, for example, the M6I mutant, which crystallises into four distinct conformers (typically labelled by the chain identifier: a, b, c, or d), one of which (M6Ia) is more closed than the WT protein, and another of which is significantly more open (M6Id).

A study involving the assignment of $^{15}N$, $^{13}C$ and $^{1}H$ resonances in solutions of pH 5.3 to 5.7 using a range of NMR techniques specifically searched for evidence of a hinge bending motion.[126] However, no such evidence was detected through either the doubling or broadening of resonances, and the work concluded that either the motion does not exist in solution, or that it is so rapid that line-broadening effects are not caused by it. Results are in agreement to an earlier solution NMR study,[127] which saw no evidence of a difference between the solution and crystalline states. However, an NMR study which looked at $^{15}N$-$^{1}HN$ residual dipolar couplings concluded that the average solution conformer was more open than that obtained using X-ray crystallography.[53] Crystal packing forces were suggested to be responsible for the closed crystal structure.

Solution NMR of T4 lysozyme in the presence of a substrate showed that a significant large-scale movement occurs upon substrate binding.[128] Residues on either side of the cavity opening were spin labelled and analysis of spin-spin interactions clearly gave significant differences between the bound and free conformations. The free solution structure was again determined to be more open than the crystal structure. More evidence of an opening event upon substrate binding was gained by use of single-molecule spectroscopy of T4 lysozyme during the hydrolysis of polysaccharide chains.[129] A donor-acceptor pair of dye molecules were attached to non interfering sites of T4L and the intramolecular fluorescence resonance energy between the dye molecules was measured. A change of the intensity of fluorescence was determined on the same timescale as enzyme turnover (a few milliseconds), and this was attributed to an opening event that altered the distance between the dye molecules. In the absence of substrate, no opening event was observed.

The L99A mutant of T4 lysozyme, with alanine substituted for leucine, is known to bind small hydrophobic ligands such as substituted benzenes. A study using methyl relaxation dispersion NMR concluded that L99A exists predominantly in a closed form that known ligands cannot bind to, with an open, ligand-receptive state that is 2.0 kcal $mol^{-1}$ higher in free energy.[130,131] The population of this state was estimated to be less than a few percent. Studies concluded that this low percentage was likely to be the reason why previous work had not seen evidence for the opening event in the absence of substrates.

### 7.1.2 Theoretical evidence for an opening event of T4 Lysozyme

Again, the small size of T4 lysozyme has made it a common target for investigation by theoretical methods. Also the large number of X-ray structures make it a suitable system for essential dynamics and domain analysis. The set of X-ray structures was an early target of the DynDom (dynamical domains) package, which locates rigid domains, and analyses the motions required to

move between structures.[45] For T4 lysozyme and its mutants, this motion is described as a 'door closing motion,' in other words; it consists of two major domains with connecting regions forming a hinge axis. Residues 11-13 are identified as important for such a motion to occur.[60] However, the considerable differences between X-ray structures and solution conformers proposed by experimental techniques should limit the conclusions drawn from the X-ray structure set.

Arnold *et al.* reported three molecular dynamics simulations of 100 ps in 1994, of WT, the most open of the M6I mutants (M6Id) and of the most closed of the M6I mutants (M6Ia).[132] These simulations were extended to 500 ps and reanalysed in 1997.[133] Nine chloride counter ions were added to neutralise each system, countering the +9 charge determined by the program Network.[138] The CVFF force field[139] was used and the proteins were solvated using a water-droplet model with a 10 Å layer of solvent. This solvent model was used to reduce the computational expense of simulation in comparison to that of a periodic system, and 2 % of water is seen to evaporate over the simulation timescale. Each simulation reported a significant closing motion, with rising RMSD values as time progressed, and all structures converging to a similar conformation after 500 ps. The most extreme change was that of the M6Id structure which closed by over 50 degrees, while the WT simulation closed by 20 degrees. This is in complete disagreement with experimental data, which suggests a more open solution conformer for both WT and M6I.

In 1997, coarse-grain simulations were reported by Bahar *et al.*[134] that used low resolution MC in which only torsional motion between α-carbon atoms and motion of rigid side chains were allowed. A hinge bending motion was observed, producing strongly correlated intra-domain fluctuations and anti-correlated inter-domain motions. Configurations fluctuated about the starting conformation, and there was no clear evidence of an average conformer that was more or less open than the starting structure.

In 1998, 1 ns MD simulations of WT, M6Id and WT starting from the M6Id conformation (hereafter referred to as WTd) were published by de Groot *et al.*[135]

The GROMOS[18] force field was used and simulations were performed with periodic boundary conditions, and explicit SPC water molecules. Unfortunately, an overall protein charge of +8 was reported, and 8 chloride ions were added to the simulation. Histidine 31 was simulated in its unprotonated state (confirmed by personal communication with authors); an error by comparison to the experimental studies of the histidine-aspartic acid salt bridge. The effects of this are impossible to determine, but removing the salt bridge is likely to destabilise the protein to some extent. Simulations are seen to over-estimate a twisting motion between domains in comparison to an essential dynamics study of X-ray structures. In disagreement to the findings of Arnold *et al.*, conformations more open than the WT crystal structure are preferentially sampled by simulation. This inconsistency between the results of the two studies is suggested to be due to protocol differences, and the water-droplet solvation model is considered to be inadequate. de Groot *et al.* conclude that the hinge bending properties of WT and M6I are similar in solution. The work does not support evidence of the two state model suggested experimentally for T4 lysozyme, with a dominant closed conformer and a small population of a more open state.

A study published in 2003 by Zhang *et al.* performed WT simulations using MD and an amplified collective motions (ACM) method.[136] In this method, velocities that correspond to the slowest collective motions, as determined by the anisotropic network model (ANM), are coupled to a higher temperature heat bath than the rest of the system. The ANM, as implemented by Zhang *et al.*, links the centre of mass of residues using harmonic potentials with force constants of 1, 0.001 or 0, depending on the inter-residue distance. A single configuration from simulation is then used to generate eigenvectors described by this coarse-potential model, which are periodically recalculated as simulation evolves. Residues 1 to 162 of T4 lysozyme were solvated using the GROMOS force field with SPC water molecules, and 3 ns of production was performed for both conventional MD and ACM simulations. Similar to the work of de Groot *et al.*, a +8 protein charge is reported, presumably also corresponding to an unprotonated histidine 31. Simulations were projected onto the two main eigenvectors determined by

essential dynamics analysis of X-ray T4 lysozyme and associated mutant structures. Molecular dynamics simulations sampled motions within the conformational space seen by X-ray structures but the ACM simulation sees increased sampling that exceeds the X-ray subset in both of the two main eigenvectors.

The validity of ACM results rely on the calculation of the slow collective modes by the ANM, from static simulation structures. The method of amplification is not linked to motions produced by simulation, as with RDFMD, but by motions produced by the description of harmonically-linked residue mass centres. Zhang *et al.* discuss the arbitrarily chosen number of modes that are coupled to the heat bath as a limitation of the method. Neither ACM or MD results show evidence for discrete hinge angle distributions that would be required to reproduce experimental results.

### 7.1.3 Summary

Experimental evidence is divided on whether T4 lysozyme in solution undergoes an opening event in the absence of substrate, although most of the more recent studies indicate that the solution structure is more open than that seen with X-ray crystallography. The suggestion that a very small proportion of an excited state could have been missed by many methods seems reasonable, and there is evidence for a ligand-accessible solution state occurring in only a few percent of the conformer population.

Theoretical studies include detailed analysis of the set of structures obtained by X-ray crystallography of T4 lysozyme and its mutants, suggesting a hinge motion between the two domains. Molecular dynamics simulations show conflicting motions, but the timescale sampled is identified as an issue in every case. No simulations support the hypothesis of an equilibrium between closed and open states.

## 7.2 Molecular dynamics

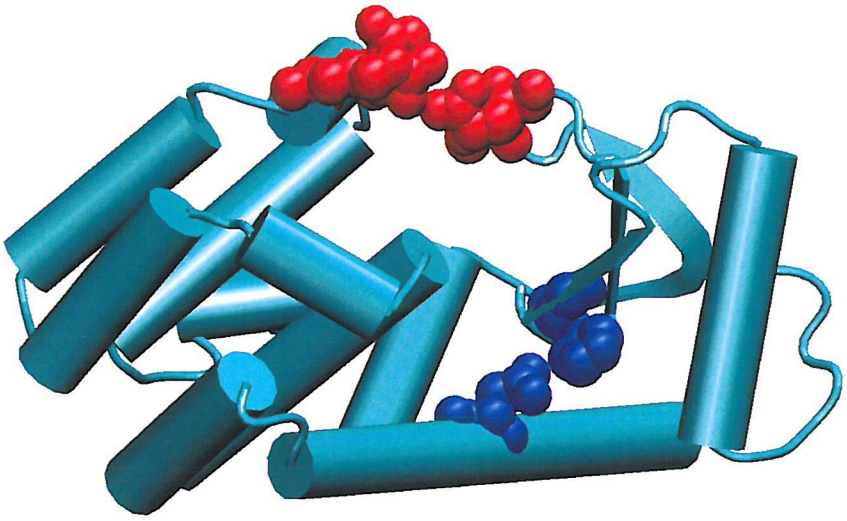
The investigation of T4 lysozyme follows the work of de Groot *et al.*, with 20 ns of NPT simulation from WT, M6Id and WTd. The pdb structures 3LZM[140] (WT) and 150Ld[141] (M6Id) are used as starting points, and these are shown in Figure 7.2. Simulation lengths are therefore an order of magnitude longer than those previously reported on this system. Simulations are described by secondary structure analysis, the trajectory of the protein hinge angle, RMSD analysis, and by several key inter-atomic distances.

As computational power becomes more readily available, it is increasingly common in literature to report duplicated simulations starting from identical structures, but using different random velocity seeds. Results are frequently very different, and some statistical confidence can be gained regarding events that occur. This has been attempted for each system presented here, changing the random seed at the beginning of the heating stage of equilibration so that a different set of velocities are assigned to the protein after minimisation. Completion of the second simulations are beyond the timescale of this project, and the 20 ns simulations are accompanied by a 10 ns duplication. The 10 ns simulations are being extended, and shall be reanalysed once they reach 20 ns. In each case the 20 ns simulation is referred to as the first simulation, and the 10 ns simulation, as the second.

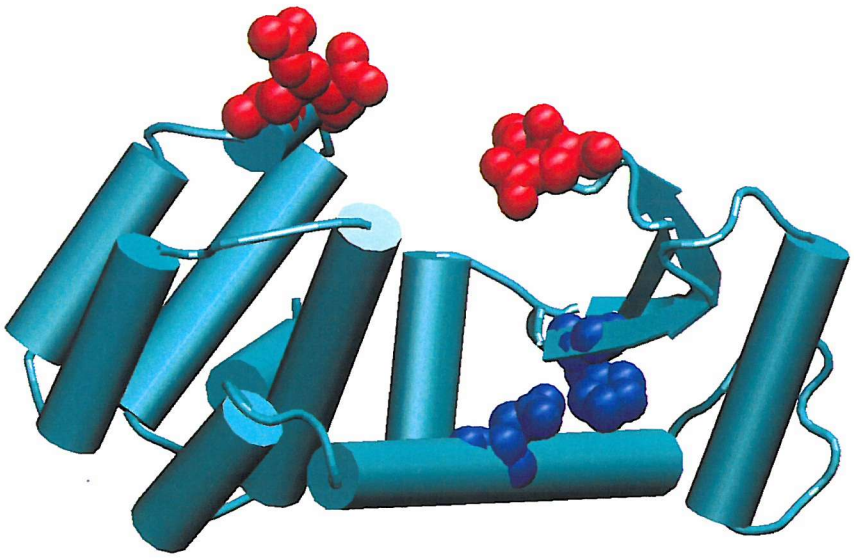
### 7.2.1 Computational details

The WT system was setup from pdb entry 3LZM with polar hydrogen atoms placed by WHAT IF[142] to maximise the hydrogen bonding network. WHAT IF was also used to check the pdb structure and assign protonation states. X-ray waters were included in the system which was then solvated using the XLEAP utility provided with the AMBER package. The +9 charge of the protein was neutralised by chloride ions placed by XLEAP.

Unless otherwise stated, all simulations in this chapter were performed using the NAMD package with cuboid periodic boundary conditions, a particle mesh



(a) 3LZM.



(b) 150Ld.

Figure 7.2: Cartoon representation of pdb structures 3LZM and 150Ld. The van der Waal radii of bridging residues are shown in red, and the histidine-aspartic acid salt bridge is shown in blue.

Ewald treatment of electrostatics (with an interpolation order of 6) and a switching function applied to Lennard-Jones interactions between 9 Å and the 11 Å cutoff. SHAKE is applied to all bonds containing a hydrogen atom, with a tolerance of $10^{-8}$ Å. A Langevin thermostat (with an associated Langevin damping parameter) is used to control temperature and a Nosé–Hoover Langevin barostat (with associated piston period and piston decay parameters) is applied to control the pressure.

The solvated system was minimised to reduce any problems with initial atomic placements. The water was minimised for 30 000 steps, followed by the ions for 5 000 steps, the water and ions for 20 000 steps, the protein for 20 000 steps, and finally the entire system for 40 000 steps.

The minimised system was gradually heated at 50 K intervals from 50 K to 300 K, with 20 000 steps of NVT MD simulation at each temperature using a 2 fs timestep (i.e. 40 ps of simulation) and a Langevin damping parameter of 10 $ps^{-1}$. At 300 K an extra 20 000 steps of NVT simulation with the same parameter set was then performed.

NPT MD simulation was then used to adjust the pressure to the desired 1 atm target. 60 000 NPT steps were performed with a piston period of 200 fs, a piston decay of 100 fs and a Langevin damping parameter of 10 $ps^{-1}$. 60 000 NPT steps were then run using a piston period of 400 fs, a piston decay of 200 fs and a Langevin damping parameter of 5 $ps^{-1}$. Finally 200 000 steps of NPT simulation with the production parameters (a 400 fs piston period, a 200 ps piston decay and a Langevin damping parameter of 1 $ps^{-1}$) was performed.

The resulting WT system consisted of 2644 protein atoms, 9955 water molecules and 9 chloride counter ions, giving a total of 32518 atoms. The equilibrated simulation cell had the dimensions 75.89, 60.64 and 69.33 Å.

The M6Id system was prepared in an identical fashion to that of WT, starting from chain 'd' of the pdb structure 150L (150Ld). The resulting system contained 2646 protein atoms, 8486 water molecules and 9 chloride ions, and had cell

dimensions of 63.40, 74.42 and 58.23 Å.

To produce the WTd system, the side chain of methionine 6 was deleted and replaced by that of isoleucine using WHAT IF prior to addition of polar hydrogen atoms. Minimisation and equilibration were otherwise identical to the WT and M6I systems. The resulting system contained 2644 protein atoms, 8486 protein atoms and 9 chloride counter ions, and had cell dimensions of 63.33, 74.33 and 58.16 Å.

The second simulations of each system, are produced in the same manner as that described above, using a different random number seed to assign initial velocities at the heating stage. Cell dimensions for these simulations are similar to those reported above.

### 7.2.2 Molecular dynamics results

DynDom has been used to generate the trajectories of the protein hinge angles for each simulation. Snapshots sampled every 1 ps have been compared to either the 3LZM or 150Ld pdb structures using all default parameters for DynDom (a maximum of 20 clusters, 100 iterations for the clustering routine, a window length of 5 residues, a domain size of at least 10 % of residues and a minimum ratio of external to internal displacement of 1.0). For DynDom to locate domains between two structures, they must be sufficiently different, and so simulations beginning from 3LZM are compared to 150Ld, and vice versa. In each case the hinge angle reported is therefore relative to the hinge position of the structure with which the system is compared. Although any domains located are equally valid, a trajectory of hinge angles can only be presented if each angle is calculated from two similar domains. Therefore the hinge angle is only reported if the two domains each contain more than 80 or 90 % (referred to as the domain cut off) of the residues determined from an analysis of the 3LZM and 150Ld crystal structures.

An RMSD analysis is also presented for each simulation, with superposition over the N-terminal domain and the RMSD of the C-terminal domain measured. This is performed in comparison to both the 3LZM and 150Ld structures.

The 3LZM crystal structure shows two pairs of residues in contact at the top of the protein cavity: arginine 137 forms a salt bridge with glutamic acid 22, and polar interactions link threonine 21 and glutamine 141. These are not in contact in the 150Ld structure, and are clearly important to judge the accessibility of the cavity for substrate binding. The α-carbon distances between these pairs are reported, as are the inter-atomic distances between key side chain atoms that can be used to determine whether the residues are in contact. These residues shall hereafter be referred to as the bridging residues.

Secondary structure analysis has been performed using the DSSP algorithm[106] with default options. It is worth noting that the salt bridge previously mentioned between protonated histidine 30 and aspartic acid 71 has been monitored throughout all simulations and is always present.

**T4 lysozyme wild-type (WT)**

Analysis of the 20 ns NPT simulation of the WT system is presented in Figure 7.3. Hinge angles shown have been produced using DynDom with a domain cut off of 90 %. The molecular hinge is generally around 30 degrees from the open 150Ld structure, suggesting a similar conformation to that seen in the 3LZM crystal form (which is 31 degrees more closed than the 150Ld structure). Three events of particular interest occur corresponding to the separation of the bridging residues at the top of the protein cavity.

The first event takes place between 3 to 4 ns of simulation (already beyond the timescale of previously reported simulations) and involves an opening motion of over 15 degrees. This follows the breaking of the polar interactions of bridging residues, and an increase in the α-carbon distance between threonine and glutamine of over 10 Å. The RMSD in comparison to the open structure reduces, and against the closed structure increases, until the protein is more similar to the open structure of 150Ld. At 2.83 ns, the most open conformation sampled over the 20 ns trajectory is seen, and this is shown in Figure 7.4. Following the opening event, the

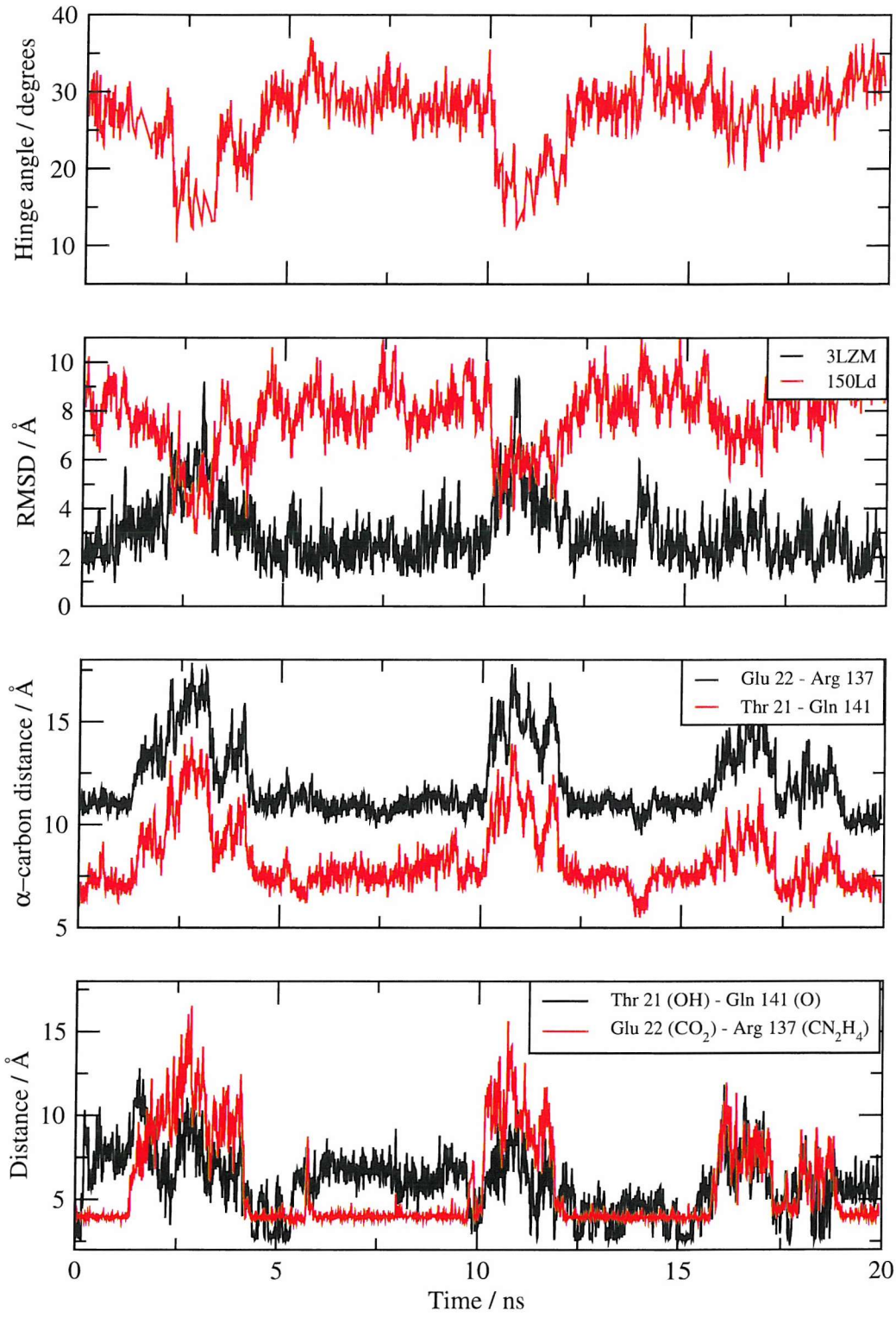



Figure 7.3: Analysis of the 20 ns NPT MD simulation of WT. Top: Hinge angle when compared to structure 150Ld using DynDom and a domain cutoff of 90 %. Upper middle: RMSD of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower middle: α-carbon distances between bridging residues. Lower: Inter-atomic distances used to determine contact between bridging residues.

protein closes back to its original conformation.

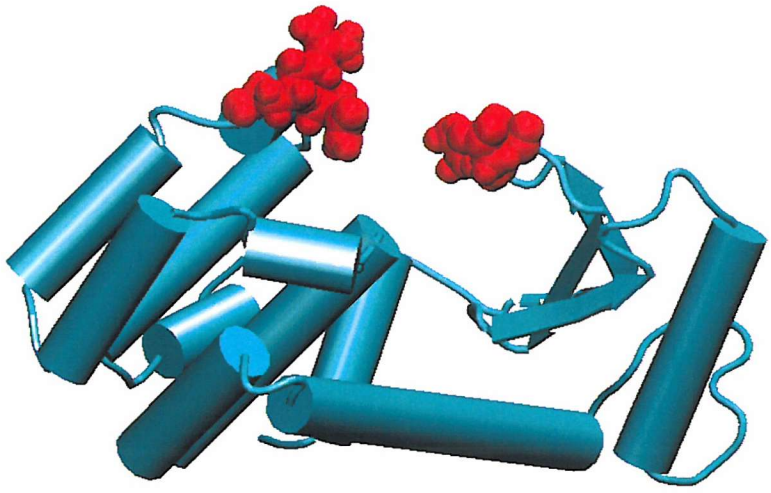


Figure 7.4: The conformation most similar to the 150Ld structure obtained during the 20 ns MD simulation of WT T4 lysozyme. The representation shown is as described in Figure 7.2.

After 10 ns of simulation, a second event occurs, similar to the first, but initiated over a shorter period of time. Here the bridging residues move apart at the same time as the protein hinge dramatically opens, again by over 15 degrees. In the first event, the bridging residues are seen to begin to move apart prior to the hinge angle change. Again, the RMSDs show a transition to a conformation more like that of the M6Id structure. It is interesting to note that the conformational change exhibits the same magnitude of motion for each opening event. This suggests possible discrete conformers previously not reported by MD studies.

A third event takes place between 15.5 and 17 ns of simulation, once again showing a separation of the bridging residues. This time however, the separation is of a much lower magnitude, and the change in RMSDs and hinge angle do not show a large-scale conformational change. This implies a lesser event, and suggests that the separation of the bridging residues is not sufficient to cause the conformational change to a more open structure.

The secondary structure analysis of the 20 ns simulation is shown in Figure 7.5, showing no significant changes to the secondary structure reported in the 3LZM crystal structure. The fact that conformational change is not associated with a secondary structure change is in agreement with experimental results.

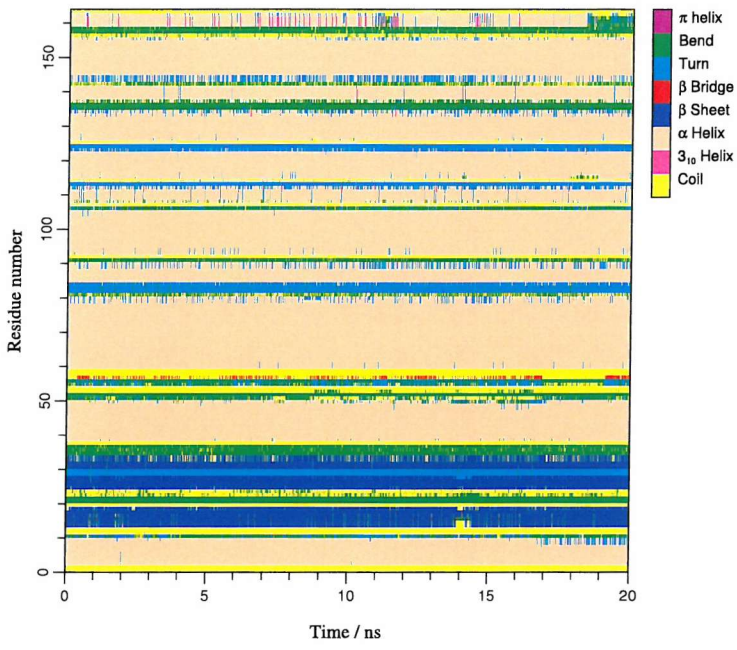
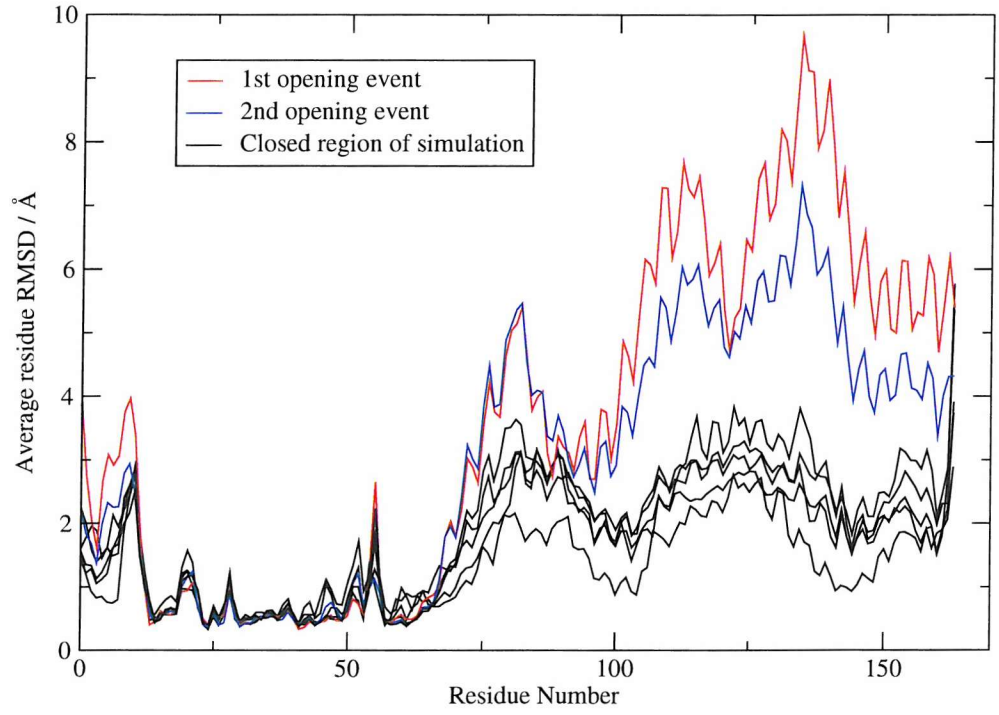



Figure 7.5: Secondary structure analysis of the 20 ns NPT MD simulation of the WT system.

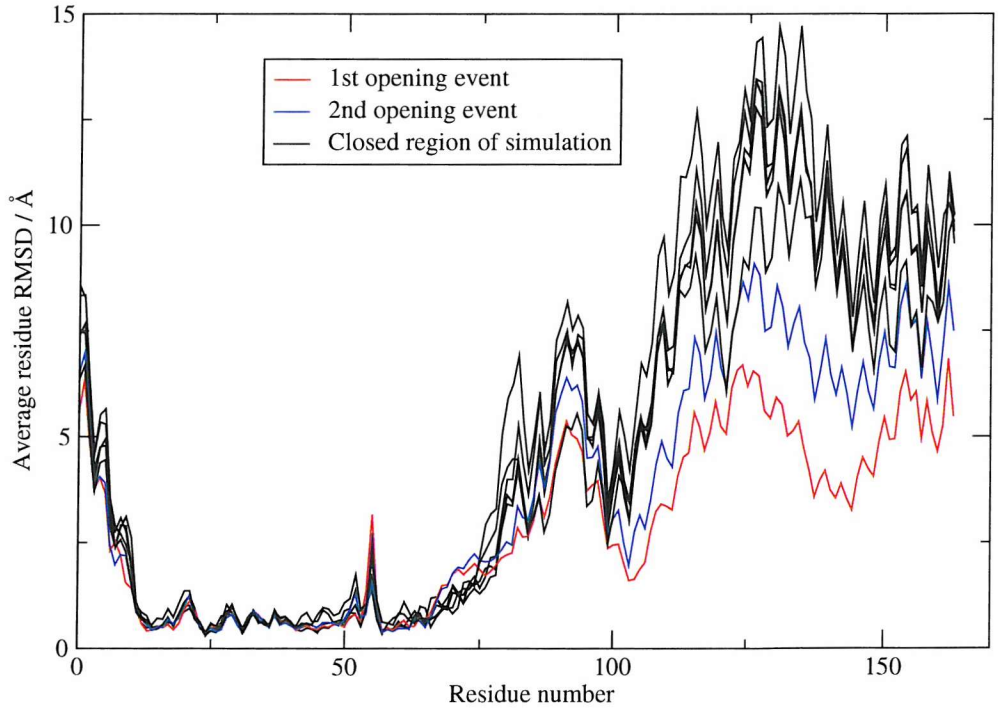
At this stage it is possible to discuss the nature of the opening events seen in the 20 ns NPT MD simulation of the WT system. The 20 ns simulation has therefore been split into blocks, each containing either one of the two domain opening events, or a maximum of 2.5 ns of simulation in which only the closed conformation was sampled.

Figure 7.6 shows the average RMSD of residues over these blocks of simulation. Very little differentiation is seen between regions of simulation that sample a closed conformation. However, a significant difference in the C-terminal domain is clear for both the major opening events. RMSDs of residues in the C-terminal domain are increased against the closed 3LZM pdb structure, and are reduced against the M6I structure. The detail that provides the RMSD shift in Figure 7.3 can therefore be seen.

It is also possible to locate the backbone torsions responsible for the conformational changes. Figure 7.7 shows the RMSD of $\psi$ and $\phi$ angles for each residue, taken against average values from the 20 ns simulation. The red blocks




(a) RMSD against the 3LZM pdb structure




(b) RMSD against the 150Ld pdb structure

Figure 7.6: Average residue RMSD analysis over periods of interest during the 20 ns NPT WT simulation. Superposition is performed on the N-terminal domain (residues 13 to 63).

indicate residues that link secondary structure units, and it is in these that the opening events occur. In particular, residues 11 to 13 show large changes in the backbone torsions, as suggested by the DynDom analysis of T4 lysozyme crystal structures.[135] Interestingly, there are significant differences between the two opening events, with the first event inducing larger changes in residues 11 to 13 and 90 to 93, and the second event inducing more significant rearrangements in residues 53 to 54. This is of particular interest to the application of RDFMD to the system, as there is no single set of backbone torsions responsible for the opening motion.

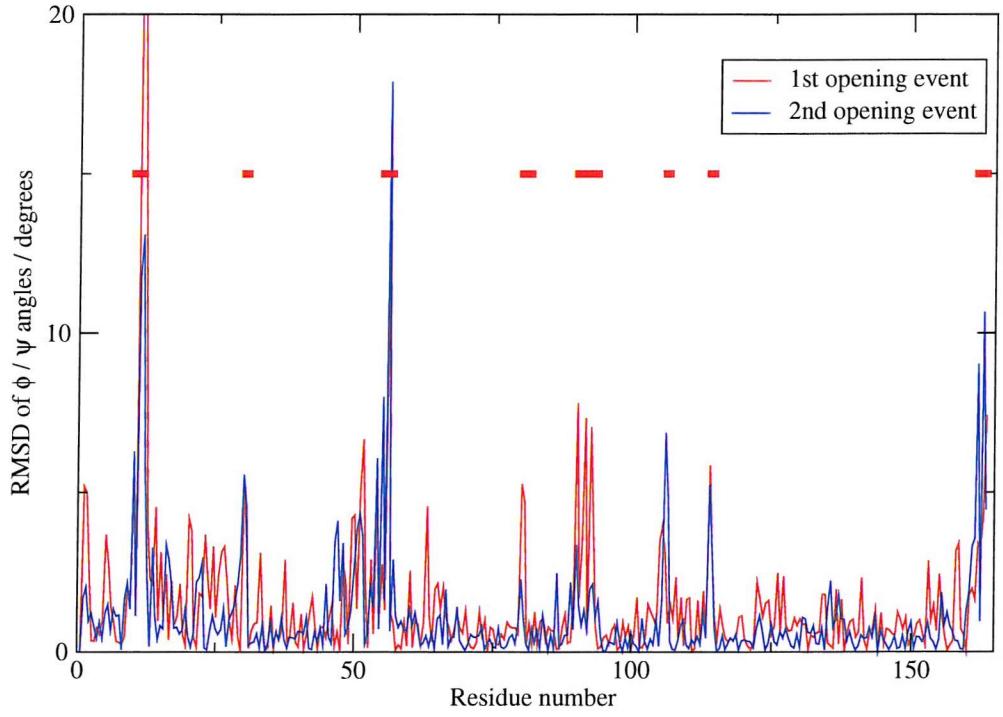



Figure 7.7: RMSD of ψ / ϕ angles against average simulation values for the 20 ns NPT simulation of the WT system. Red blocks indicate loop regions between secondary structure elements that observe significant motion.

The 10 ns WT simulation, shown in Figure 7.8 begins after equilibration to a similar structure to that seen in the first simulation, however the RMSD against the 3LZM crystal form is slightly increased. This could be caused by one of the bridging residue pairs (threonine and glutamine) not being in contact at the start of the simulation.

An 80 % domain cutoff is used to accept DynDom hinge angles, and three

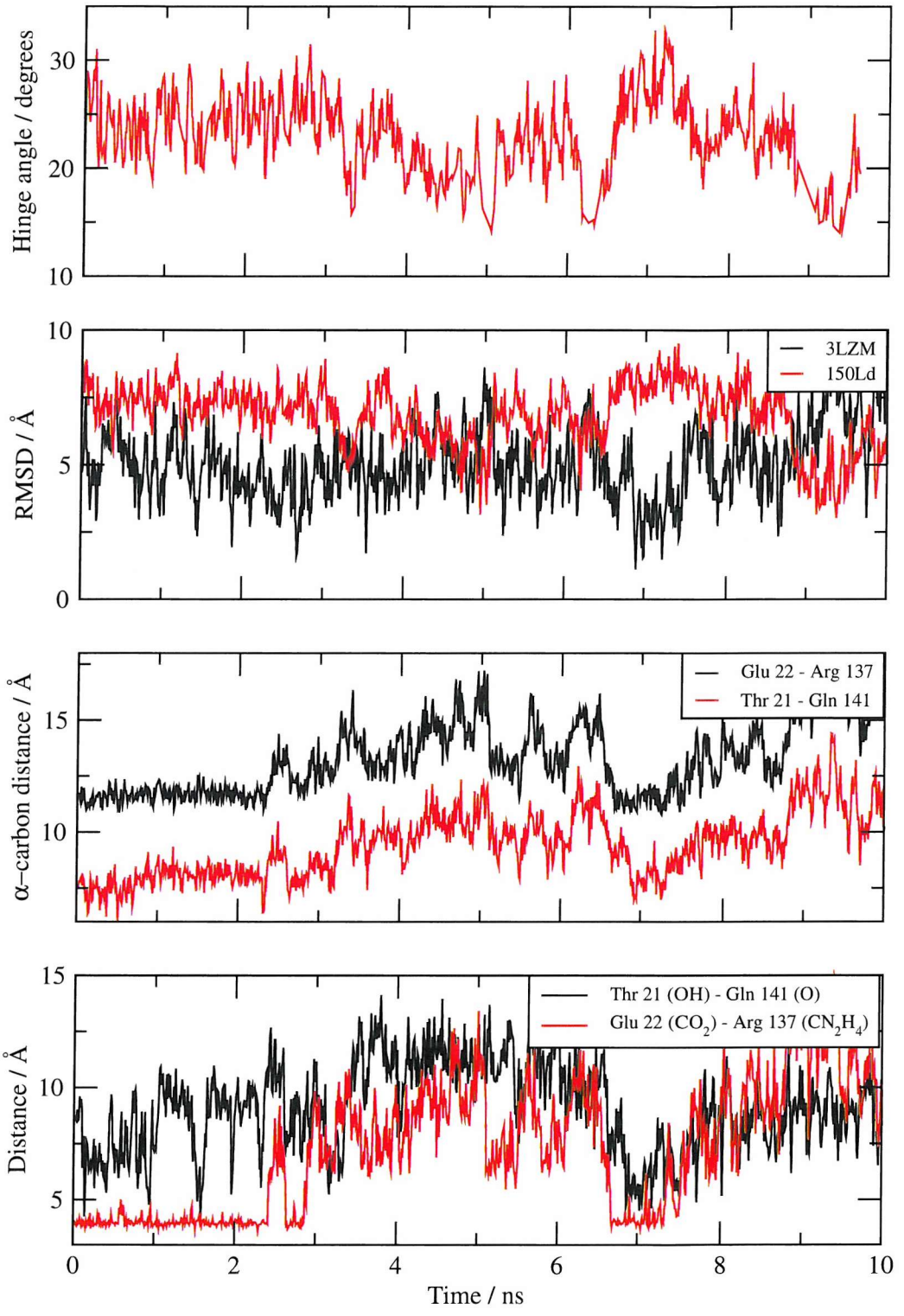



Figure 7.8: Analysis of the 10 ns NPT MD simulation of WT. Top: Hinge angle when compared to structure 150Ld using DynDom. Upper middle: RMSD of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower middle: α-carbon distances between bridging residues. Lower: Inter-atomic distances used to determine contact between bridging residues.

events occur for which the relative hinge angle is decreased (i.e. the hinge angle becomes more like that seen in the 150Ld structure). The first two of these are short lived, at 4.5 ns and 6.2 ns, and the third occurs after 8.8 ns, and continues to the end of the simulation. For each event, hinge angles are reported by domains outside the cutoff and the angle trajectory is therefore discontinuous. However, the opening events are well recorded by the overlap of the RMSDs, with the conformation being more similar to that of the 150Ld structure than the 3LZM pdb. The simulation again gives evidence of a distinct, stable, closed conformer, however this is less clear than in the 20 ns simulation.

The secondary structure analysis of the 10 ns simulation is very similar to that shown in Figure 7.5 and is therefore not shown.

### Most open conformation of the M6I mutation (M6Id)

The results of a 20 ns NPT simulation starting from the most open structure of the M6I mutant (pdb entry 150Ld) are presented in Figure 7.9. The hinge bending angle has been calculated using an 80 % domain cutoff, and structures are compared against the closed 3LZM pdb. The simulation shows an open conformer, with one significant closing event at 3.6 ns. Here, the hinge angle reported by DynDom falls outside the domain cutoff, and is not shown. The distance between the bridging residues decreases, as does the RMSD of the C-terminal domain compared to the 3LZM structure. The structure quickly reopens, and no permanent closing motion is seen as in the simulations of de Groot *et al.* There is no connection of the bridging residues, although this could require simulation over a longer timescale.

The secondary structure analysis of the 20 ns NPT simulation of M6Id shows several important difference to that of WT. A short helix close to the C-terminus is no longer found, and the main helix that links the two domains is now split in two, with a turn introduced in residue 70. This split is spatially close to the mutation site and the separated helices are slightly bent in relation to each other in a manner that compliments an open conformer. The stability of the broken helix suggests a barrier to closure that is not encountered in the WT system.

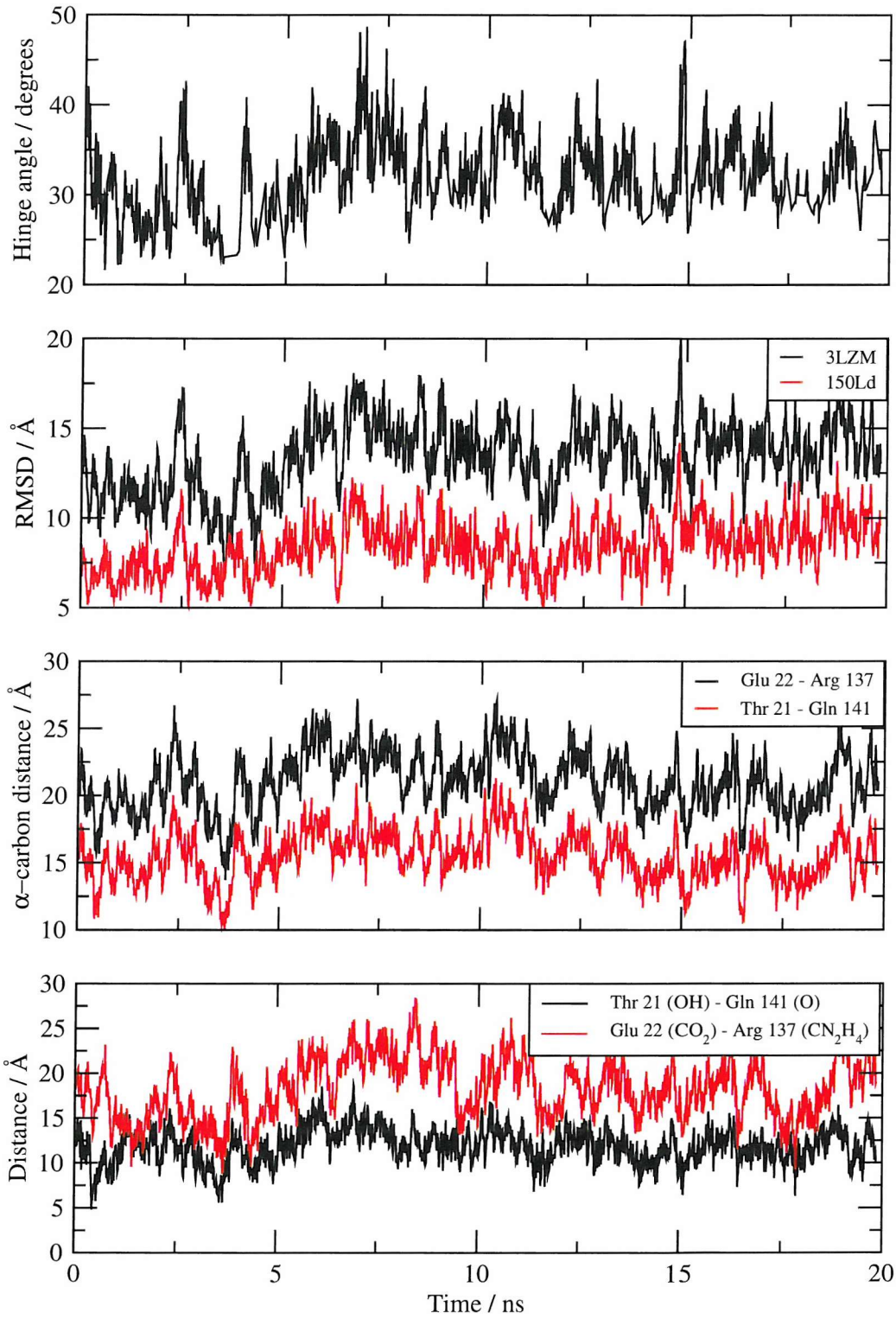



Figure 7.9: Analysis of the 20 ns NPT MD simulation of M6Id. Top: Hinge angle when compared to structure 3LZM using DynDom. Upper middle: RMSD of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower middle: α-carbon distances between bridging residues. Lower: Inter-atomic distances used to determine contact between bridging residues.

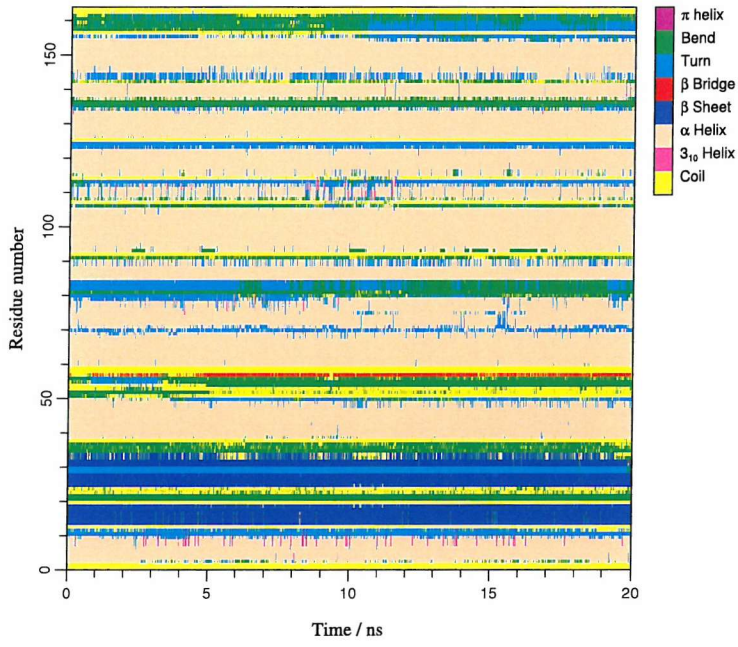



Figure 7.10: Secondary structure analysis of the 20 ns NPT MD simulation of M6Id.

The 10 ns simulation of M6Id samples several events that move to a more closed conformation, as shown in Figure 7.11. Between 2.1 and 2.4 ns, the hinge angle (described by an 80 % domain cutoff) decreases by approximately 10 degrees, bringing the side chains of threonine and glutamine into close proximity. The hinge angle at the extreme of the opening event is not described within the domain cutoff, and the degree of closure is greater than 10 degrees. The RMSDs briefly show a structure closer to the closed form before the hinge residues part and the protein returns to its open conformation. Similar short-lived events occurs at 4 ns and 6.2 ns.

Once again the secondary structure analysis of the 10 ns simulation are similar to that of the first simulation, and results are not shown.

### Wild-type T4 lysozyme starting from an open structure (WTd)

The results of the 20 ns NPT simulation are shown in Figure 7.12. A single closing event is identified between 8.5 and 9.8 ns in which the bridging residues move almost 5 Å closer, and the hinge angle, which otherwise has a similar value to

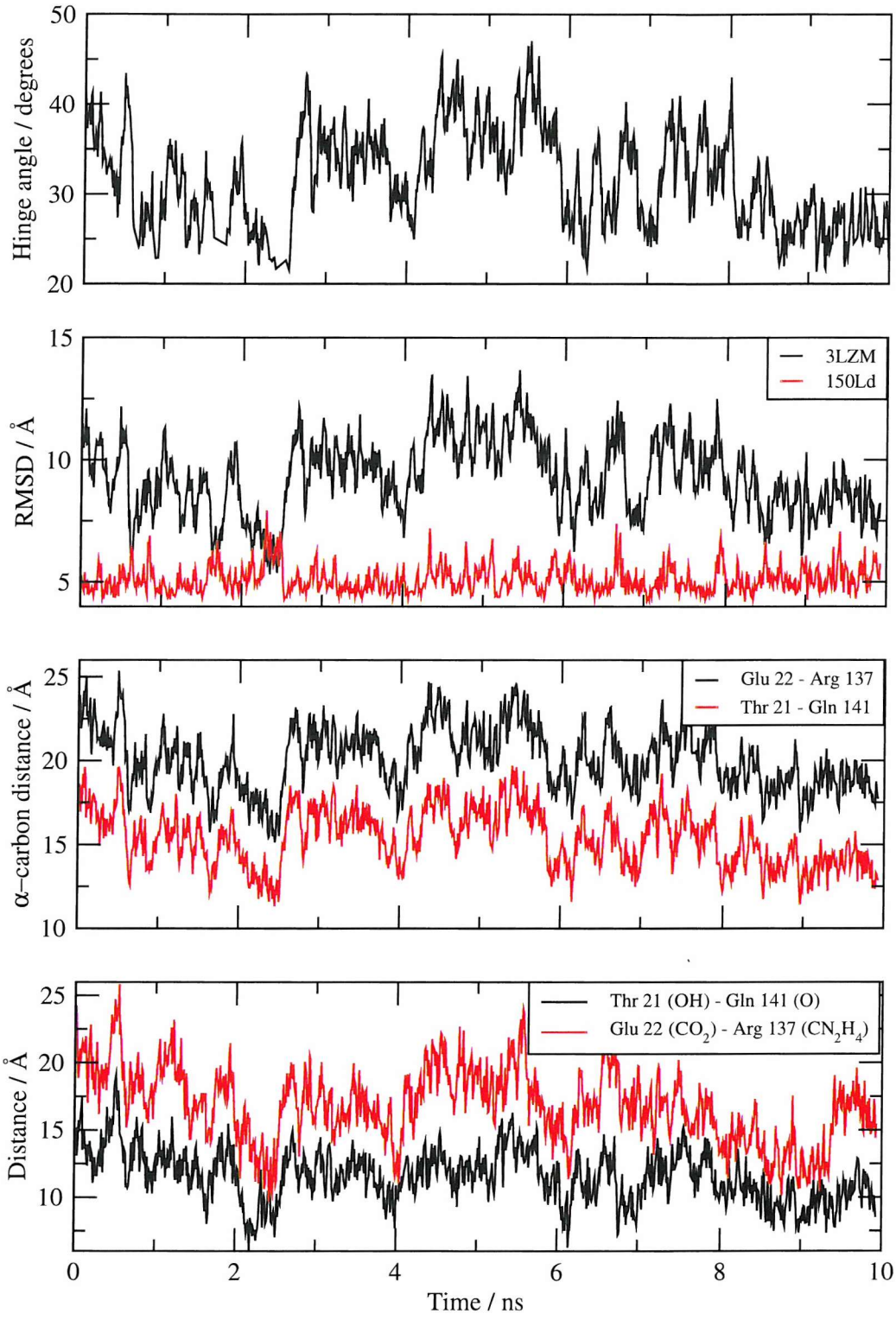



Figure 7.11: Analysis of the 10 ns NPT MD simulation of M6Id. Top: Hinge angle when compared to structure 3LZM using DynDom. Upper middle: RMSD of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower middle: α-carbon distances between bridging residues. Lower: Inter-atomic distances used to determine contact between bridging residues.

that of the M6Id hinge angle, falls outside the 80 % domain cutoff used, and is not included in the plot. During this event, the RMSD against the closed 3LZM structure is significantly reduced, falling briefly below the RMSD against the open, 150Ld structure. The protein then reopens, and the rest of the simulation returns to a state similar to that sampled by the M6Id system.

Even though an open conformer of WTd is sampled by simulation, the secondary structure analysis shown in Figure 7.13 is almost identical to that seen by the WT simulations. This supports the earlier finding that no alteration of secondary structure is required to change the molecule's hinge axis.

Analysis of the 10 ns NPT simulation of WTd shown in Figure 7.14 contains possibly the most important result shown so far. After only 200 ps a full closing event is sampled, resulting in a stable closed structure exhibiting all the structural features observed in simulations of WT. This observation was not seen in the work of de Groot *et al*, for which a separation of sampling between WT and WTd was maintained. Between 3.6 and 3.8 ns a brief opening event is seen, after which the structure returns to the obviously stable, closed conformer.

The secondary structure analysis for the 10 ns WTd simulation is very similar to that of the 20 ns WTd simulation, and is not shown.

### 7.2.3 Summary

Even though the simulations presented so far are of significantly increased length to any previously published on this system, sampling is still an issue, with duplicate simulations showing significant differences. However, the simulations appear to fall into two categories: those of an open conformer more similar to the M6Id structure, and those of a closed conformer, similar to the 3LZM pdb.

The simulations of a generally closed structure include both WT simulations, and the 10 ns WTd simulation. From these, infrequent, partial opening events are seen, but the closed structure, particularly the salt bridge formed by the bridging glutamic acid and arginine residues, exists for significant periods of time. The

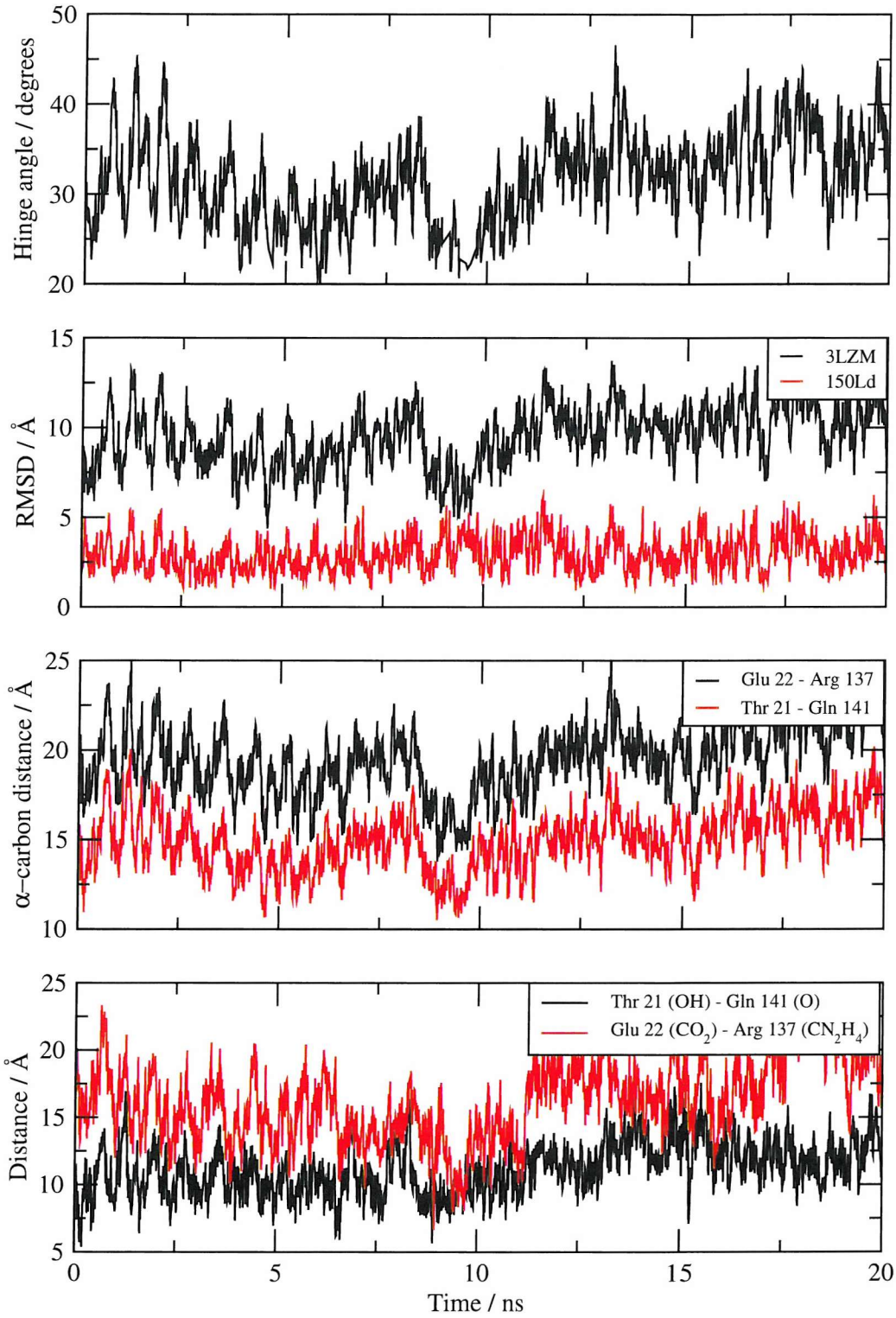



Figure 7.12: Analysis of the 20 ns NPT MD simulation of WTd. Top: Hinge angle when compared to structure 3LZM using DynDom. Upper middle: RMSD of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower middle: α-carbon distances between bridging residues. Lower: Inter-atomic distances used to determine contact between bridging residues.

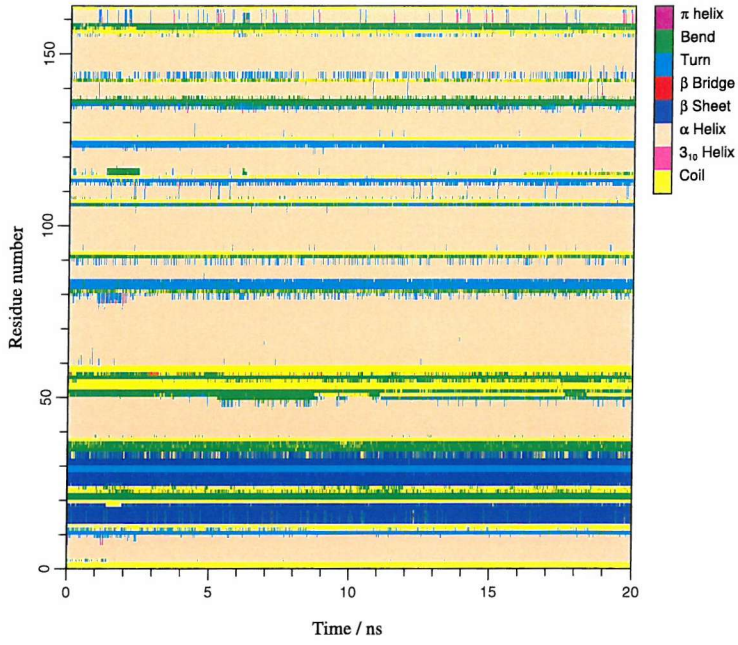



Figure 7.13: Secondary structure analysis of the 20 ns NPT MD simulation of WTd.

opening of the bridging residues is not sufficient evidence of a domain-scale motion, as shown by the third event discussed from the 20 ns WT simulation. This subset of simulations supports the experimental evidence of a dominant, ligand-inaccessible solution state, with a small population of a more open conformer. It is possible that a larger opening event occurs on ligand binding, but the investigation of this is outside the scope of the current project. Importantly, comparison to previously reported MD simulations, suggests that previous studies did not see the stable glutamic acid–arginine connection extensively sampled here. This connection appears to restrict the hinge angle to a value that is 25 to 30 degrees more closed than that of the 150Ld structure, and produces a distinct, closed conformer.

The simulations of generally open structures include both the M6Id simulations, and the 20 ns simulation of WTd. The M6Id mutant shows a secondary structure change, that may be a barrier to closure. This change is not present in WTd. Analysis of this subset of simulations indicate a more mobile hinge angle, possibly corresponding to the results of Zhang *et al.* and de Groot *et al.* However, the closing seen in the 10 ns WTd simulation suggests that if the closed

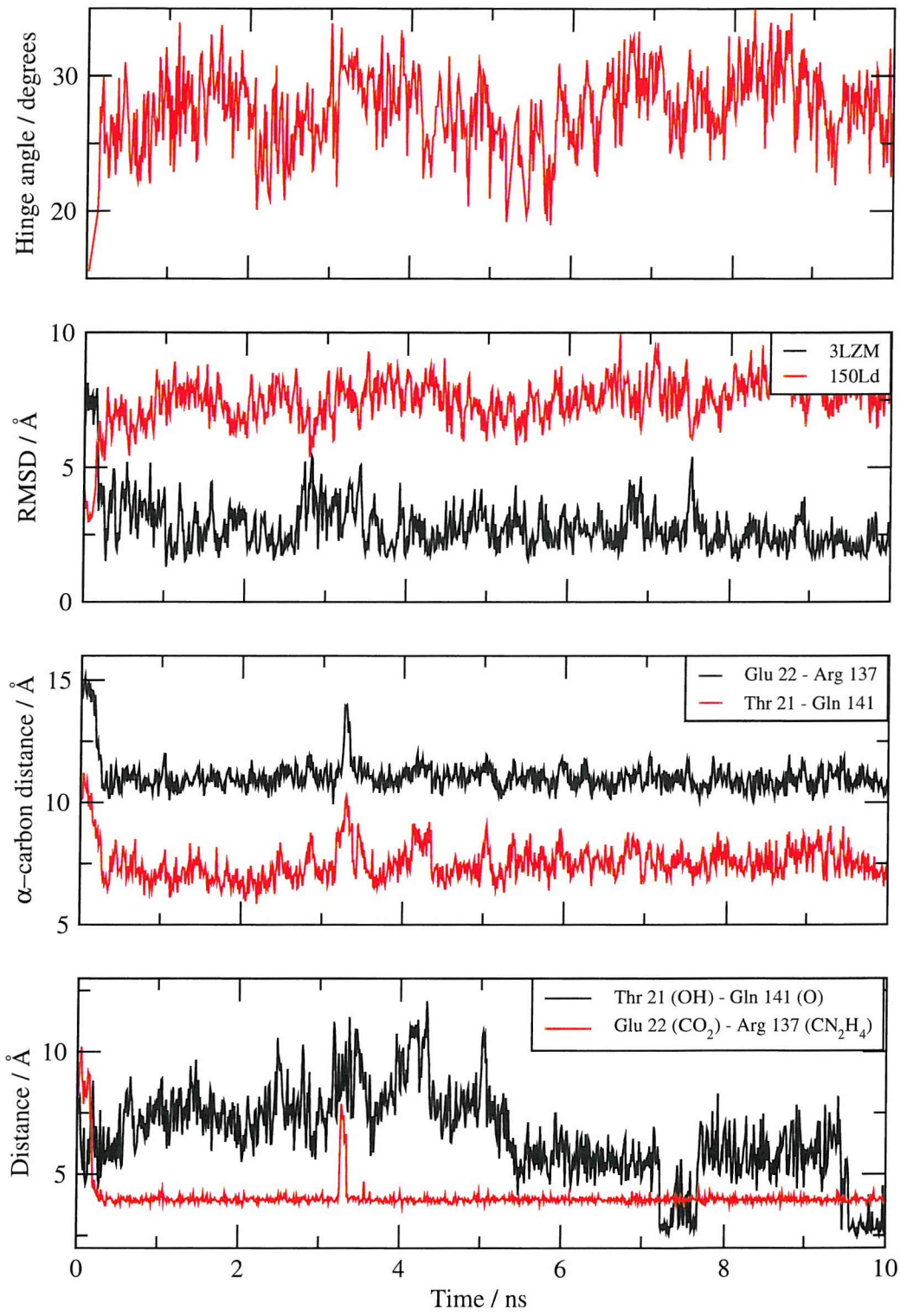



Figure 7.14: Analysis of the 10 ns NPT MD simulation of 150Ld WT. Top: Hinge angle when compared to structure 150Ld using DynDom. Upper middle: RMSD of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower middle: α-carbon distances between bridging residues. Lower: Inter-atomic distances used to determine contact between bridging residues.

conformation is reached, it will exhibit significant stability. There is insufficient evidence to suggest that this is the same for the M6I system.

A possible explanation for the simulation differences to those previously reported could be the use of the GROMOS force field by Zhang *et al.* and de Groot *et al.* A study of the movements of the HIV-1 protease flaps using the GROMOS force field saw dramatically more motion than other MD studies,[143,144] with the separation of bridging residues similar to those of T4 lysozyme. These simulations are discussed further in Chapter 9. It is unclear whether these results are force field dependent.

As mentioned, there are significant differences between duplicate simulations, and limited sampling is still an important issue. To further investigate the distribution of open and closed conformations, parallel tempering shall be used.

## 7.3 Thermal simulations

By increasing the temperature of simulations, some indication of possible motions can be gained. Results at raised temperatures are not necessarily relevant however, due to the optimisation of the force field for a single temperature. As previously discussed, the high energy regions of the potential energy surface may not be reproduced accurately.

Thermal simulations are also used to assign a maximum temperature to the parallel tempering performed on the WT system. The maximum temperature needs to be sufficient to frequently sample the desired conformational motions, if information on their distribution is to be obtained. 6 ns simulations of the WT, M6Id and WTd systems have therefore been performed at 50 K intervals from 300 to 500 K.

The Langevin thermostat coupling parameter is tightened to 5 $ps^{-1}$ and simulations are performed in the NVT ensemble, to be consistent with the parallel tempering methodology presented in Chapter 6. Simulations are started from the same equilibrated states used for the 20 ns NPT simulations, and velocities are

randomly assigned and scaled to reproduce the correct temperature. It is worth noting that the differences between the first part of the 20 ns simulations previously discussed, and the 300 K results over 6 ns presented here, is due only to the reassignment of velocities.

Increased sampling is seen by the higher temperature simulations, and the DynDom analysis is often unable to recognise the same domains using an 80 % domain cutoff. Hinge angle trajectories are therefore not included for the thermal simulations. The MD simulations at 300 K previously reported are sufficiently characterised by the α-carbon distance between the bridging glutamic acid and arginine residues coupled with the RMSD analysis against the 3LZM and 150Ld pdb structures. Again the RMSD is calculated by superposition over the C-terminal domain and the RMSD of the N-terminal domain measured.

**T4 lysozyme wild-type (WT)**

The results of the thermal simulations of the WT system are shown in Figure 7.15 and confirm the previous conclusion that the closed conformation of T4 lysozyme is a distinct, highly populated state in solution. At 300 and 350 K, only small deviations from the closed structure are seen within 6 ns. At 400 and 450 K, several opening events are seen, however in each case the simulations return to the closed conformation. The 500 K simulation however shows a dramatic opening event, beyond the scale of those measured previously.

A secondary structure analysis on the five simulations is shown in Figure 7.16. The cause of the dramatic opening of the 500 K simulation is clear; major secondary structure disruption within the N-terminal domain. Such changes are also seen to several helices at 450 K, but to a lesser extent. It can be seen therefore, that without reaching temperatures at which the protein denatures, the closed conformation is a stable state, with only brief, infrequent opening events.

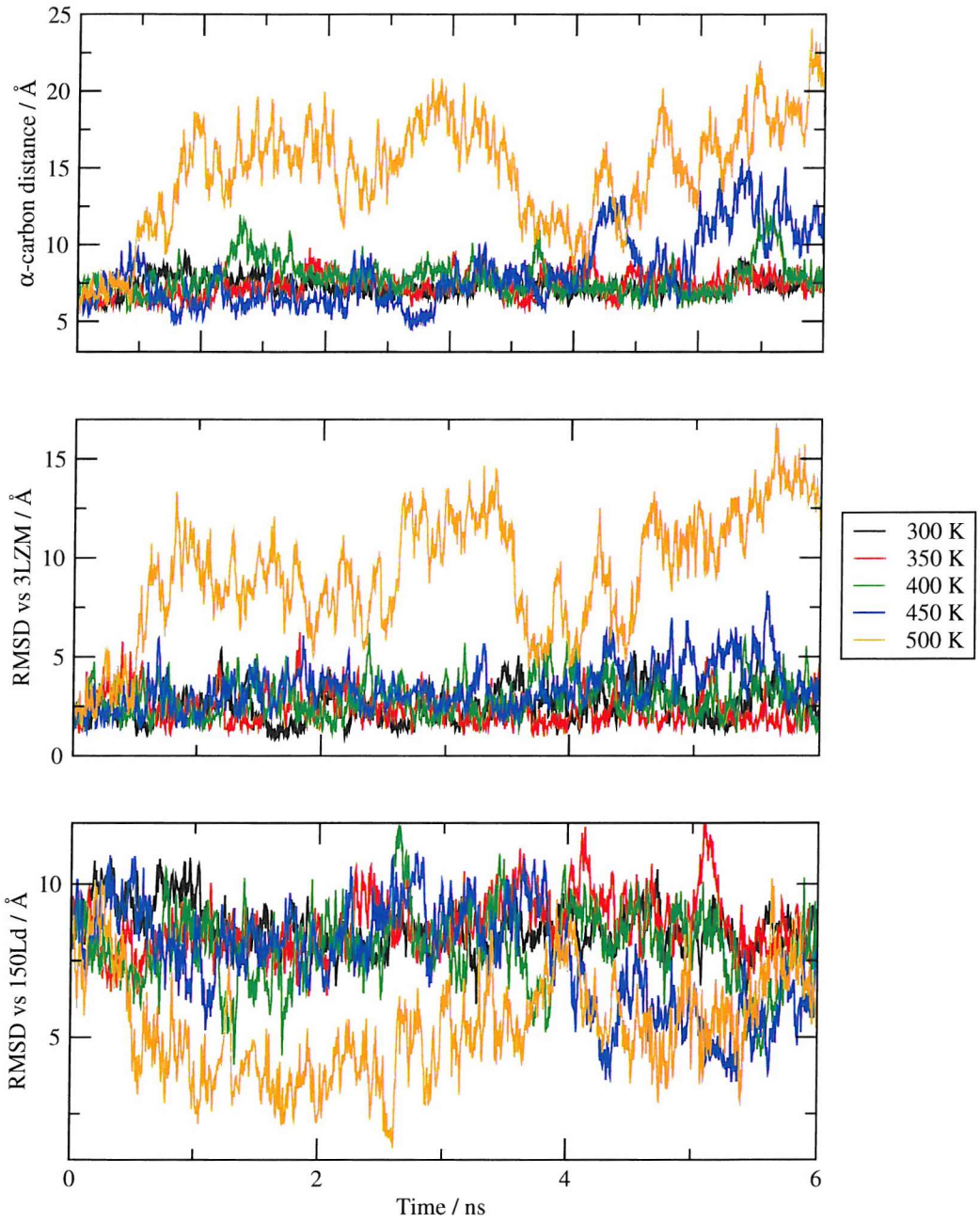
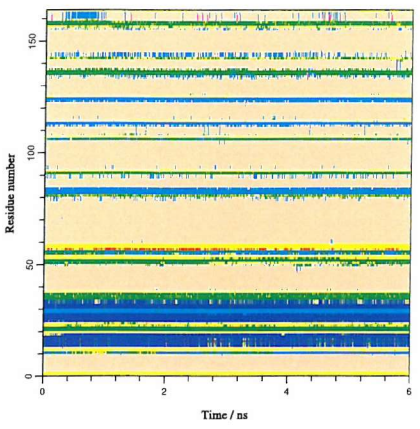
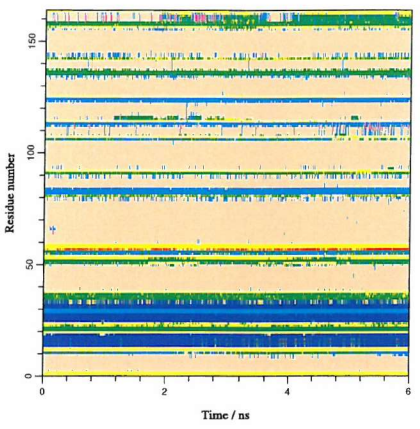


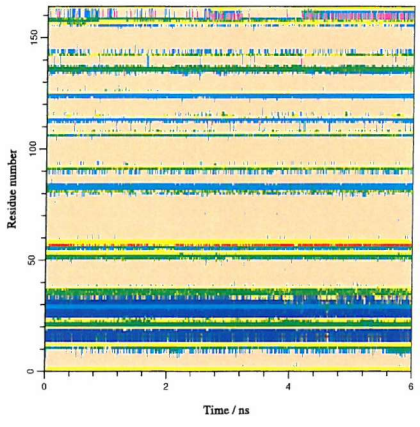

Figure 7.15: Analysis of the 6 ns NPT MD simulation of WT at a range of temperatures. Top: α-carbon distances between bridging residues. Middle: RMSD with respect to structure 3LZM of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower: RMSD with respect to structure 150Ld of C-terminal domain calculated by superposition on N-terminal domain.



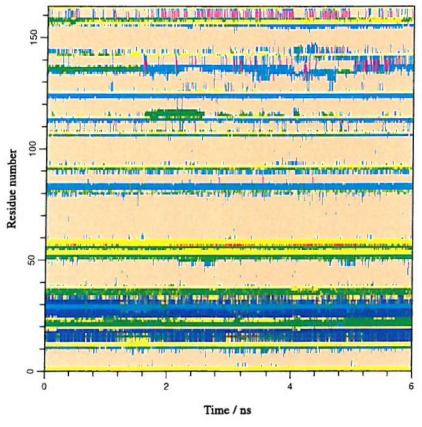

(a) 300 K



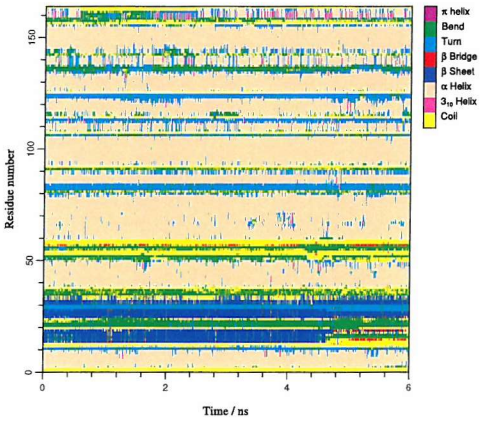

(b) 350 K




(c) 400 K




(d) 450 K




(e) 500 K

Figure 7.16: Secondary structure analysis of the 6 ns NPT MD simulation of WT at a range of temperatures.

### Most open conformation of the M6I mutation (M6Id)

The five thermal simulations of the M6Id system are shown in Figure 7.17. Several, significant closing events are seen at and above 400 K, particularly at 500 K, which closes and opens several times. At 300 K there are no closing events, and results are consistent with the NPT simulations previously reported.

The 450 K simulation contains several occasions for which the RMSDs show a structure more similar to pdb entry 3LZM, than the open conformer, 150Ld. Interestingly, these are not accompanied by a close α-distance, as previously seen. The domain opening motion is therefore not directly induced by the proximity of the bridging residues.

The secondary structure analysis of the five thermal simulations for the M6Id system show that at 500 K, much of the helical and β-sheet structure is lost. At 450 K, damage to the helical structure is seen to a lesser extent. These results are not shown.

### Wild-type T4 lysozyme starting from an open structure (WTd)

The results of the thermal simulations of the WTd system are shown in Figure 7.18. Closing events are seen at 400 and 450 K, and briefly, at 3 ns, for 500 K. Interestingly once closed, both the 400 and 500 K simulations re-open and return to conformations similar to the 150Ld pdb structure. However, at 450 K the structure remains closed from 4.3 ns until the end of the simulation, with a conformation shown by the RMSDs to be closer to the 3LZM structure. This suggests that the M6I system also sees a distinct closed conformation, although more simulation is required to confirm this hypothesis.

The secondary structure analysis of the WTd thermal simulations is similar to that shown in Figure 7.16, and is not included. Significant disruption of both domains is seen at 500 K, and to a lesser extent, at 450 K.

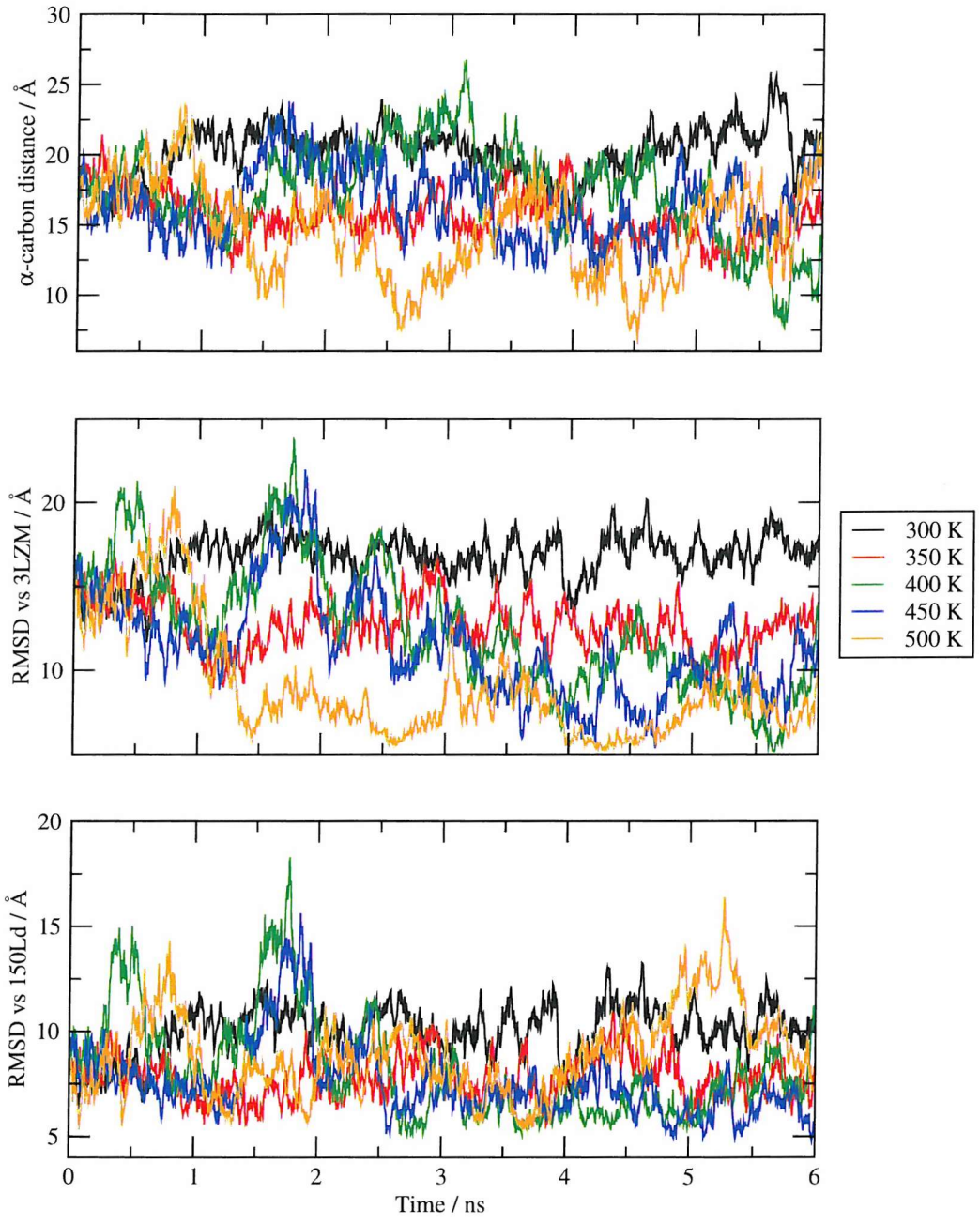



Figure 7.17: Analysis of the 6 ns NPT MD simulation of M6Id at a range of temperatures. Top: α-carbon distances between bridging residues. Middle: RMSD with respect to structure 3LZM of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower: RMSD with respect to structure 150Ld of C-terminal domain calculated by superposition on N-terminal domain.

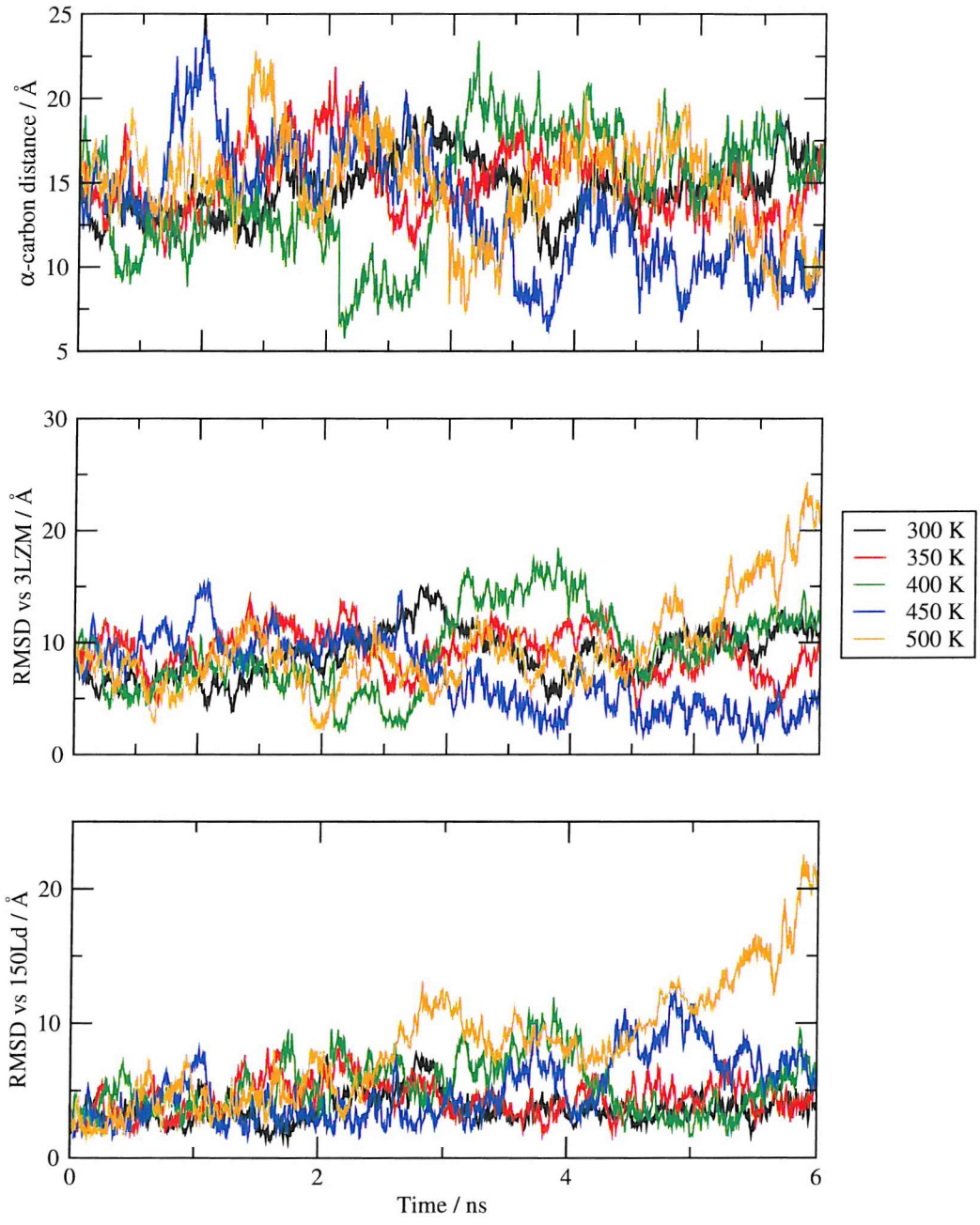



Figure 7.18: Analysis of the 6 ns NPT MD simulation of WTd at a range of temperatures. Top: α-carbon distances between bridging residues. Middle: RMSD with respect to structure 3LZM of C-terminal domain (residues 75 to 164) calculated by superposition on N-terminal domain (residues 13 to 63). Lower: RMSD with respect to structure 150Ld of C-terminal domain calculated by superposition on N-terminal domain.

### 7.3.1 Summary

Several opening and closing events are sampled by the thermal simulations presented here. Results are in agreement with those from the long NPT simulations, suggesting a stable, closed conformation for the WT system, which is obtainable for the initially open WTd system. A stable, closed conformation for the M6I system is suggested, although more simulation is required to confirm this.

Significant disruption of the secondary structure is seen above 400 K, suggesting this as a maximum temperature for PT simulations. Since the conformational transition between open and closed states is sampled at and below 400 K, the distribution between these conformations should be obtainable with this temperature cap.

## 7.4 Parallel tempering

A parallel tempering simulation of the WT system has been performed reaching 400 K, as suggested by the thermal simulations previously presented. PT parameters are taken from those found to be suitable for the YPGDV system, with 1 ps of NVT MD between attempted PT moves and a 5 $ps^{-1}$ Langevin thermostat damping parameter.

The T4 lysozyme system is considerably larger than that of YPGDV, and a low acceptance probability of 0.2 is used to minimise the computational expense of simulating the PT ensemble. The temperature distribution has been calculated as described in Chapter 6, using the means and variances of seven 100 ps simulations at 25 K intervals from 275 to 425 K. A 2 fs timestep is used for all of the 42 temperatures, which span the range 291.1 to 402.0 K.

### 7.4.1 Parallel Tempering results

2 ns of PT simulation has been completed for the WT system, giving a total of 82 ns of NVT simulation. This is a short timescale in comparison to the 20 ns PT simulation of YPGDV, and PT will be extended and reanalysed, hopefully targeting 10 ns

of PT simulation, a total of 420 ns of simulation time. Completion of this simulation is outside the timescale of this project, and the results so far are presented here.

The probability profile of the 2 ns PT simulation is shown in Figure 7.19. With no changes in the timestep used, a smooth profile is obtained that reproduces the desired acceptance probability. It is interesting to note a slight increase in the acceptance probability as temperature is raised. This could be due to an inadequate description of the variance in potential energy as a function of temperature, used for the temperature distribution calculation. However, the deviation from the desired acceptance probability is small.

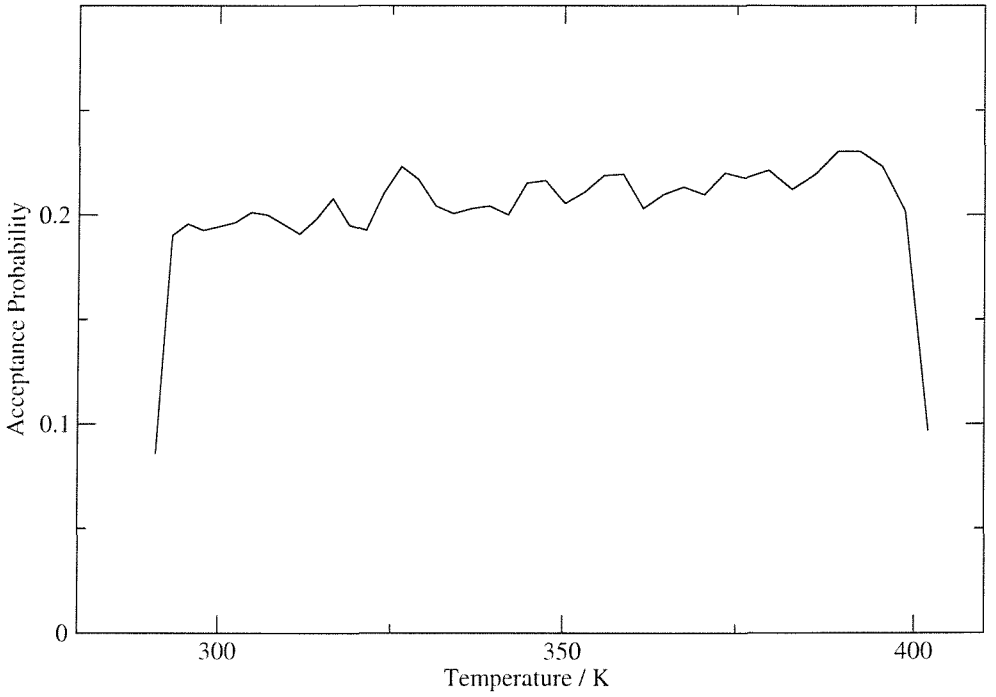



Figure 7.19: Probability profile from 2 ns of WT PT.

It is important to confirm that the replicas are sufficiently mobile across the PT ensemble. The mobility of replicas that started simulation approximately 25 K apart are shown in Figure 7.20. Significant movement in temperature space is seen, particularly for replica E, which begins at 300.0 K, reaches over 375 K after 1.2 ns and falls back to 300.0 K by 2 ns. Replicas are clearly sufficiently mobile, but more simulation time is required to allow all replicas to contribute to sampling at every temperature.

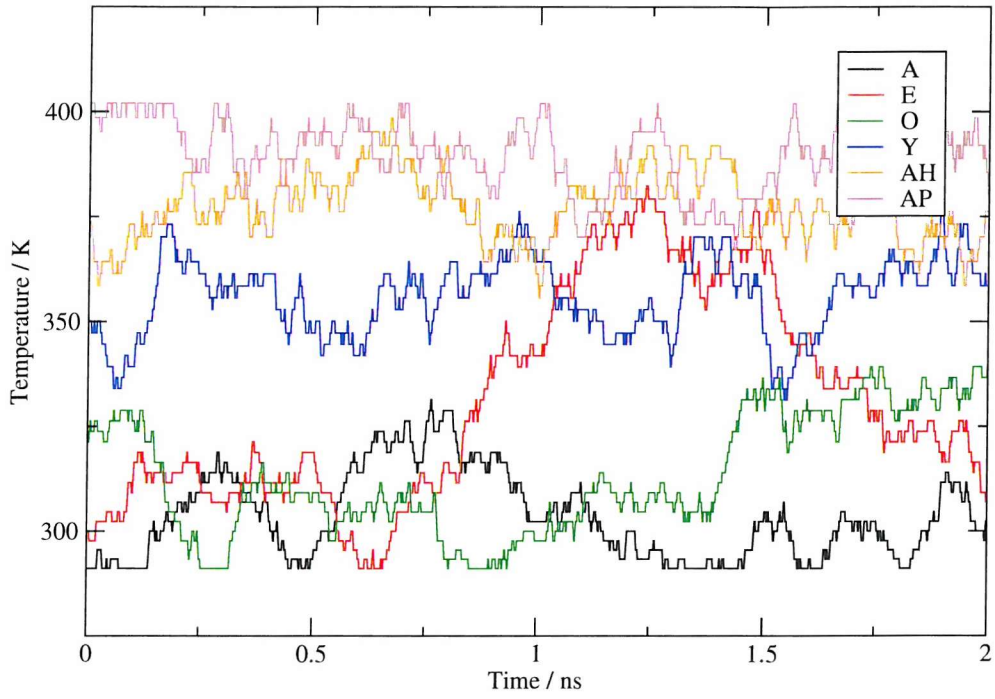



Figure 7.20: Temperature mobility of replicas over 2 ns of WT PT.

The 400 K temperature cap was chosen to minimise secondary structure disruption, and a secondary structure analysis of all replicas reaching the maximum temperature is shown in Figure 7.21. Results are as expected, with little deviation from the structure reported for the long 300 K NPT simulations of the WT and WTd systems.

It is sufficient to characterise the protein conformation using the α-carbon distance of the glutamic acid-arginine bridging residues, and the RMSDs of the C-terminal domain with superposition of the N-terminal domain against known closed (3LZM) and open (150Ld) structures. These results are shown, for a range of temperatures, by frequency plots in Figure 7.22. Comparison to results from the 20 ns WT NPT simulation suggest that the structure can be considered as open if the glutamic acid-arginine distance is over 10 Å, and if the RMSD against the open structure is less than 5 Å. The frequency plots in Figure 7.22 show a minority of states with these characteristics, as expected by experimental evidence.

The population of the open conformation is small in comparison to that of the closed state, and Figure 7.23 shows an enlargement of the the low-frequency

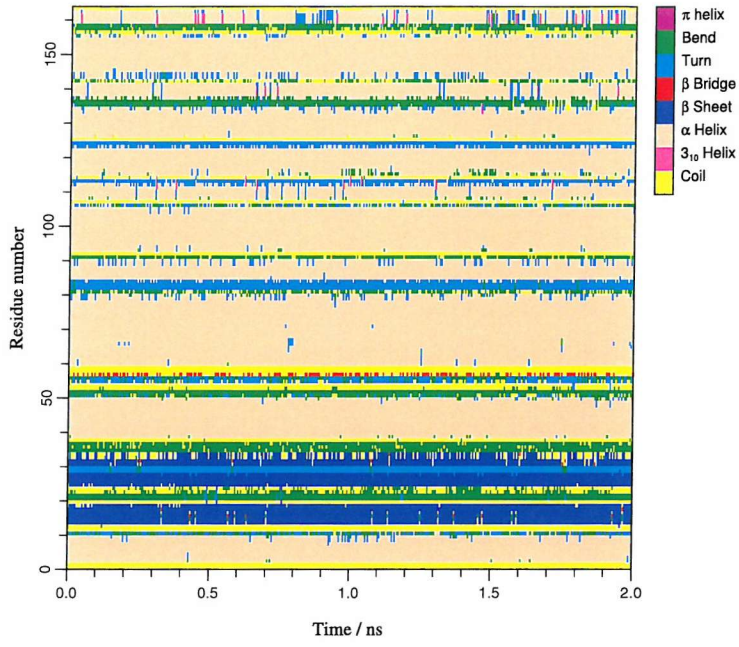



Figure 7.21: Secondary structure analysis of replicas at 402.0 K for 2 ns of WT PT.

regions of Figure 7.22. The open state has not yet been seen at 300.0 K.

It is worth discussing the top plot of Figure 7.23 in some detail. All of the starting structures were initially closed, with an α-carbon separation between the bridging residues of 5 to 10 Å. Throughout simulation, this closed conformation remains the dominant state at every temperature. However, at high temperatures, the α-carbon separation is also sampled between 10 to 15 Å, suggesting the presence of more open states, and the transitions to these from the closed conformation. For 323.8 K, there is no sampling between 10 and 11 Å, but a conformer is seen with a bridging residue separation between 11 and 12.5 Å. This suggests the energy barrier between the closed and open states was overcome at a higher temperature, after which the opened replica moved down to 323.8 K. At 350.1 K, increased sampling of the 11 to 12.5 Å α-carbon distance conformation is seen, including sampling of the transition from the closed conformation. Conformations with a bridging residue separation above 13 Å are not seen at 350.1 K, but are present by 376.2 K. These results suggest that more open conformations have higher potential energies, in agreement with experimental evidence of an excited,

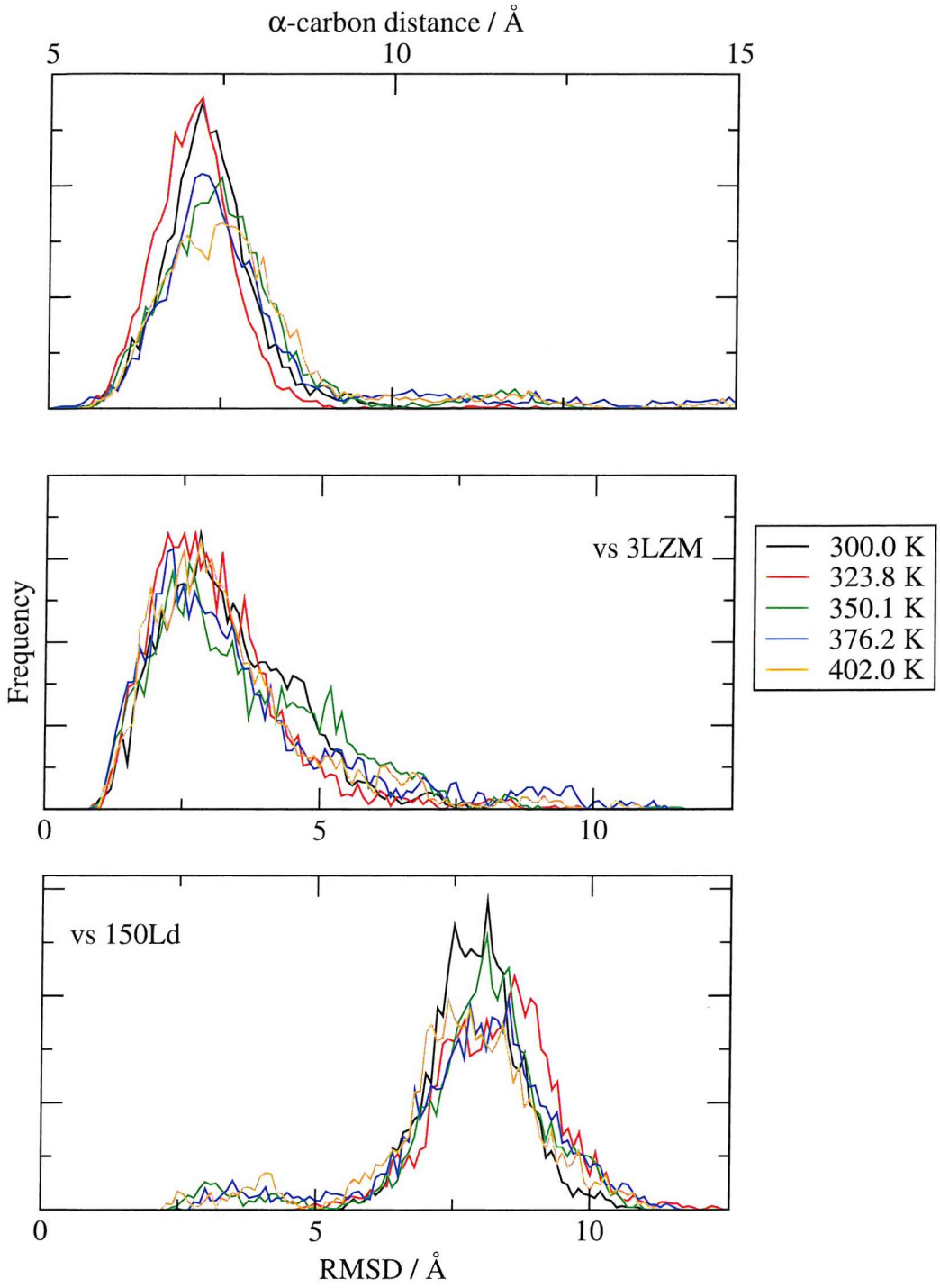



Figure 7.22: Analysis of replicas from 2 ns of WT PT simulation. Top: α-carbon distance between the glutamic acid- arginine bridging residues. Middle: RMSD against the 3LZM structure. Bottom: RMSD against the 150Ld structure. RMSDs are taken of the C-terminal domain with superposition over N-terminal residues.

ligand accessible state.

Whilst these conclusions serve as an excellent example of expected parallel tempering results considering the available experimental evidence, the data cannot be shown to have converged. It is likely that sampling of the open state will increase at all temperatures given more time.

### 7.4.2 Summary

The PT results in this section show the success of the PT algorithm in delivering conformers generated at high temperatures to lower temperatures, however the ensemble has not converged, given that the closed to open transition is expected in a small population of states at 300.0 K. The results are very encouraging however, and should be of significant interest once a longer timescale has been achieved.

There is no doubt that the PT simulations support the hypothesis of a dominant population of largely closed conformers, in good agreement with experimental evidence, and in direct disagreement with the work of Zhang *et al.* and de Groot *et al.*, whose short simulations identified no closed conformer of increased stability.

## 7.5 Reversible digitally filtered molecular dynamics

The parallel tempering results that have been presented in this chapter required extensive simulation to generate, and require significantly more before convergence is reached. If an investigation was desired into the hinge mobility of a range of T4 lysozyme mutants, or the WT system from a number of starting conformers, the computational expense of long MD and PT simulations would be prohibitive. However, RDFMD simulations are quick to produce, and have been shown to increase conformational sampling in peptide chains. The bridging residues of T4 lysozyme are clearly of significant importance in maintaining the closed conformation, and in this section RDFMD is applied to all of the atoms belonging to three of the four bridging residues, to increase the rate of opening events seen. RDFMD is not applied to the fourth residue, arginine 137, as it forms one end of a

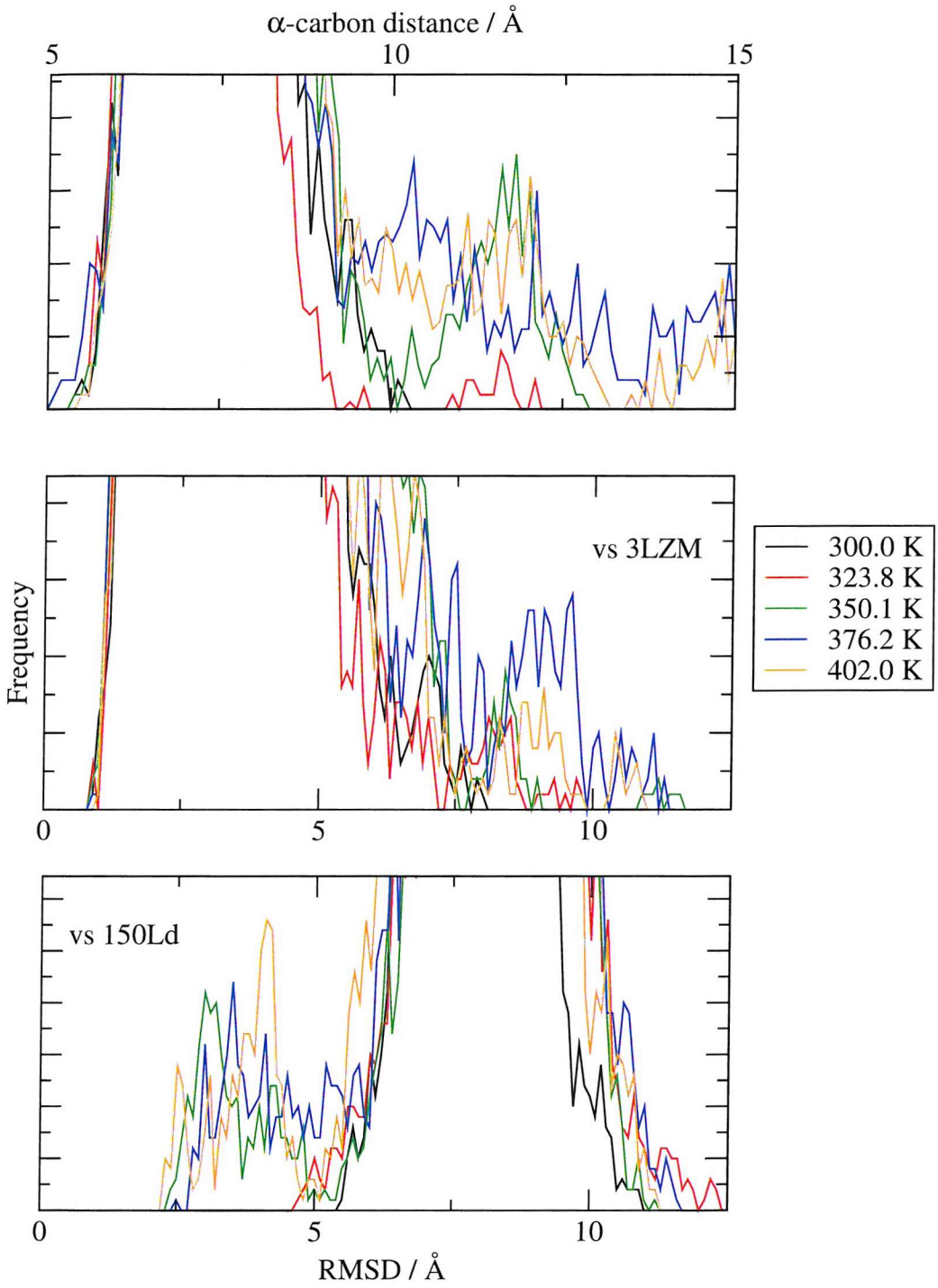



Figure 7.23: Analysis of replicas from 2 ns of WT PT simulation. Only regions with low frequencies are shown. Top: α-carbon distance between the glutamic acid-arginine bridging residues. Middle: RMSD against the 3LZM structure. Bottom@ RMSD against the 150Ld structure. RMSDs are taken of the C-terminal domain with superposition over N-terminal residues.

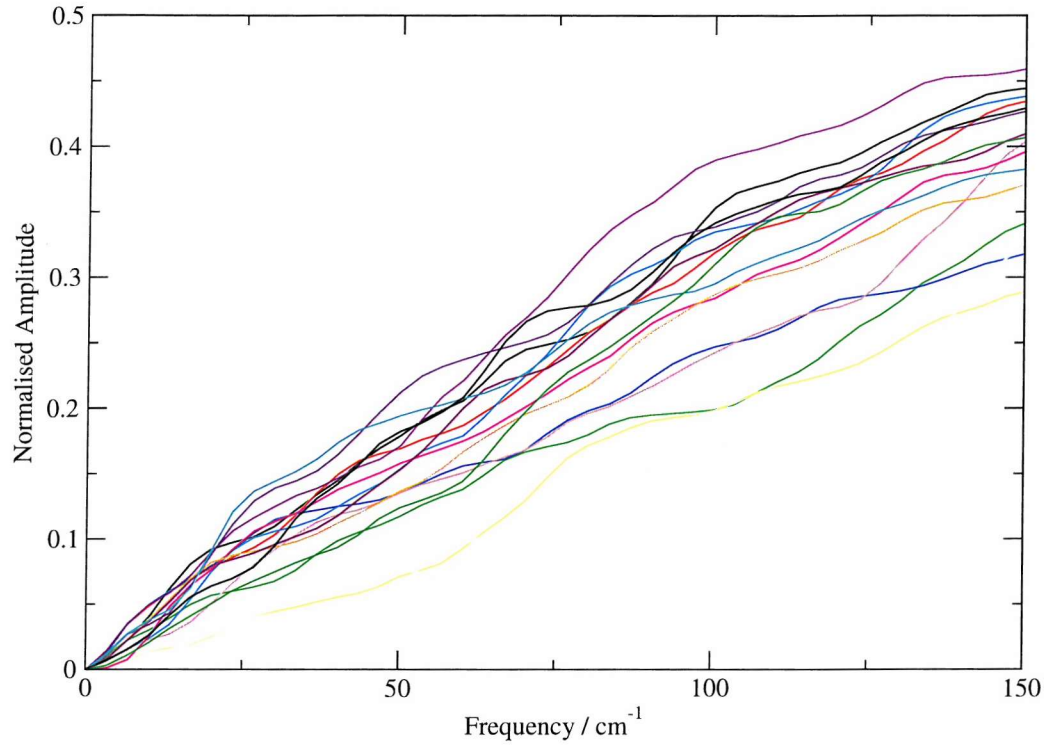
helix. It is not desired to disrupt the protein's secondary structure.

Cumulative amplitude plots of the spectral components from a short YPGDV simulation were used in Chapter 5 to indicate the frequency separability of the $\psi$ and $\phi$ backbone torsions from those of the $\omega$ angles. This has been reproduced for the WT system, using a 10 ps NVE simulation, with sampling of the torsion angles every 2 fs for bridging residues and their neighbours. The cumulative amplitude plots generated from this simulation are shown in Figure 7.24. There are significant low frequency motions for several of the $\omega$ angles sampled, and no frequency separability of the $\psi$ and $\phi$ torsions is seen. The previously identified limitation of being unable to target specific degrees of freedom whilst excluding those that have a similar frequency and involve similar atoms, is clearly of significance to biological applications.

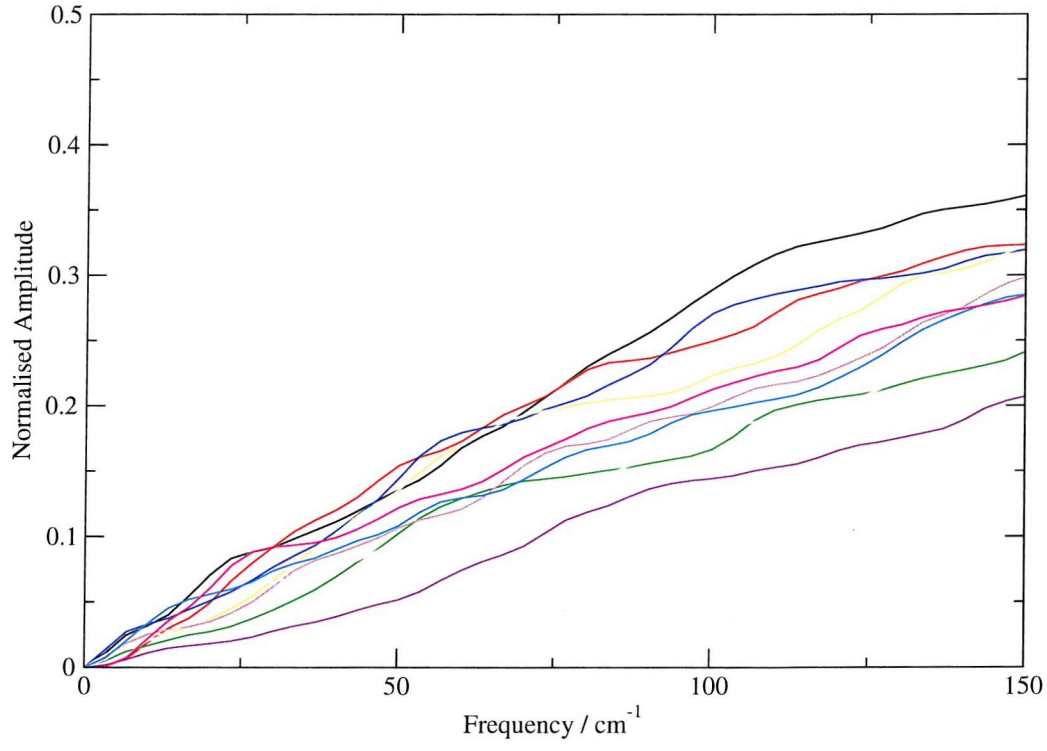
$\omega$ angle transitions are known to involve higher energy barriers than those for $\psi$ and $\phi$ transitions, and the internal temperature cap shall therefore be used to maintain all-trans peptide bonds. No RDFMD simulations presented in this chapter include $\omega$-angle transitions.

### 7.5.1 Optimised protocol results

RDFMD simulations have been performed using the equilibrated WT system previously described as the starting point for the 20 ns NPT MD simulation. The RDFMD protocol used is that determined by optimisation on the YPGDV system as described in Chapter 5, including a filter delay of either 50 or 100 steps, an amplification factor of 2, 4 ps of NPT simulation between filter sequences and a 201 coefficient, 0–100 $cm^{-1}$ filter. There is no maximum on the number of filter application allowed before reaching the internal temperature cap, although this has been monitored, and is never greater than 10. The internal temperature cap of 900 K for the flexible YPGDV system, as suggested by the conclusions of Chapters 5 and 6, is likely to be below that required for the application to the more constrained peptide regions of T4 lysozyme targeted here. Simulations are therefore performed with a range of internal temperature caps: 900 K, 1100 K




(a) ψ and ϕ angles




(b) ω angles

Figure 7.24: Cumulative amplitude plots of backbone torsions for residues targeted by RDFMD, and their neighbours. Each colour represents a different backbone torsion.

and 1500 K. Trajectories for which a transition occurs in any peptide bond are excluded from analysis.

The results of RDFMD simulations of 100 filter sequences (a total of 400 ps of NPT simulation) are shown in Figure 7.25. As previously discussed, an α-carbon distance between the glutamic acid and arginine bridging residues of 10 Å suggests an opening event, and this is achieved on several occasions by each of the simulations performed. The protocols used also sample closing events from the opened structures. Simulations with an internal temperature cap of 900 K are less mobile than those of 1100 K, and do not show such dramatic opening events.

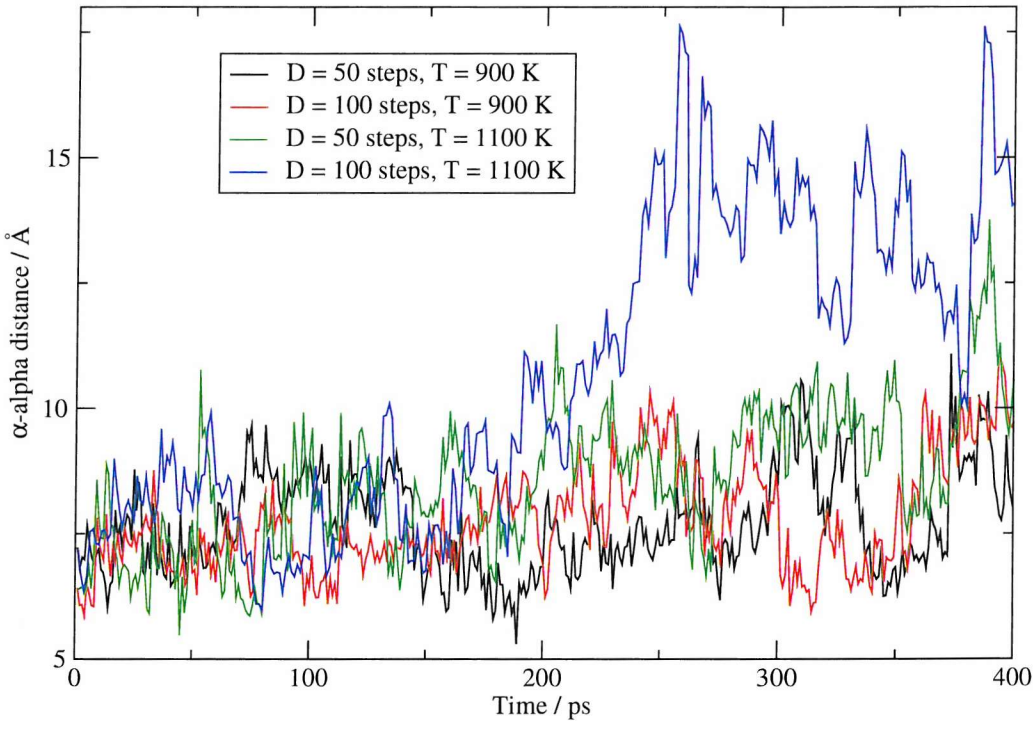



Figure 7.25: Inter-residue distance across T4L opening for RDFMD simulations. Filter delay, D, and internal temperature cap, T, are shown in legend.

It is necessary to show that the simulation trajectories produced by RDFMD sample relevant conformational space, and without the convergence of PT results (as were available for the YPGDV system), it shall be assumed sufficient to demonstrate the validity of simulations that show the largest motions. Figure 7.26 therefore shows an RMSD and secondary structure analysis for the simulation with a filter delay of 100 steps and an internal temperature cap of 1100 K.

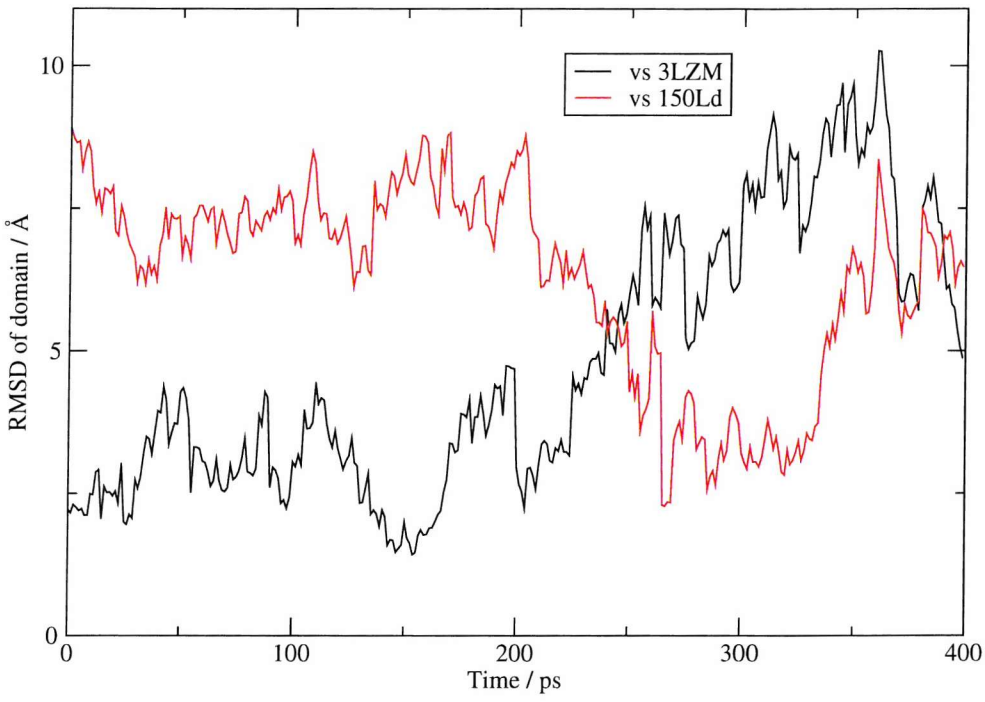
This simulation reports an opening of 10 Å from the starting conformation. The RMSD against the open structure (150Ld) reaches values slightly lower than those obtained by the long NPT MD simulations of the WT system, and just before the end of the simulation, the open structure begins to close. The conformation most similar to the open 150Ld structure is sampled at is shown in Figure 7.27 for comparison to Figure 7.2. The secondary structure analysis shows a small increase in the disruption to the α-helix formed between residues 137 and 141, in comparison to simulations at 300.0 K, although the helix is not removed, and is only intermittently disturbed. Residues 137 and 141 are involved in the bridging of the T4 cavity, and so this result may not be avoidable with the current protocol. Analysis therefore suggests the sampling of a reversible event, with an opening motions similar to the largest seen by the 20 ns NPT MD simulation.

RDFMD simulations have therefore achieved the desired results, with increased mobility of the distance between bridging residues, via the disturbance of the polar interactions that hold the cavity closed. RMSD results for the largest motion confirm domain-scale events are occurring, and analysis shows that disruption to secondary structure near the targeted region is intermittent. The protocol does not involve excessive heating of the protein.

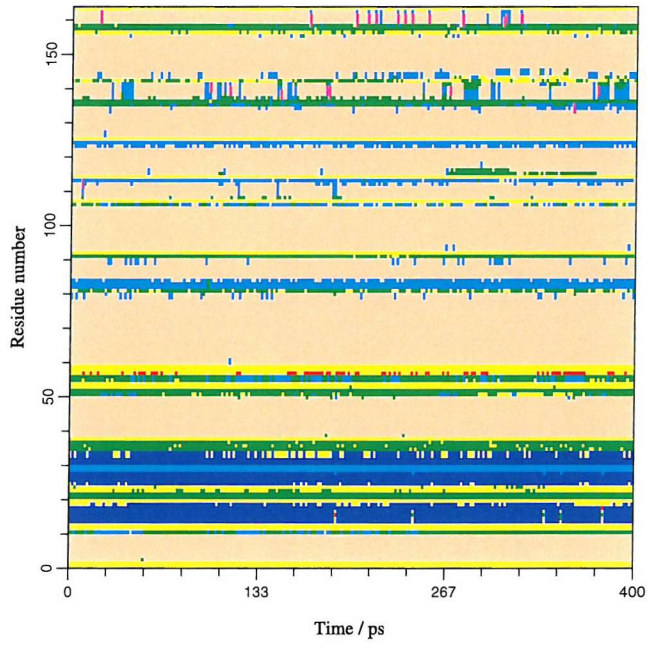
The computational expense of producing an RDFMD simulation with 400 ps of NPT MD is less than 4 % of that required to produce 20 ns MD, and RDFMD has revealed an opening event similar to the largest sampled by the 20 ns or 10 ns NPT simulations presented on the WT system. The use of RDFMD for conformational searching has therefore been successfully applied, using the RDFMD protocol generated using the YPGDV system. This demonstrates the portability of the parameter set, as claimed in Chapter 5.

### 7.5.2 Filter sequences

As seen in Chapter 5, the efficiency of the RDFMD method is heavily dependent on the protocol used. It would be computationally expensive to perform optimisation of the parameters for each system to which RDFMD is applied, and the portability




(a) RMSD analysis.




(b) Secondary structure analysis.

Figure 7.26: Analysis of the WT RDFMD simulation using an internal temperature cap of 1100 K and a filter delay of 100 steps.

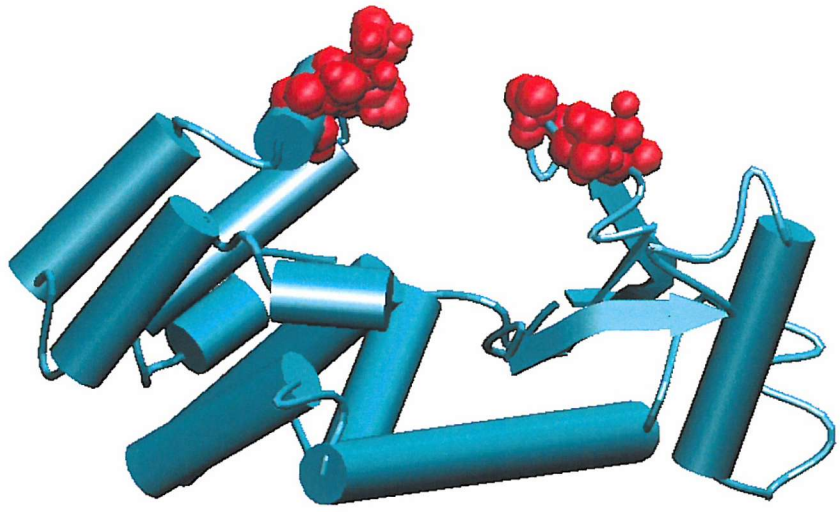


Figure 7.27: The conformation most similar to the 150Ld structure (with the lowest RSMD against 150Ld as shown in Figure 7.26 (a)) obtained during an RDFMD simulation with an internal temperature cap of 1100 K and a filter delay of 100 steps. The representation shown is as described in Figure 7.2.

of the protocol optimised on YPGDV suggests that this is not necessary. However, it is useful to develop a method of quickly testing different RDFMD protocols, to test protocol adjustments specific to a system, such as the choice of residue targets, or alternative amplification techniques.

To do this, similar techniques to those used in Chapter 5, to optimise the RDFMD protocol on YPGDV, are returned to. These involve using a large number of RDFMD filter sequences run from the same starting structure, but with randomised velocities. To explore the effects of significantly amplified low frequency motions, a filter sequence can be defined to contain a fixed number of filter applications. Should the application of a filter take the internal temperature of targeted residues above the desired temperature cap, the amplification factor can be dynamically reduced, guaranteeing that the internal temperature cap is not reached. The specified number of filter applications are therefore always completed.

Using this method, the resulting conformer is not required for further simulation, and should an undesirable conformation be obtained (such as one that includes an $\omega$ transition), the trajectory can be identified by analysis, and discarded. In the work presented here, only simulations containing $\omega$ transitions are discarded, but

the methodology could be simply extended to a range of checks, for example, of RMSD against known structures.

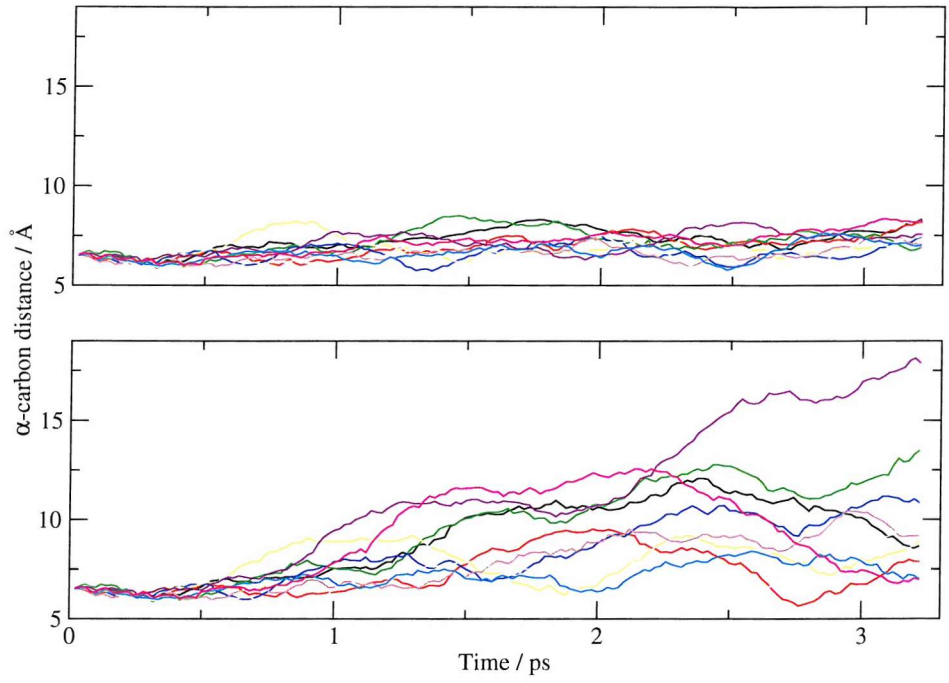
Figure 7.28 shows the results of RDFMD filter sequences for a range of internal temperature caps and a filter delay of 50 steps, targeting the same residues as used previously. 10 sets of randomised velocities are used as starting points, thirty filter buffers are completed for each simulation, and analysis is performed on the continuous trajectory created by following the progression of the system forward in time.

The simulations in Figure 7.28 (a) were generated using an initial amplification factor of 2. Those for Figure 7.28 (b) used an amplification factor that is dynamically chosen to increase the system temperature of the targeted residues by 700 K with each filter application. This is similar to the methodology used to test different frequency targets in Chapter 5. The value of 700 K was determined as optimal in the same manner as the amplification factor of 2.

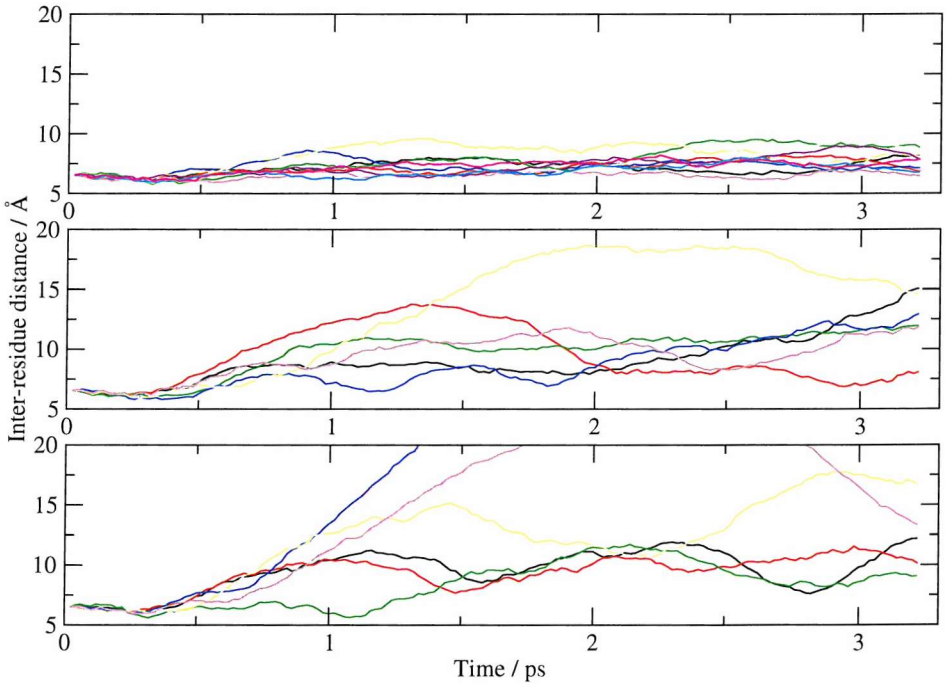
For each protocol, a 500 K internal temperature cap produces little significant sampling over the thirty buffers. The RDFMD method is only progressive if sufficient energy is put into the system so that it does not dissipate between filter applications. For a 500 K internal temperature cap, the initial amplification factor of 2 (and obviously an amplification of 700 K) is however always reduced. The energy put into the system dissipates between filters and a non-progressive method is achieved.

For an internal temperature cap of 1000 K, some simulations see isomerisation of the peptide bond. In the all-trans simulations, significant opening events are seen for each protocol, including many where the distance between the bridging residues is seen to return to a more closed conformation. These are of particular interest, suggesting reversible sampling of the conformational motions.

Using an internal temperature cap of 1500 K, all simulations with an initial amplification factor of 2 see $\omega$ transitions. However, many of those that use an amplification of 700 K do not. An amplification factor of 2 initially produces small




(a) An amplification factor of 2 is used, which is reduced if required to keep the temperature below 500 K (top) or 1000 K (bottom)




(b) Each filter application increases the temperature of the system by 700 K, which is reduced if required to keep the temperature below 500 K (top), 1000 K (middle), or 1500 K (bottom)

Figure 7.28: Inter-residue distance across bridging residues using filter pulse of 30 filters and a filter delay of 50 steps. Only simulations that did not see isomerisation of the peptide bond are included.

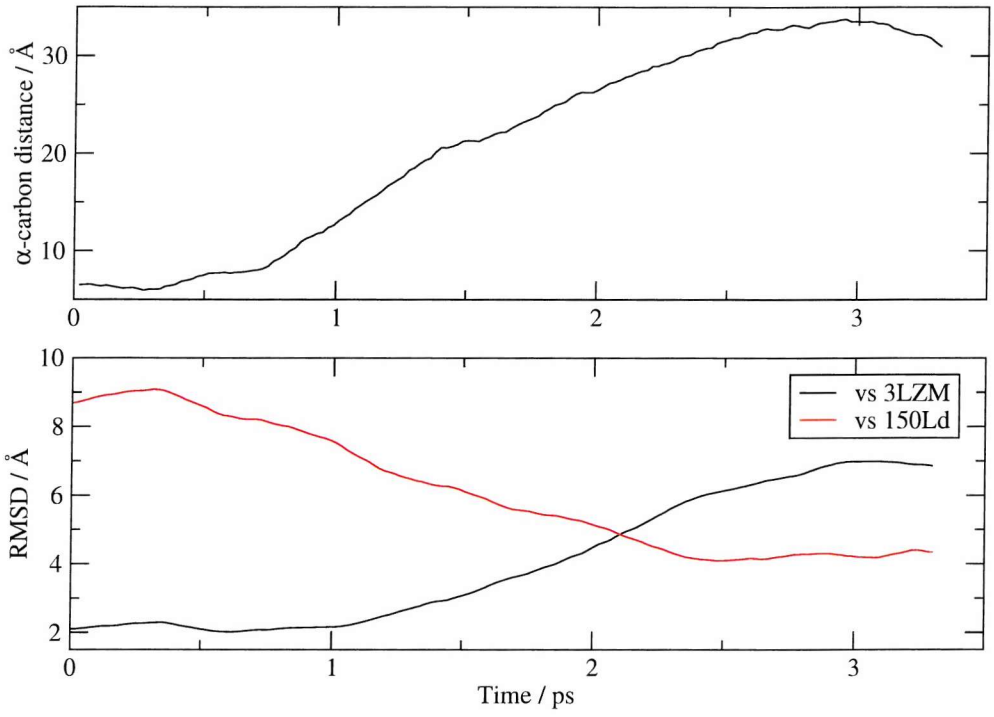
increases in temperatures, of 200 to 300 K, but, once the low frequency energy in the targeted residues has been increased, an amplification factor of 2 can produce increases of over 1000 K. By maintaining an amplification of 700 K, a more gentle protocol is therefore produced for high temperatures, resulting in fewer ω transitions. Many significant motions are seen using an internal temperature cap of 1500 K and several show opening and closing motions.

Again the simulation containing the largest motion shall be validated, although the RMSDs and secondary structures of all simulations have been checked and are acceptable. The largest motion is sampled by an amplification of 700 K by each filter application, and an internal temperature cap of 1500 K. This simulation is shown in blue in Figure 7.28 (b), the scale of which is truncated to show the detail of lesser opening events. The full plot is shown in Figure 7.29 which also includes RMSD and secondary structure analysis. RMSD values show a conformer is achieved closer to the open 150Ld structure, and values are within those seen for previous simulations. The secondary structure analysis again shows limited disruption to the α-helix close to the target residues, but the rest of the protein appears unaffected.

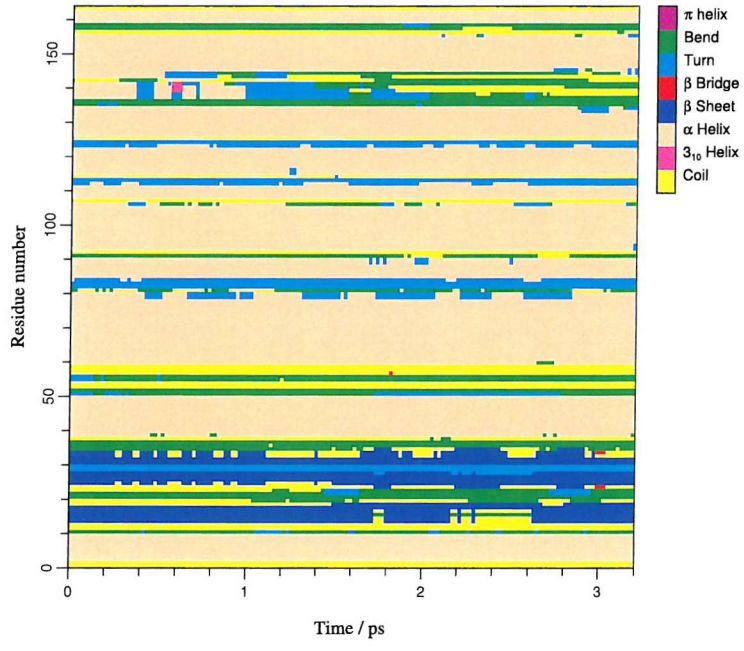
Filter sequences have therefore been used to quickly generate a number of significant conformational events, and are suitable for testing alternative protocols, such as the amplification by a set temperature increase.

### 7.5.3 Further RDFMD work

There are a number of possibilities for the continuation of RDFMD work on the T4 lysozyme system, testing protocols with filter sequences, and then producing simulations of many filter sequences separated by MD for comparison to the long NPT MD simulations previously shown. To demonstrate the potential of further RDFMD work, filter sequences have been performed using a low internal temperature cap of 500 K, and an initial amplification factor of 2. All residues that are not involved in the secondary structure of the equilibrated system have been targeted. No prior knowledge about the nature of the conformational event is




(a) α-carbon distance across bridging residues and RMSD analysis.




(b) Secondary structure analysis.

Figure 7.29: Analysis of the RDFMD simulation with the largest opening motion. An internal temperature cap of 1500 K was used, and the temperature was increased up to this limit by a maximum of 700 K with each filter application.

therefore used.

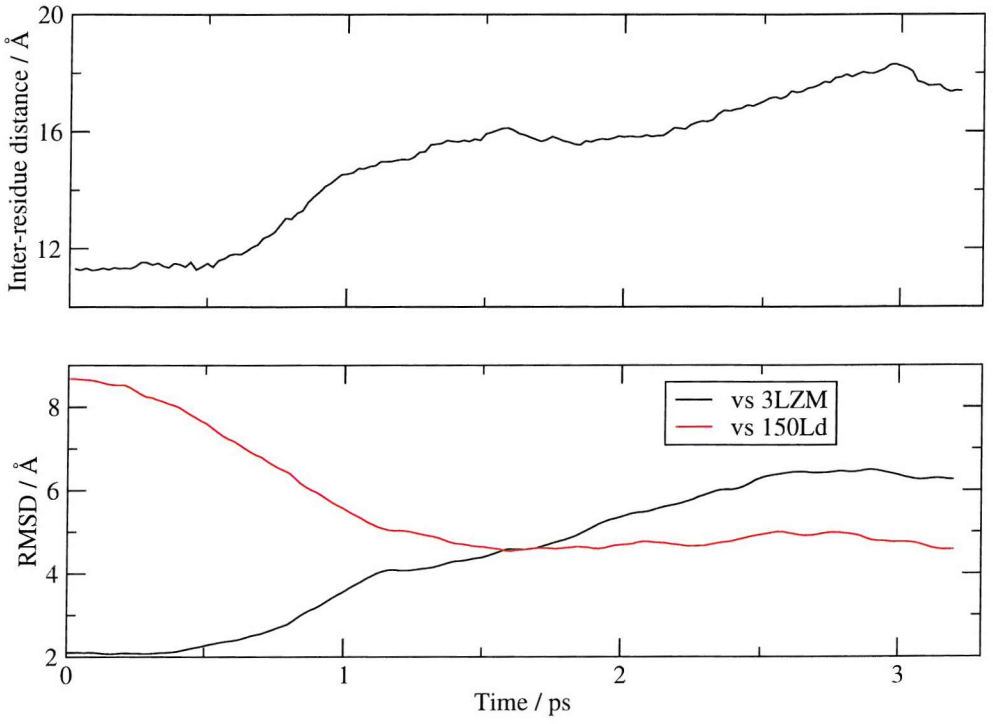
The largest induced motion from ten simulations started from random velocities is shown in Figure 7.30. A significant opening event is seen, reaching a conformation closer to the open 150Ld structure. An amplification factor of 2, applied to so many residues in the system, increases the temperature of the protein more significantly that previous protocols that target a small number of residues. It is unsurprising therefore, that some secondary structure is disrupted, with the loss of the three β sheets in the N-terminal domain. The final conformation is shown in Figure 7.31. More work is required to optimise a protocol that can be applied to an entire protein in this manner, without such significant disruption in secondary structure. However, this encouraging result suggests that such work may be of great use for increasing conformational sampling, with no prior knowledge of expected events.

## 7.6 Conclusions

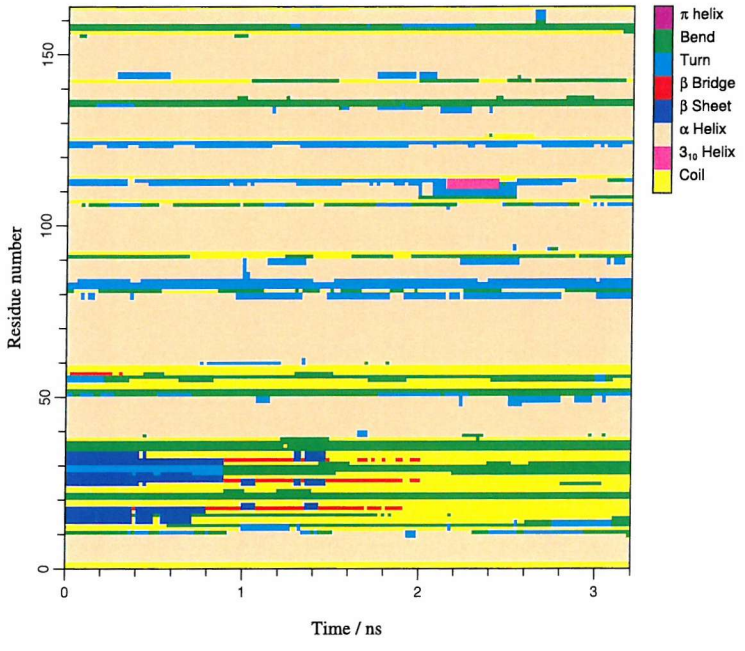
A thorough investigation into the hinge motion of T4 lysozyme and the M6I mutant has been performed, consisting of long MD simulations, thermal simulations at a range of temperatures, parallel tempering, and reversible digitally filtered molecular dynamics.

An overview of all results indicates the presence of a dominant, closed conformation, and a small population of an open state in which cavity residues known to be catalytically important are more accessible. These results are in good agreement with the reviewed experimental work, and disagree with previous, short MD simulations presented by Zhang *et al.* and de Groot *et al.*, who did not identify a distinct closed conformation. The evidence for a closed conformation, and infrequent opening events in this work is undeniable, and it is likely that the differences to past work is due to use of a different force field, or insufficient sampling of the shorter simulations.

The parallel tempering simulations have shown to be producing expected




(a) α-carbon distance across bridging residues and RMSD analysis as previously described.




(b) Secondary structure analysis.

Figure 7.30: Analysis of an RDFMD simulation that targeted all non-secondary structure residues of T4L.

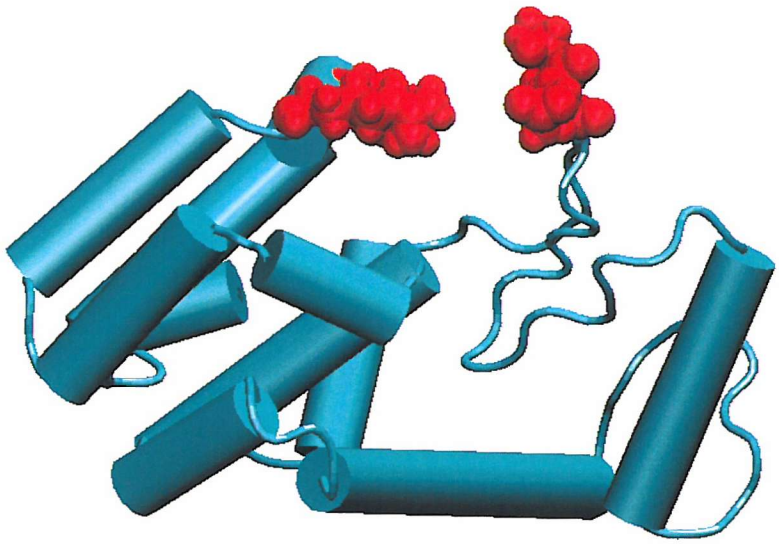


Figure 7.31: The final conformation of an RDFMD filter sequence targeting all non-secondary structure residues. The representation shown is as described in Figure 7.2.

results, but more time is required to converge the computationally demanding simulation. The probability profile is very close to that targeted, and sufficient temperature mobility is obtained within the PT ensemble.

RDFMD has been applied to the WT system in a variety of ways, sampling significant motions using a fraction of the computational expense required to produce the long MD simulations. The protocol generated using the YPGDV system is shown to be applicable to peptide regions of a larger system.

Finally, the possibility of further RDFMD simulations that require no prior knowledge of the protein system have been discussed. A promising result targeting all residues not involved in the protein's secondary structure has been obtained, producing a significant opening event.

# Chapter 8

# E. coli Dihydrofolate Reductase

## 8.1 Introduction

The apoenzyme, dihydrofolate reductase (DHFR), and its cofactor NADPH, catalyse the reduction of 7,8-dihydrofolate (DHF or $H_2F$) to 5,6,7,8-tetrahydrofolate (THF or $H_4F$). THF is essential for the production of thymidylate, a building block of DNA, making DHFR a drug target for inhibiting DNA synthesis in cancer cells. Escherichia coli DHFR (EcDHFR) contains only 159 residues, making it a popular system for experimental and theoretical studies. EcDHFR consists of one domain that can be split into the adenosine binding subdomain (36 to 108), which binds the adenosine portion of NAPDH, and the loop, or major, subdomain (residues 1 to 37 and 107 to 159), so named as it contains three loops of significant size. The loop subdomain is sometimes further split into the N-terminal subdomain (residues 1 to 37), known to bind to DHF, and the large subdomain (residues 107 to 159).

The catalytic cycle in which EcDHFR is involved has been well characterised[145] and involves conformational changes in the M20 loop (defined as residues 10 to 24,[146] 9 to 24,[147] 14 to 24[148] or more specifically as residues 15 to 20,[149] also referred to as loop I). This loop will be the subject of this conformational investigation. It is known to close over the bound substrates, and is mobile in the apoenzyme.[150]

In this chapter current literature on the EcDHFR system shall be reviewed, and

analysis is performed on several trajectories obtained in the presence of different substrates. A range of simulations of the apoenzyme are then presented, including results from molecular dynamics at a range of temperatures, parallel tempering and reversible digitally filtered molecular dynamics.

### 8.1.1 The catalytic cycle of EcDHFR

In 1987, Fierke *et al.* published the kinetic pathway of the EcDHFR catalytic cycle, determined using stopped-flow fluorescence.[145] The reaction can be summarised in five steps, beginning from the EcDHFR - NADPH complex (the holoenzyme). Initially, DHF associates to the complex (forming the Michaelis complex), followed by a hydride transfer step that forms a ternary product complex (EcDHFR - NADP+ - THF). NADP+ then dissociates, and is replaced by NADPH before the rate-determining dissociation of THF.

To further understand this kinetic pathway, an investigation was published in 1997 by Sawaya *et al.*, involving the assignment of structures to steps in the catalytic cycle using X-ray crystallography.[146] The conformation of the M20 loop is extensively discussed, existing in a "closed", "open" or "occluded" state. The closed conformation of the M20 loop dominates the first section of the reaction cycle, wrapping round the NADPH cofactor in both the holoenzyme and the Michaelis complex. The stabilisation of the closed state can be attributed to the binding of the nicotinamide-ribose moiety of NADPH, which interacts with residues 16 to 20 of the M20 loop.[151] On the dissociation of NADP+, the loop occludes towards the THF substrate, forming a $3_{10}$ helix. Sawaya *et al.* note that the temperature factors of several X-ray structures indicate substantial motion of the occluded M20 loop. Movement into the occluded conformation requires significant rearrangement of backbone torsions, and of hydrogen bonding with neighbouring loops. The cycle completes with the binding of NADPH and release of THF, leaving the M20 loop once again in a closed conformation.

Sawaya *et al.* assign structures to the five steps of the catalytic cycle, and to one proposed intermediate. Interpolation is used to construct a movie between

these static conformations.[152] Although the open conformation of the M20 loop is not assigned to any of the six structures, it is considered an intermediate for the binding of DHF and NADPH, and for the release of NADP+. It is also proposed that the occluded M20 loop is only accessible from the closed state via a more open conformation. Sawaya *et al.* clearly indicate the importance of M20 opening motions as vital to the progression of the catalytic cycle.

### 8.1.2 Experimental studies of the M20 loop

The M20 loop is disordered in structures obtained by X-ray crystallography of the apoprotein, and a number of experimental studies have been carried out on this obviously important region of EcDHFR. In 1992, the $^{1}H$ NMR of wild-type EcDHFR, and that of a mutant with residues 16 to 19 replaced by a glycine residue, were compared.[150] The mutation reduced the hydride transfer rate of the reduction reaction from 950 $s^{-1}$ to only 1.7 $s^{-1}$, showing the crucial role of the M20 residues. A slow exchange process, on the NMR time scale, was seen for the WT protein, but this was not present with the mutant system. The study concluded that active residues behind the M20 loop are accessible in the apoprotein, with the M20 loop exhibiting a slow motion that includes an open state. The loop then seals upon binding of the NADPH cofactor, with a polar link between the side chain of asparagine 18 and the carbonyl of histidine 45.

The nature of the slow motion seen for the apoenzyme was investigated in 1994 using $^{1}H$ NMR.[153] The rate of this motion was assigned to be approximately 35 $s^{-1}$. No similar motion was observed in the EcDHFR - folate complex, but a later study using high pressure $^{15}N/^{1}H$ NMR in 2000, confirmed the presence of two conformations of this complex in solution; the occluded and a more open, hydrated state, with a population of approximately 10 %.[154] In 2001, a $^{15}N$ spin relaxation NMR study confirmed fluctuations on the microsecond/millisecond timescale for the generally closed EcDHFR - NADP+ - folate complex, involving the M20 loop and two neighbouring loops in the major domain.[155] This complex is thought to closely resemble that of the Michaelis complex.

A recent study by Osborne *et al.*, of $^{1}H$, $^{15}N$ and $^{13}C$ NMR assignments of the EcDHFR - 5,6-dihydroNADPH - folate (dihydroNADPH is a reduced cofactor analogue), EcDHFR - folate, EcDHFR - NADPH and EcDHFR - NADP+ - folate complexes, reports only the closed and occluded conformations in solution,[156] with no evidence of a significantly populated open state. The publication misleadingly suggests that Sawaya *et al.* argue that only the occluded and closed structures prevail in the catalytic cycle, with the open conformation seen in X-ray crystal structures due to crystal packing contacts. Although only the closed and occluded conformations are assigned to discrete states of the cycle, Sawaya *et al.* discuss the open conformation as, "proposed to be a transient intermediate in loop motion between closed and occluded conformations," "proposed to play a role in protonation of N5 of DHF," and "included in the DHFR movie to demonstrate how the loop may inter-convert between closed and occluded extremes."

NMR studies have therefore confirmed conformational fluctuations of the M20 loop for the apoenzyme, and for a variety of substrates. These motions are linked to the position of the enzyme on the catalytic cycle proposed by Sawaya *et al.*, but whether motions occur only between the closed and occluded states, or between closed, occluded and open conformations, is not clear. The work of Osborne *et al.* suggests no significant proportion of the open conformation exists for various DHFR complexes, although from this the possibility of a transient open state should not be excluded. A number of open X-ray structures have been resolved, and the open conformation of the M20 loop has previously been assumed to be an important component of the EcDHFR catalytic cycle.

### 8.1.3 Theoretical studies of EcDHFR

Experimental evidence for an open conformation of EcDHFR complexes is limited by the possibility of crystal packing effects on X-ray structures, and the lack of atomistic detail of NMR experiments due to the complexity of chemical shift assignments. Theoretical methods therefore offer the potential for significant insight into the EcDHFR catalytic cycle.

Radkiewicz *et al.* published 10 ns simulations of three solvated EcDHFR complexes in 2000: EcDHFR - DHF - NADPH (the Michaelis complex), EcDHFR - THF - NADP+ (the ternary product complex) and DHFR - THF - NADPH (formed prior to the dissociation of THF).[148] Simulations all started from pdb structure 1RX2 (with a closed M20 loop), and the CHARMM22 force field was used with TIP3P solvent and periodic boundary conditions. Extensive discussion showed the generated trajectories to be stable over the time scale simulated. This adds confidence in the force field and protocol that differ little from that used extensively in this thesis. Several conformational changes of the M20 loop were seen, including the open and closed structures. These changes were dependent on the bound substrates, and differences between simulations were noted to be significant, especially considering that the systems differed only in the positions of a few hydrogen atoms. Insufficient sampling is discussed as a deficit of the study and, importantly, an open conformation of the M20 loop was sampled in several simulations that is not found in X-ray structures of EcDHFR. This does not necessarily suggest the conformation is unreasonable, and indicates a stable state as defined by the force field. Unfortunately this conformation is not fully described.

The thermal unfolding of EcDHFR was studied by Sham *et al.* in 2002[149] using molecular dynamics simulations at 300, 400, 500 and 600 K. Solvent effects were modelled using the Effective Energy Function,[157] and potential energies were described using the CHARMM19 force field. The conformation on the M20 loop was not investigated, but the collapse of the adenosine binding subdomain (ABD) was identified as the first step of the unfolding pathway. The stability of this subdomain is therefore monitored during the RDFMD simulations that will be reported later.

In 2004, Lei *et al.* investigated the proposal of Osborne *et al.*, that the open conformation is not required for the transition between the closed an occluded conformations.[158] A combined study using the FIRST[158] and ROCK[158] algorithms is used; FIRST (Floppy Inclusion and Rigid Substructure Topography) identifies flexible and rigid regions of a protein, and ROCK samples conformations for

the flexible region of the protein using small, random changes in torsion angles. Six trajectories of the apoenzyme from the occluded structures are performed, using ROCK to generate conformations that are only accepted if closer to, or slightly further away from, the closed conformation. Results indicate that the open conformation is not required as an intermediate between the closed and occluded states. This study must be regarded as limited due to the restricted flexibility of the protein, and the significant differences between the closed state reached by simulation, to that found in crystal structures. Since the trajectory is generated in a manner that accepts states according to their similarity to a known target, it is not surprising that this target is reached, and the open conformation, which differs significantly from the closed and occluded forms, is not sampled.

A range of further theoretical studies exist that are less relevant to an investigation of the mobility of the M20 loop, including analysis of the protonation steps of the catalytic cycle,[159–162] coupled motions in the protein by a mutation far from the active site,[147,163] the access of water molecules to bound substrates,[164] and the stability of various protein fragments in solution.[165] The work of Radkiewicz *et al.* provides the only recent simulation of significant length on the EcDHFR system that reports the M20 conformation. This work samples closed and open conformations for several EcDHFR complexes, in apparent disagreement with the results of Osborne *et al.*

### 8.1.4 Summary

The catalytic cycle in which EcDHFR participates has been fully described and experimental evidence for conformational motions of the M20 loop has been discussed. It is unclear whether the open conformation sampled in the majority[146] of EcDHFR X-ray structures is due to crystal packing forces, or whether it is an intermediate in the catalytic cycle. The work of Osborne *et al.* reports only closed and occluded conformations are present in solutions of EcDHFR complexes, and a limited study by Lei *et al.* suggests that a direct pathway between these states is possible without an open intermediate proposed by Sawaya *et al.*

However, molecular dynamics simulations of significant length, starting from the closed conformation, sample the closed and open conformations of several EcDHFR complexes.[148] A new conformation is also seen that is not similar to any X-ray structure. It is unclear whether the open and new conformations are sampled due to deficiencies of the CHARMM force field used.

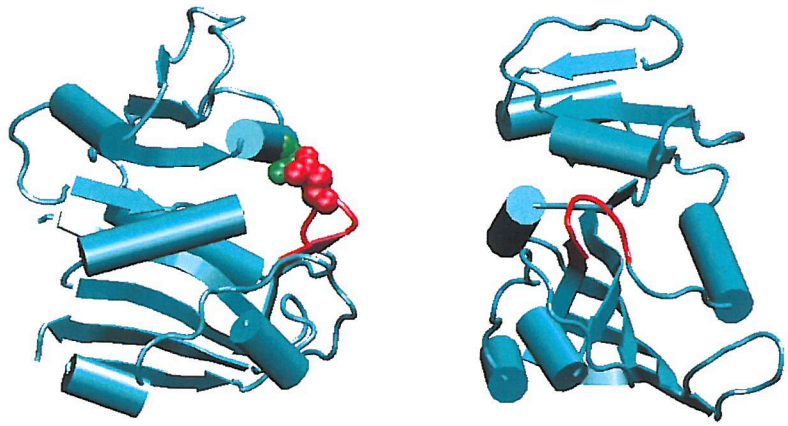
NMR evidence suggests motions of the M20 loop in the apoenzyme,[150] which have not been studied by long molecular dynamics simulations. Apo-EcDHFR shall therefore be a particular target of this investigation, and motions that occur in the M20 loop may indicate the accessibility of conformations inherent to the EcDHFR system.

## 8.2 Molecular dynamics

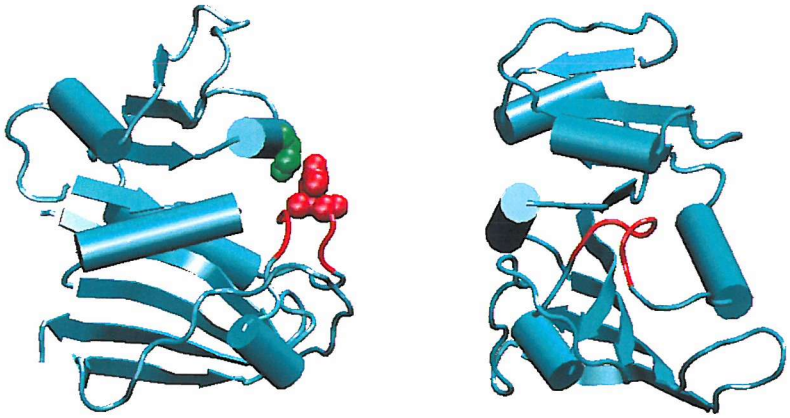
Several trajectories,[166] and equilibrated starting structures,[166] for EcDHFR systems, have been made available for use by the author. All analysis presented here has been performed by the author, and the computational details used to generate the obtained starting structures and trajectories are described in full.

An estimation of how open the M20 loop is can be viewed by the α-carbon distances between asparagine 18 and serine 49, and asparagine 18 and histidine 45. Analysis of simulation trajectories is also shown for the RMSD of residues 15 to 20 of the M20 loop, against three known pdb structures, 1RX2[146] (crystallised in a closed conformation with folate, as a DHF analogue, and NADP+ bound), 1RA9[146] (NADP bound but with the nicotinamide-ribose moiety not in the active site, leaving the M20 loop in an open conformation), and 1JOM[167] (crystallised in an occluded conformation with folonic acid). These structures, and the importance of the asparagine 18 to serine 49 distance, are shown in Figure 8.1. Superposition is performed over all residues involved in secondary structure elements in the closed 1RX2 structure.

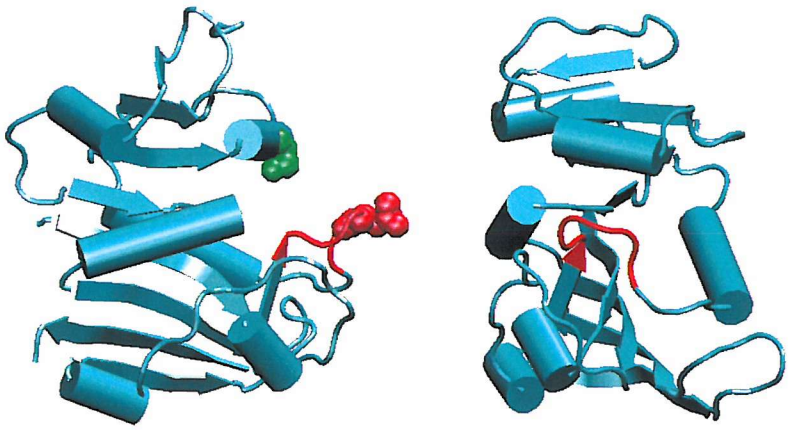
Polar interactions referred to in the literature that describe the closed, open and occluded conformations are also included in analysis. The closed conformation



(a) 1RX2 (closed)



(b) 1JOM (occluded)



(c) 1RA9 (open)

Figure 8.1: Cartoon representation of EcDHFR pdb structures. Substrates have been omitted for clarity. Residues 15 to 20 of the M20 loop are shown in red by the tube representation. The heavy atoms of residues 18 (red) and 49 (green) are shown (when it is useful to do so) by van der Waals radii.

is characterised by the close proximity of the histidine 45 carbonyl to the side chain of asparagine 18[150] and by the short distances between aspartic acid 122 and the backbone of the glutamic acid residues 15 and 17.[147,155] In the occluded conformation, the distance between aspartic acid 122 and the M20 loop backbone is increased, and interactions between asparagine 23 and serine 148 form.[155] The M20 loop in the open conformation forms hydrophobic contacts between methionines 16 and 20.[146]

Where possible, results of simulation trajectories are presented on the same scale for comparison purposes.

### 8.2.1 Computational details

Four simulation trajectories and starting points have been obtained. These are for the EcDHFR - NADP+ - folate complex (an analogue of the ternary product complex, starting from the 1RX2 pdb structure), the EcDHFR - folate complex (an analogue of the binary complex, starting from the 1JOM structure), and for the apoenzyme EcDHFR, starting from both closed (1RX2) and open (1RA9) conformations. Parameters for the substrates are taken from the work of Radkiewicz *et al.*,[148] and the CHARMM22 force field is used for EcDHFR. Histidine protonation states have been assigned according to literature pKa values,[168] with a δ-hydrogen for histidines 45 and 149, and double-protonated histidines for residues 114, 124 and 141.

The setup protocol was based on that used for T4 lysozyme previously presented. Polar hydrogen locations and protonation states were determined using WHAT IF, and remaining hydrogen atoms added using XLEAP. TIPS3P solvent was added using XLEAP giving a minimum of 10 Å between the solute and cuboid cell edge. The systems were neutralised by the addition of sodium ions (twelve for the NADP+ folate complex, ten for the folate complex and eight for the apoenzyme).

Simulation was performed using the NAMD package, with periodic boundary

conditions, a particle mesh Ewald treatment of electrostatics (with an interpolation order of 6), and a switching function applied to Lennard-Jones interactions between 9 Å and the 10.5 Å cutoff. SHAKE was applied to all bonds containing a hydrogen atom, with a tolerance of $10^{-8}$ Å.

The solvated systems were minimised to remove any problems with initial atom placements. Solvent atoms were minimised for 25 000 to 30 000 steps, ions for 200 to 1 000 steps, solvent and ions for 15 000 to 20 000 steps, each substrate individually for 2 500 to 5 000 steps, and the EcDHFR complex for 3 000 to 6 000 steps. Finally the entire systems were minimised for 25 000 to 30 000 steps.

Velocities were randomly assigned at 50 K and each system heated at 50 K intervals up to 250 K, for 5, 7.5, 10, 15 and 15 ps for each temperature, using a 1 fs timestep and a Langevin damping parameter of 10 $ps^{-1}$. NVT equilibration was then performed for 100 ps at 298 K, with a damping parameter of 1 $ps^{-1}$ for the second 50 ps.

The pressure was then adjusted to the desired 1 atm target over 100 ps of NPT MD simulation, using a Langevin piston period of 400 fs, a Langevin piston decay of 200 fs and a 2 fs timestep. Simulations were performed using the final parameter set for a production period of 5 ns.

### 8.2.2 Analysis of molecular dynamics trajectories

#### The EcDHFR - NADP+ - folate complex

Analysis of the 5 ns simulation performed on the ternary product complex analogue is shown in Figure 8.2. The top plot shows the α-carbon distances that indicate the openness of the M20 loop, the upper-middle plot shows the RMSD against known closed (1RX2), open (1RA9), and occluded (1JOM) structures, and the three lower plots show the interactions described in literature that characterise the closed, occluded and open conformations respectively.

The simulation is dominated by a closed conformation, with α-carbon distances

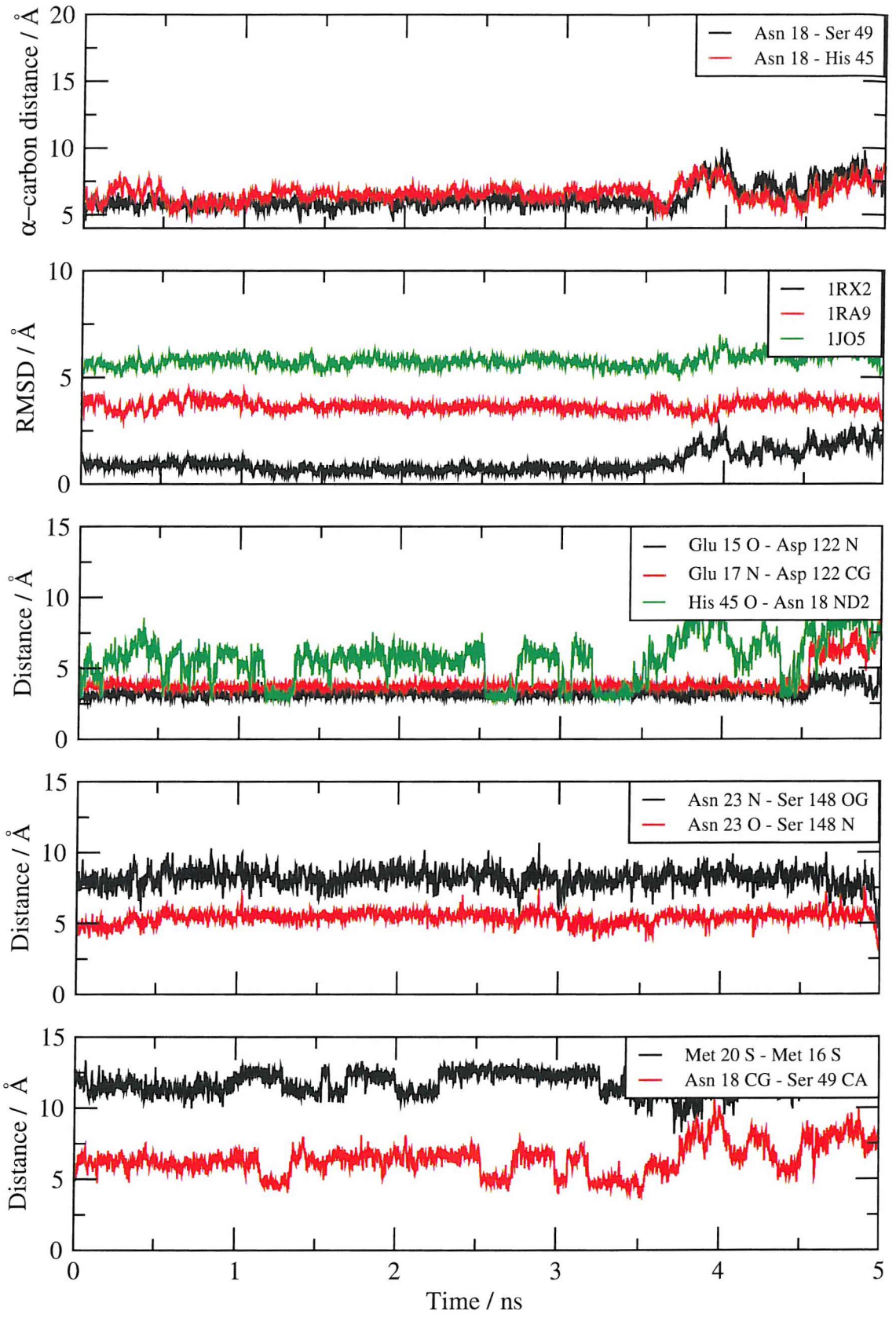



Figure 8.2: Analysis of the obtained 5 ns trajectory of the EcDHFR - NADP+ - folate complex. Top: α-carbon distances that give an indication of the openness of the M20 loop. Upper middle: RMSD of residues 15 to 20 against known 1RX2 (closed), 1RA9 (open) and 1JOM(occluded) structures. Middle: interactions that characterise the closed conformation. Lower middle: interactions that characterise the occluded conformation. Bottom: interactions that characterise the open conformation.

between the asparagine 18 on the M20 loop, and residues 45 and 49 fluctuating between 5 and 7.5 Å. The polar interactions that are known to characterise the closed conformation, those between residues 15 and 17, and 17 and 122, maintain a value of around 3 Å. The histidine 45 - asparagine 18 interaction is intermittent, and is matched by the asparagine 18 - serine 49 distance shown at the bottom of Figure 8.2.

After approximately 3.5 ns, the trajectory moves away from the closed conformation, with an increasing distance between the M20 loop and residues 45 and 49. These distances return towards the closed form briefly at 4.34 to 4.5 ns. After 4.6 ns, residues 15 and 17 move away from aspartic acid 122, breaking the last of the characteristic interactions of the closed structure.

The extent of the opening of the M20 loop seen at the end of the 5 ns trajectory is not clear, and requires more simulation. The trajectory may return to the closed form, sample a relevant open conformation, or the opening may represent an inadequacy of the protein-substrate parameters.

Figure 8.3 shows a secondary structure analysis of the 5 ns EcDHFR - NADP+ - folate simulation which maintains the structure seen in the 1RX2 crystal. Residues 15 to 20 show β-bridge encompassing a turn structure, but this breaks down to a bend during the last nanosecond, corresponding to the opening event discussed.

**The binary EcDHFR - folate complex**

The results of the 5 ns EcDHFR - folate NPT simulation are shown in Figure 8.4. The close polar interactions between asparagine 23 and serine 148 seen in the crystal structure break after 1.7 ns, and do not reform. The RMSD against the 1JOM structure does not increase, suggesting the change is localised to loop residues above 20. Again it is not clear whether the loss of these characteristic interactions indicates a deficiency of the parameter set, a poor representation of the solution conformation in the X-ray structure due to crystal packing forces, or

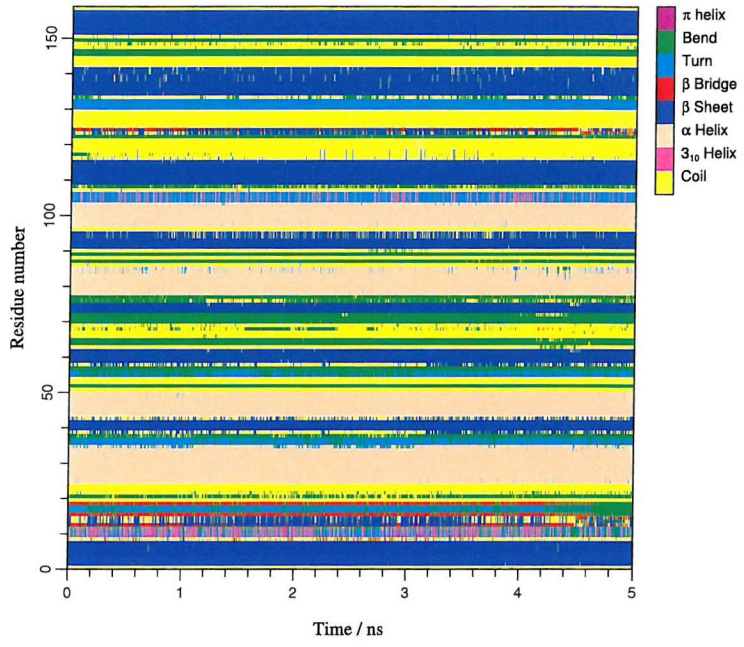



Figure 8.3: Secondary structure analysis of the 5 ns NPT MD simulation of the EcDHFR - NAPD+ - folate complex.

an inadequate time scale of simulation.

The characteristics that describe the closed conformation shown in the middle plot of Figure 8.4 are very different to the simulation with NADP+ complexed. The RMSDs against known structures clearly differentiate the simulation as being in an occluded conformation.

The secondary structure of the 5 ns EcDHFR - folate simulation is shown in Figure 8.5. The 1JOM crystal structure shows a short helix within the M20 loop which is replaced by turn residues in simulation. This represents a small movement of the loop residues away from the bound substrate. However, the 1JOM structure was crystallised with folonic acid rather than folate, which may account for this difference.

### Apoenzyme starting from a closed conformation

Figure 8.6 shows the results of the 5 ns NPT simulation of the apoenzyme, starting from the closed 1RX2 structure. Initially a closed conformation, similar to that

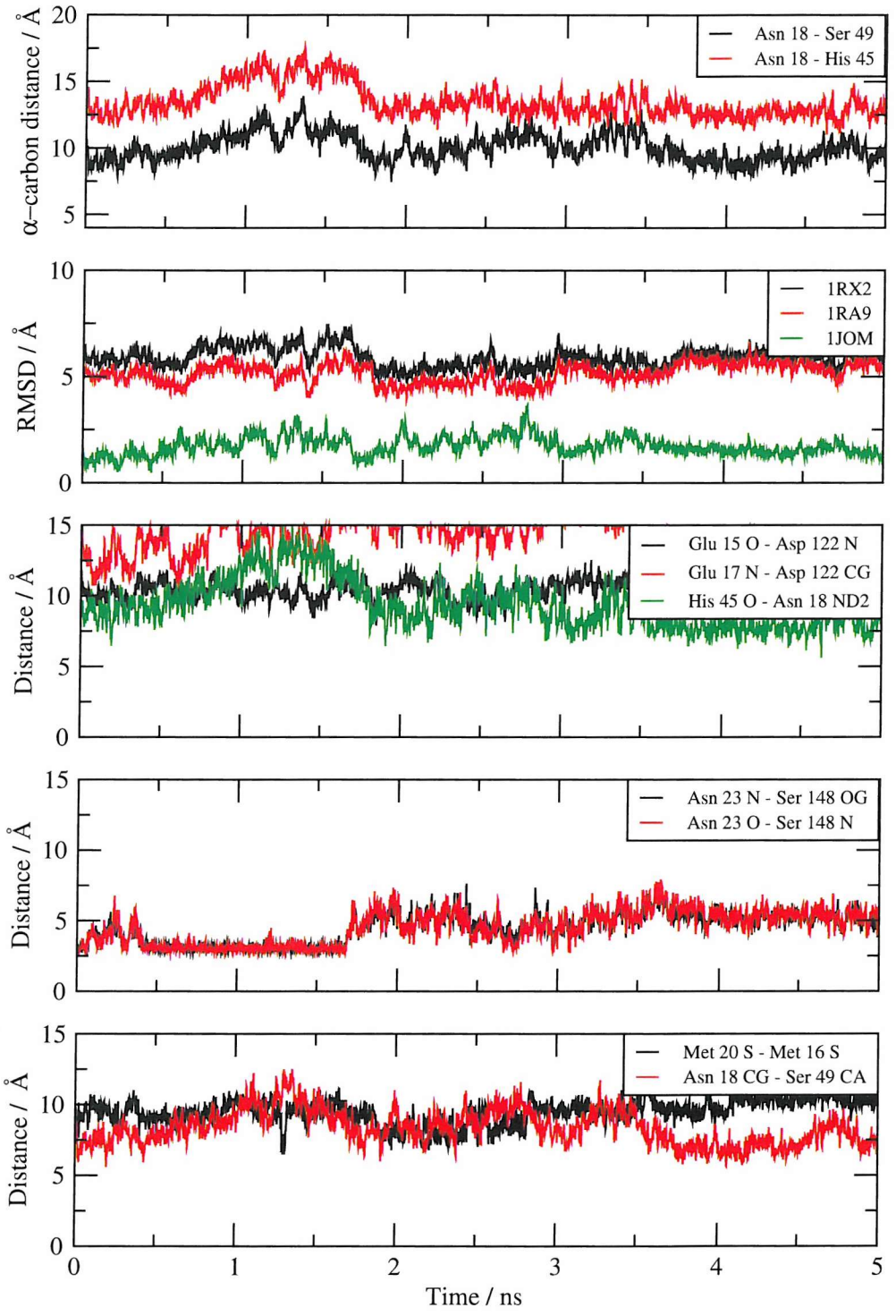



Figure 8.4: Analysis of the obtained 5 ns trajectory of the EcDHFR - folate complex. Top: α-carbon distances that give an indication of the openness of the M20 loop. Upper middle: RMSD of residues 15 to 20 against known 1RX2 (closed), 1RA9 (open) and 1JOM(occluded) structures. Middle: interactions that characterise the closed conformation. Lower middle: interactions that characterise the occluded conformation. Bottom: interactions that characterise the open conformation.

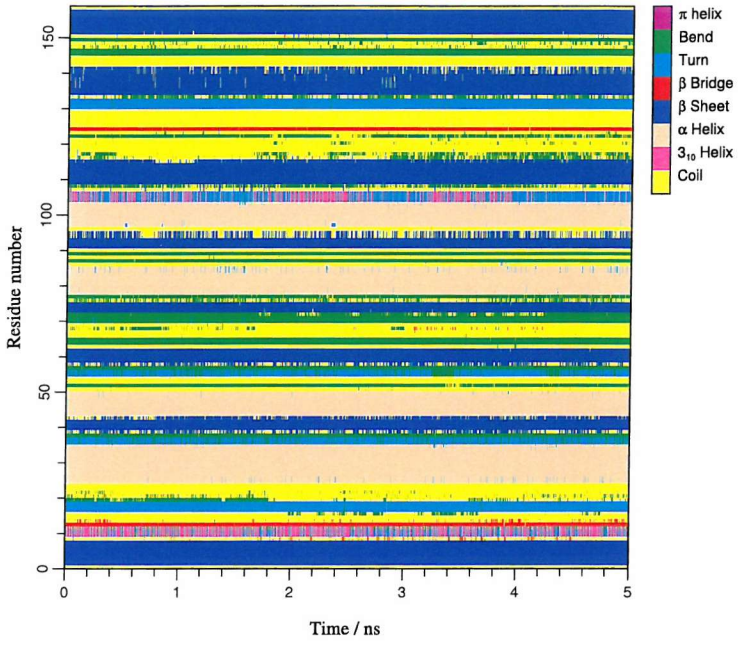



Figure 8.5: Secondary structure analysis of the 5 ns NPT MD simulation of the EcDHFR - folate complex.

seen in the majority of the EcDHFR - NADP+ - folate simulation, is sampled. Between 0.8 and 1.1 ns, a brief opening event occurs, shown in the α-carbon distances, and the asparagine 18 - serine 49 distance. The opening occurs again for a longer period between 1.6 and 2.85 ns, before returning to the closed conformation, which remains for the duration of the simulation.

Interestingly, the opening event does not include a decrease in RMSD of residues 15 to 20 against the 1RA9 structure. The distances that define the opening suggest a discrete conformer that may be similar to that seen by Radkiewicz *et al.* However, there is insufficient evidence to confirm this.

The secondary structure analysis of the apoenzyme shows little difference to that of the closed EcDHFR - NADP+ - folate trajectory, although the presence of the β-bridged turn is only intermittent. This analysis is not shown.

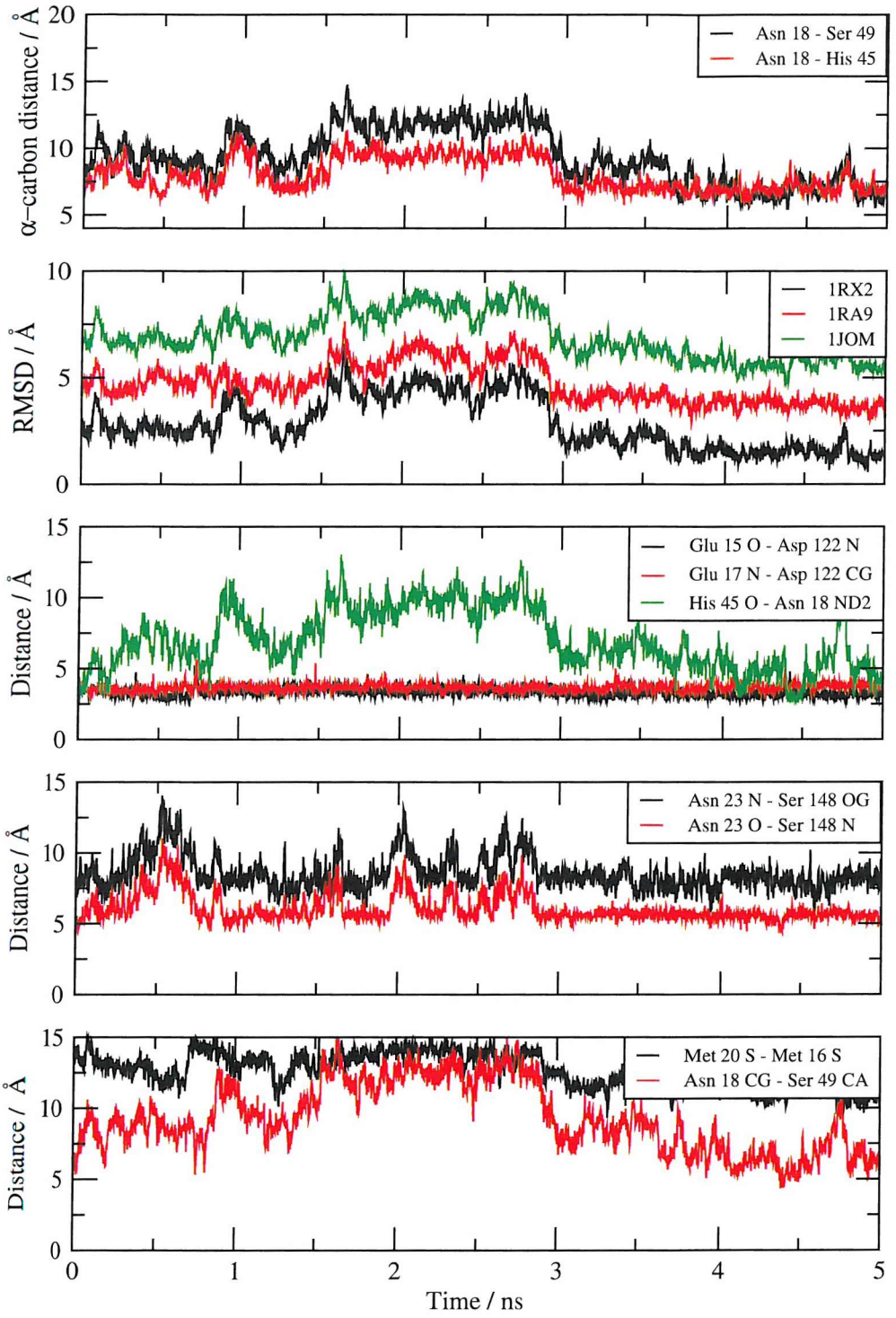



Figure 8.6: Analysis of the obtained 5 ns trajectory of EcDHFR starting from the closed 1RX2 structure. Top: α-carbon distances that give an indication of the openness of the M20 loop. Upper middle: RMSD of residues 15 to 20 against known 1RX2 (closed), 1RA9 (open) and 1JOM(occluded) structures. Middle: interactions that characterise the closed conformation. Lower middle: interactions that characterise the occluded conformation. Bottom: interactions that characterise the open conformation.

**Apoenzyme starting from an open conformation**

The results of the apoenzyme simulation starting from an open conformation are shown in Figure 8.7. A significant closing event occurs over the first nanosecond, shown particularly in the 10 Å closure of the asparagine 18 side chain to the serine 49 α-carbon. The RMSD against the 1RX2 structure falls, until the structure is more similar to the closed than the open. At 2.4 ns, an opening occurs, returning to a structure more similar to the open conformation for the remainder of the simulation.

One interesting result is that the sulphur atom distance between methionine residues 16 and 20, which is 4.11 Å in the open 1RA9 structure, begins around 10 Å, and only falls below 5 Å towards the end of the simulation. The final conformation clearly shows characteristics of both the closed and open states.

The secondary structure analysis of the apoenzyme simulation shows similar characteristics to those seen in closed conformations from previous simulations (not shown).

### 8.2.3 Summary

The analysis of the 5 ns simulations show the closed and occluded conformations as dominant in the EcDHFR - NADP+ - folate and EcDHFR - folate simulations. Each show opening events, but it is not clear whether these are reversible, or represent an inadequacy of the substrate parameters. Further simulation is therefore required.

The closed apoenzyme shows interconversion between discrete closed and open states. However, the open state is not similar to the open conformation seen in the 1RA9 crystal structure, and may be similar to that seen by Radkiewicz *et al.* The simulation from the open structure shows the 1RA9 starting conformation to be unstable, although there is insufficient sampling to conclude whether the proximity of the methionine residues is a stable feature in solution.

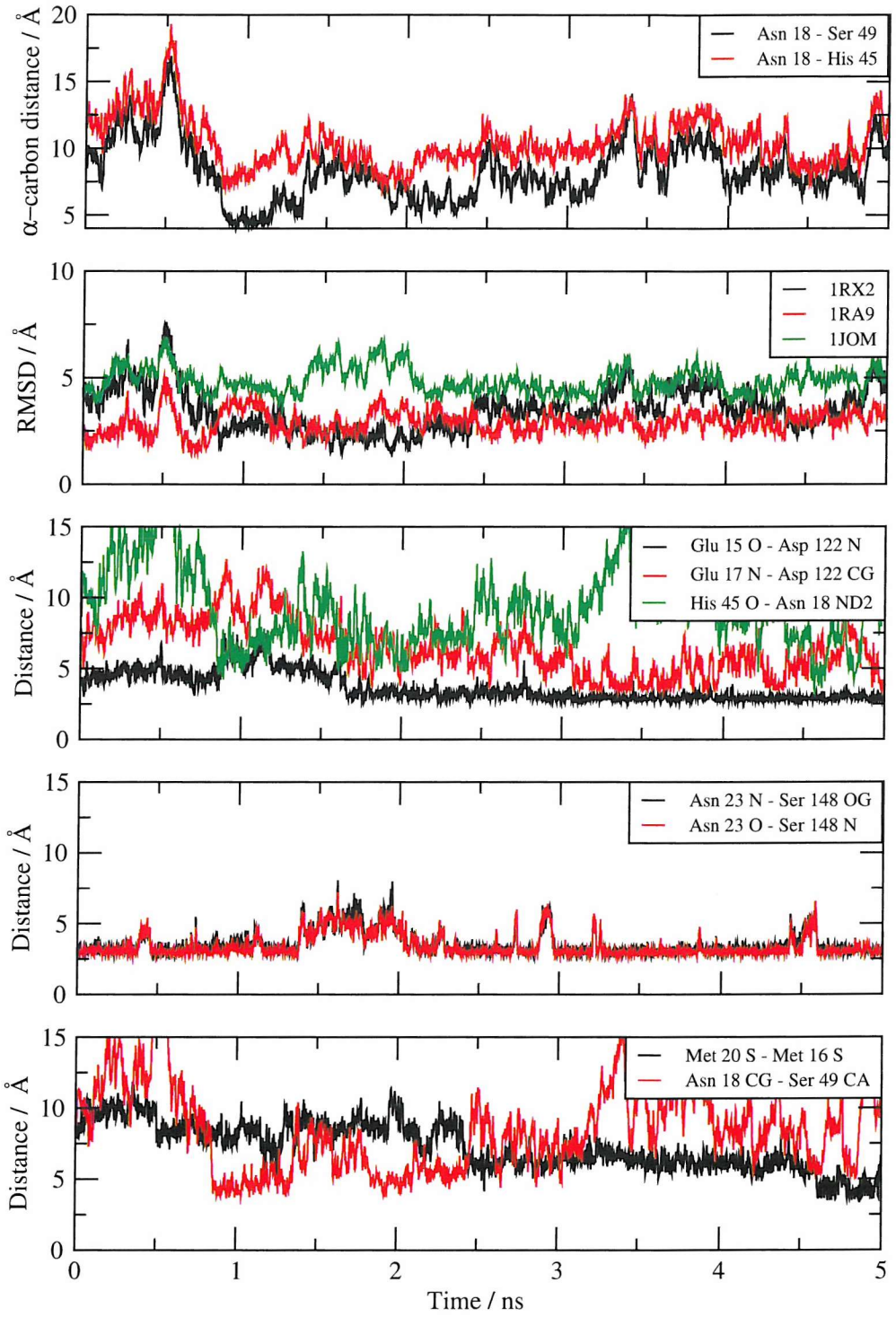



Figure 8.7: Analysis of the obtained 5 ns trajectory of EcDHFR starting from the open 1RA9 structure. Top: α-carbon distances that give an indication of the openness of the M20 loop. Upper middle: RMSD of residues 15 to 20 against known 1RX2 (closed), 1RA9 (open) and 1JOM(occluded) structures. Middle: interactions that characterise the closed conformation. Lower middle: interactions that characterise the occluded conformation. Bottom: interactions that characterise the open conformation.

## 8.3 Thermal simulations of the EcDHFR apoenzyme

Simulations have been performed at 50 K intervals from 300 to 500 K. A Langevin damping parameter of 5 $ps^{-1}$ was used, and initial velocities were randomly assigned and scaled to the desired temperature. 6 ns of NVT molecular dynamics simulation of the apoenzyme is presented for each temperature, starting from both the closed 1RX2 and open 1RA9 equilibrated structures previously described. All simulations included in this, and subsequent sections, are the work of the author.

### Apoenzyme starting from a closed conformation

Full analysis of the 300 K simulation is shown in Figure 8.8. Similar to the simulation of the apoenzyme starting from a closed conformation previously presented, the trajectory samples both a closed, and a more open state. This is shown by the α-carbon distances in the top plot of Figure 8.8, but can also be seen clearly seen in the distance between the asparagine 18 side chain and the serine backbone (shown in the bottom plot of Figure 8.8).

Figure 8.9 summarises the results of the five thermal simulations for EcDHFR starting from the 1RX2 structure. The distance between the side chain of asparagine 18 and serine 49 indicates how open the M20 loop is, and is shown alongside the RMSD analysis of residues 15 to 20, against 1RX2, 1RA9 and 1JOM structures.

The 300, 350 and 400 K simulations show similar results to those seen for the closed apoenzyme simulations previously presented. An asparagine to serine distance between 5 and 7.5 Å is abundantly sampled, indicating a closed conformation, and this opens intermittently to approximately 10 Å. The RMSD against the 1RX2 structure increases upon opening, and returns upon closing, but there is no corresponding decrease of RMSDs against the 1RA9 structure.

The 450 and 500 K simulations sampled more open structures, but the asparagine to serine distance returns to 5 Å before the end of the simulation. When sampling the more open structures, the RMSD against 1RX2 and 1RA9 increases. This suggests reversible opening events of a larger magnitude to those previously

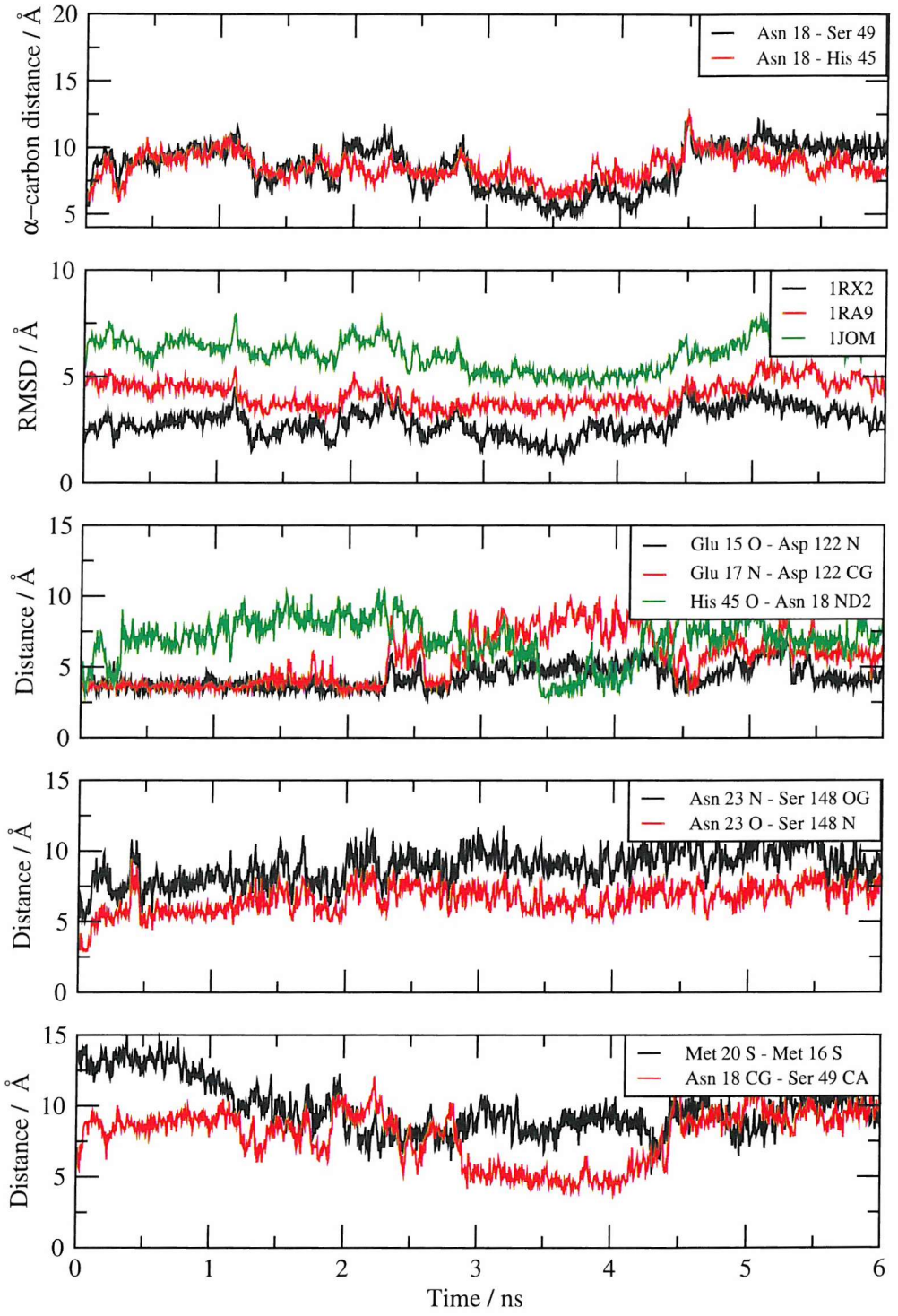



Figure 8.8: Analysis of the 6 ns, 300 K trajectory of EcDHFR starting from the closed 1RX2 structure. Top: α-carbon distances that give an indication of the openness of the M20 loop. Upper middle: RMSD of residues 15 to 20 against known 1RX2 (closed), 1RA9 (open) and 1JOM(occluded) structures. Middle: interactions that characterise the closed conformation. Lower middle: interactions that characterise the occluded conformation. Bottom: interactions that characterise the open conformation.

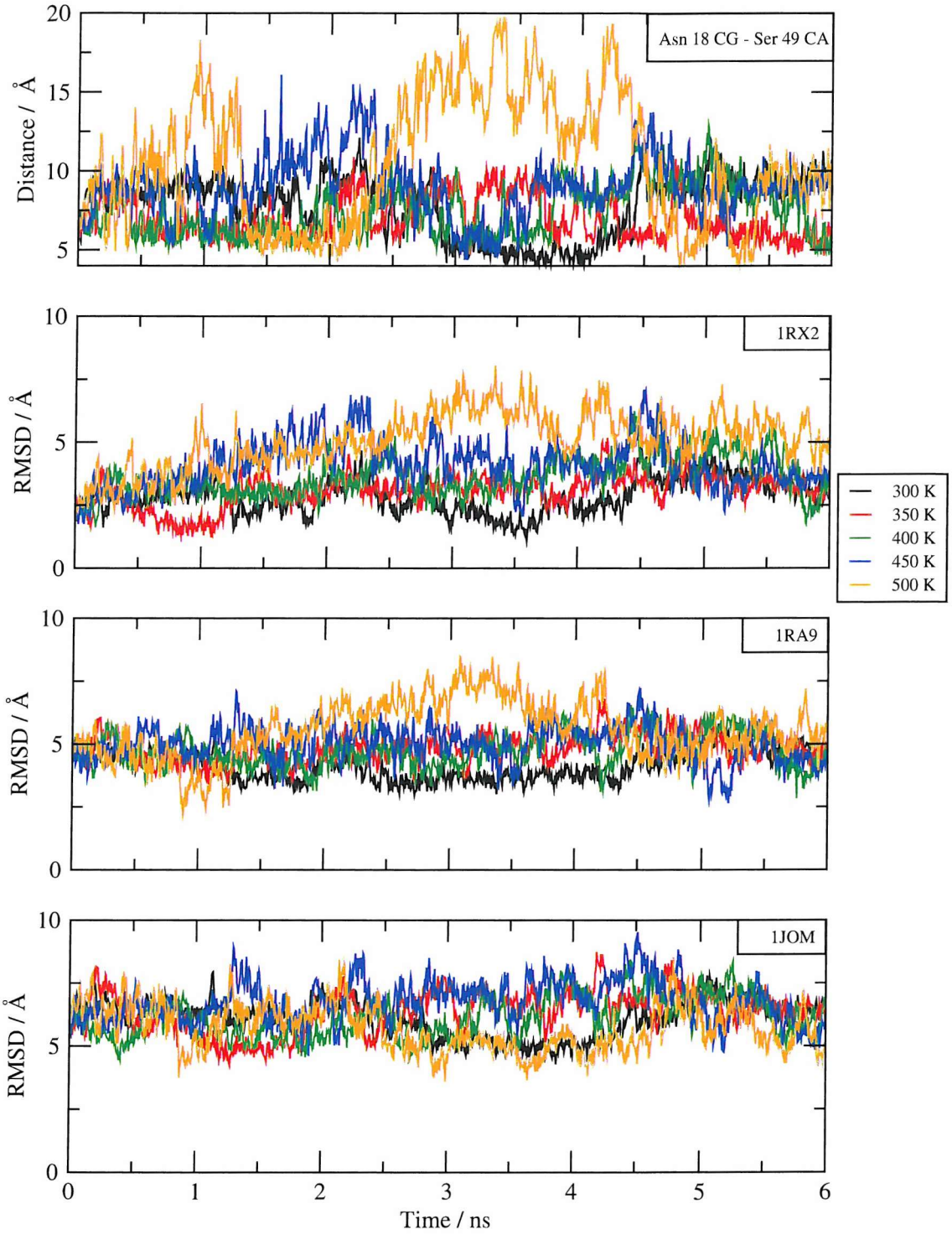



Figure 8.9: Analysis of the 6 ns, thermal simulations at a range of temperatures of EcDHFR starting from the closed 1RX2 structure. Top: Asparagine 18 CG - serine 49 α-carbon distance. RMSD of residues 15 to 20 against the closed 1RX2 structure (upper middle), the open 1RA9 structure (lower middle) and the occluded 1JOM structure is shown.

sampled, still without sampling of the 1RA9 structure.

The secondary structure analysis of the thermal simulations is shown in Figure 8.10. Only in the 500 K simulation is significant disruption seen, with the breaking down of several α-helices. This suggests a high maximum temperature could be used for parallel tempering simulations.

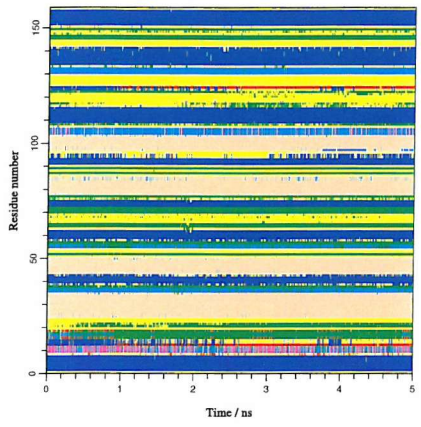
None of the thermal simulations starting from the closed conformation suggest accessibility of the open 1RA9 or occluded 1JOM structures. Sawaya *et al.* suggest that the occluded state may only accessible from the open structure, and so the open structure has also been investigated at raised temperatures.

**Apoenzyme starting from an open conformation**

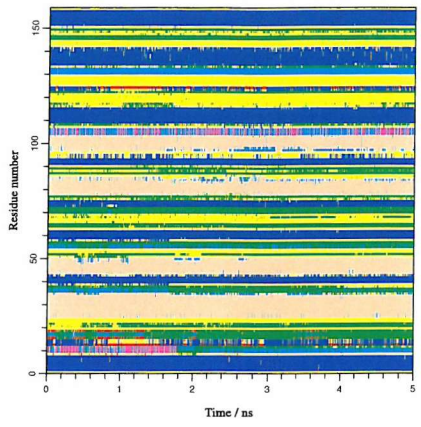
Analysis of the 300 K EcDHFR simulation starting from the open 1RA9 structure is shown in Figure 8.11. A clear closing event occurs after 0.4 ns, initially with the asparagine 18 residue closing towards serine 49, at which point the RMSD against the closed 1RX2 structure falls lower than that against the 1RA9 structure. After 2.2 ns, a further closing occurs, between asparagine 18 and histidine 45. At this point no further decrease in the RMSD against 1RX2 is seen, but the RMSD against the open and occluded structures both increase.

The open conformation is clearly unstable in solution, and the closed state is both readily accessible, and stable once formed. There is no indication of a movement towards the occluded form.

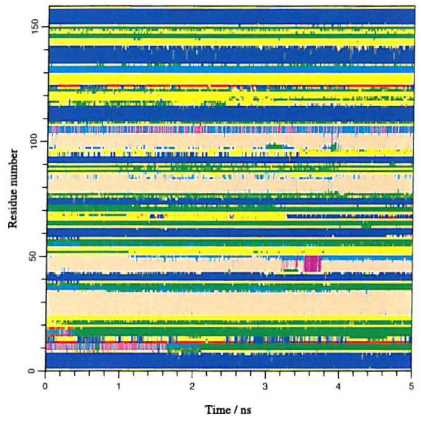
A summary of the thermal simulations starting from the open 1RA9 conformer is shown in Figure 8.12. Although all trajectories initially close, by 6 ns there is a clear separation between the 300, 350 and 400 K simulations that have remained closed, and the 450 and 500 K simulations that reopen. Again there is no evidence to suggest a movement towards the occluded conformation, and trajectories that close and reopen do not do so towards the 1RA9 structure.



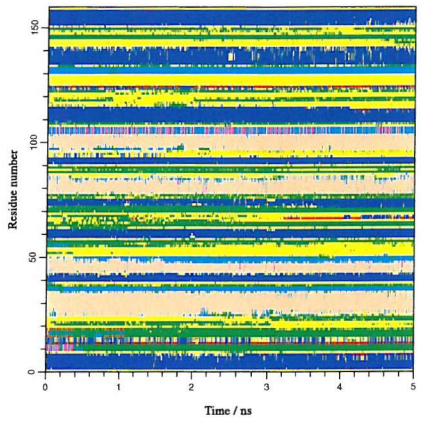

(a) 300 K



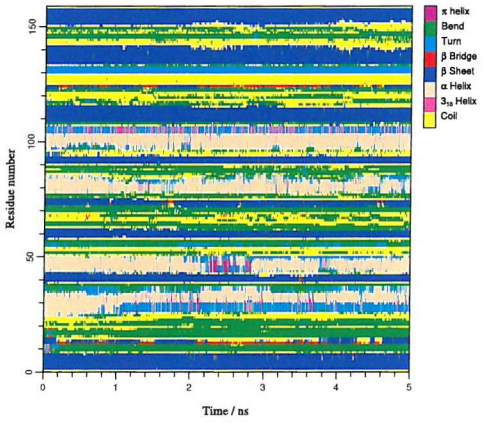

(b) 350 K




(c) 400 K




(d) 450 K




(e) 500 K

Figure 8.10: Secondary structure analysis of the 6 ns NPT MD simulations of EcDHFR, starting from a closed structure, at a range of temperatures.

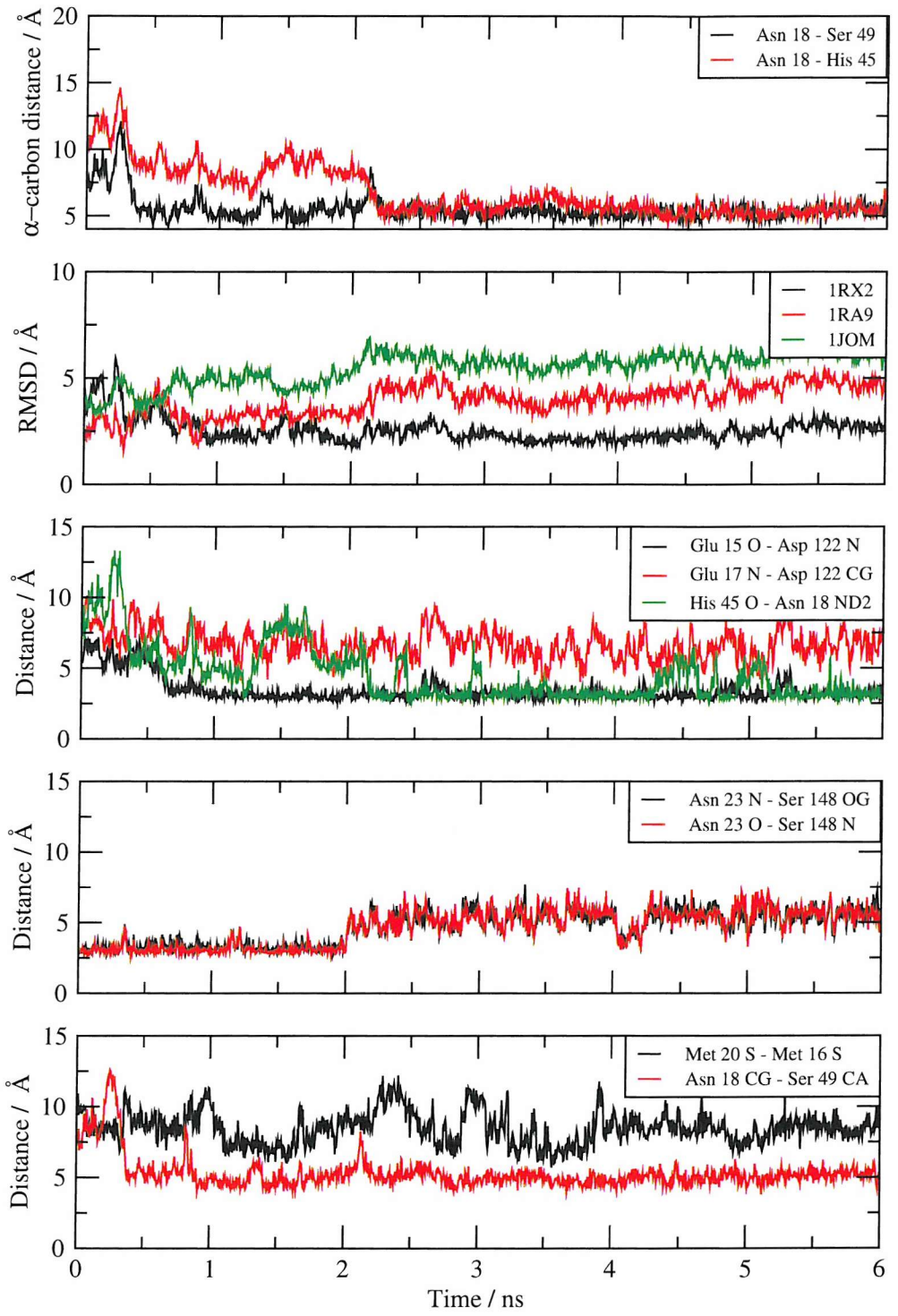



Figure 8.11: Analysis of the 6 ns, 300 K trajectory of EcDHFR starting from the open 1RA9 structure. Top: α-carbon distances that give an indication of the openness of the M20 loop. Upper middle: RMSD of residues 15 to 20 against known 1RX2 (closed), 1RA9 (open) and 1JOM(occluded) structures. Middle: interactions that characterise the closed conformation. Lower middle: interactions that characterise the occluded conformation. Bottom: interactions that characterise the open conformation.

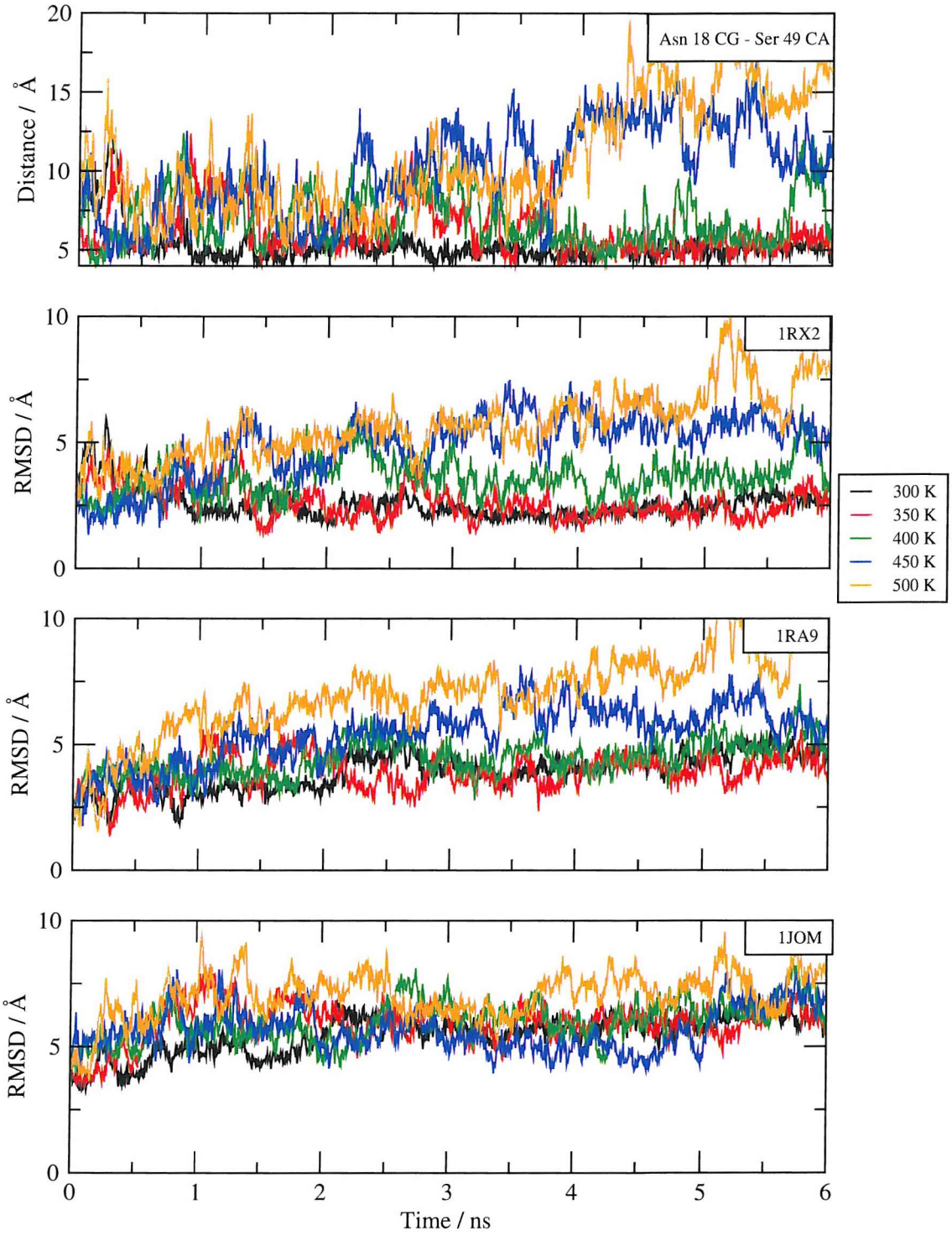



Figure 8.12: Analysis of the 6 ns, thermal simulations at a range of temperatures of EcDHFR starting from the open 1RA9 structure. Top: Asparagine 18 CG - serine 49 α-carbon distance. RMSD of residues 15 to 20 against the closed 1RX2 structure (upper middle), the open 1RA9 structure (lower middle) and the occluded 1JOM structure is shown.

### 8.3.1 Summary

Significant temperatures are clearly required to reach and maintain conformations as open as those seen in X-ray crystal structures. If the occluded conformation is indeed only accessible via the open structure, then sampling a transition to this state during a time scale accessible to simulation may require parallel tempering to be performed to significant temperatures.

However, even at high temperatures there is no suggestion of movement towards an occluded conformation. It seems likely that the transition to this state is triggered by the binding of specific substrates.

## 8.4 Parallel tempering

A parallel tempering simulation of the EcDHFR apoenzyme has been performed, starting from the closed 1RX2 structure. Parameters used are similar to those previously presented, and include a thermostat damping parameter of 5 $ps^{-1}$. PT moves are attempted with alternate neighbouring temperatures every 1 ps.

The PT simulation has been set up to investigate the parallel tempering method, as well as the EcDHFR system. An acceptance probability of 0.3 was used to determine the temperature distribution, based upon previously discussed methods and using the mean and variances from 100 ps NVT simulations at 25 K intervals from 275 to 550 K. However, an additional temperature has been added to the bottom end of the general ensemble, which has been determined using a probability acceptance rate of 0.6. The effect of this on the resulting probability profile will be discussed.

The T4 lysozyme PT simulation was performed up to the temperature at which secondary structure was disturbed. The EcDHFR ensemble is performed above these temperatures, to 550 K, determined as the maximum temperature for which a 2 fs timestep conserves energy for this system. The thermal simulations indicate significant disruption of the secondary structure at these temperatures. It is possible that the unfolding of the protein may produce structures that are unable

to refold within the simulation time scale. These replicas would be unable to adopt conformations with a low enough potential energy to swap to lower temperatures. The system is therefore checked for a reduced mobility of the higher temperature replicas.

The generalised ensemble created contains 91 replicas, spanning the temperature interval 291.1 to 553.6 K.

### 8.4.1 Parallel tempering results

2 ns of PT simulation for the EcDHFR system has been performed, giving a total of 182 ns of NVT simulation. As previously discussed in Chapter 7, the convergence of PT simulations on large systems is beyond the time scale of this project. A target of 10 ns PT simulation is desired, and results shall be reanalysed once this is reached. Results of the 2 ns completed so far are presented here.

Figure 8.13 shows the probability profile of the 2 ns PT simulation. The 0.3 acceptance probability target is well described, and, as with the T4 lysozyme PT simulation, a slight increase in acceptance probability is seen as temperature increases. This is likely to be due to a deficiency in the method of calculating the temperature distribution, but considering the computational expense of investigating temperature distribution calculations (requiring a PT simulation to analyse each alternative method), this small deviation is considered acceptable.

The lowest temperature was determined using a 0.6 acceptance probability. Being on the end of the ensemble, half the tests are automatically failed (since there exists no neighbour with which to attempt a swap), and the 0.324 acceptance probability is close to the expected value of 0.3. The next temperature in the simulation has an acceptance probability of 0.452. This is approximately equal to the average of the acceptance probability used to determine the temperature above it (0.3) and that below it (0.6). No other temperatures appear to be affected. The acceptance probability is therefore seen to behave in the mathematically expected fashion. This suggests that the PT method is sufficiently flexible to experiment with the use of non-exponential temperature distributions in a predictable manner.

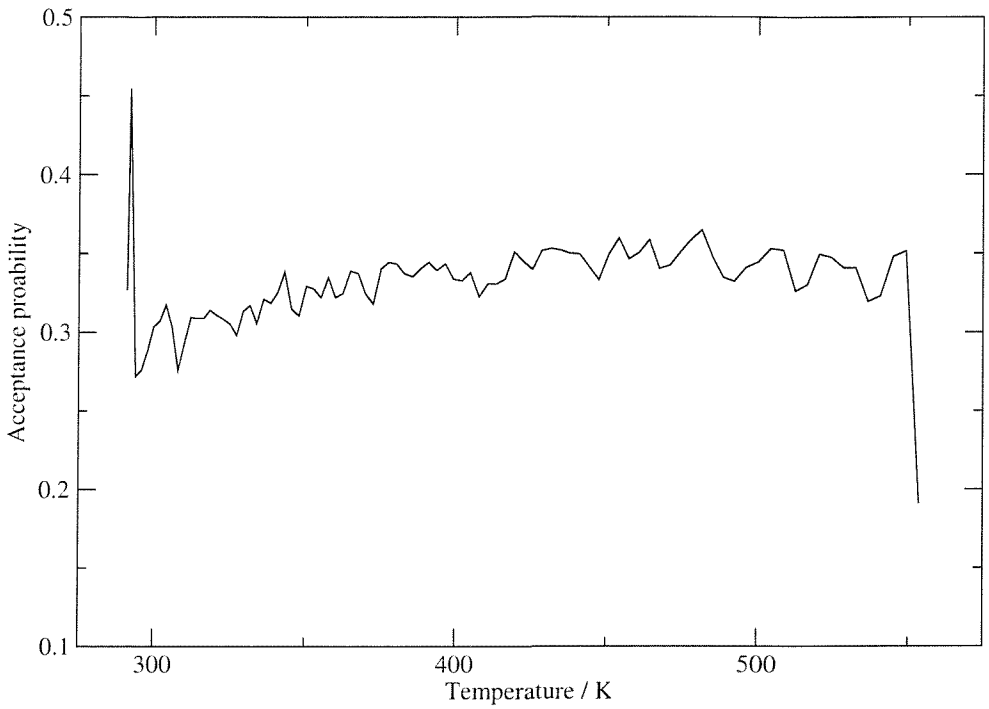



Figure 8.13: Probability profile from 2 ns of EcDHFR PT.

For example, the PT of the T4 lysozyme system was set up to reach temperatures between 375 and 400 K, at which the opening of the molecular hinge is well sampled. Above this, disruption of the secondary structure occurs, and it was considered desirable to avoid this. The simulation could have been set up with a higher probability of acceptance in the 375 to 400 K range, increasing the proportion of replicas in which a conformational change was likely, with predictable effects on the probability profile.

The temperature mobility of replicas initially spaced at 50 K intervals is shown in Figure 8.14. Significant movement in temperature space is seen, but the temperature interval covered by the ensemble (over 250 K) clearly requires further simulation. Replica CN is the only replica to sample both the highest and lowest temperatures.

A summary of the results at a range of temperatures is shown in Figure 8.15. The asparagine 18 to serine 49 distance suggests a bimodal distribution representing the closed (5 to 7.5 Å) and previously sampled more open (7.5 to 11 Å) conformations. The RMSD frequency distributions in Figure 8.15 all show

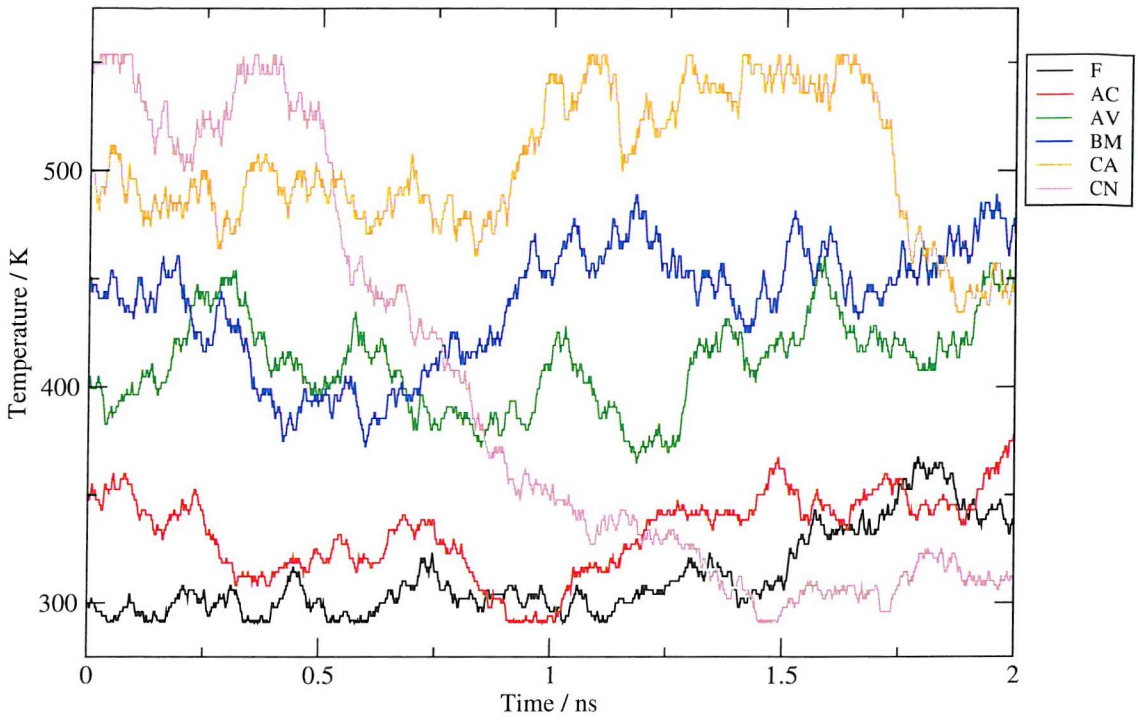



Figure 8.14: Temperature mobility of replicas over 2 ns of EcDHFR PT.

right hand tails corresponding to structures less like any of the known crystal forms. At higher temperatures, a few conformations more similar to the open 1RA9 structure are seen, but the population of these is too low to account for the more open conformations sampled with a 7.5 to 11 Å asparagine 18 to serine 49 distance. There is no evidence of an occluded state similar to that of the 1JOM structure.

### 8.4.2 High temperature replica mobility

The probability profile was approximately uniform, suggesting no restriction of the mobility of high temperature replicas due to unfolding events. This is investigated further in Figure 8.16, which shows the temperatures visited by replicas that occupied the five highest and five lowest temperatures after 1 ns of simulation. There is no evidence to suggest that high temperature replicas are less mobile, and replica CM drops significantly in temperature towards the end of the 2 ns simulation.

The secondary structure of the trajectory sampled by replica CM is shown in

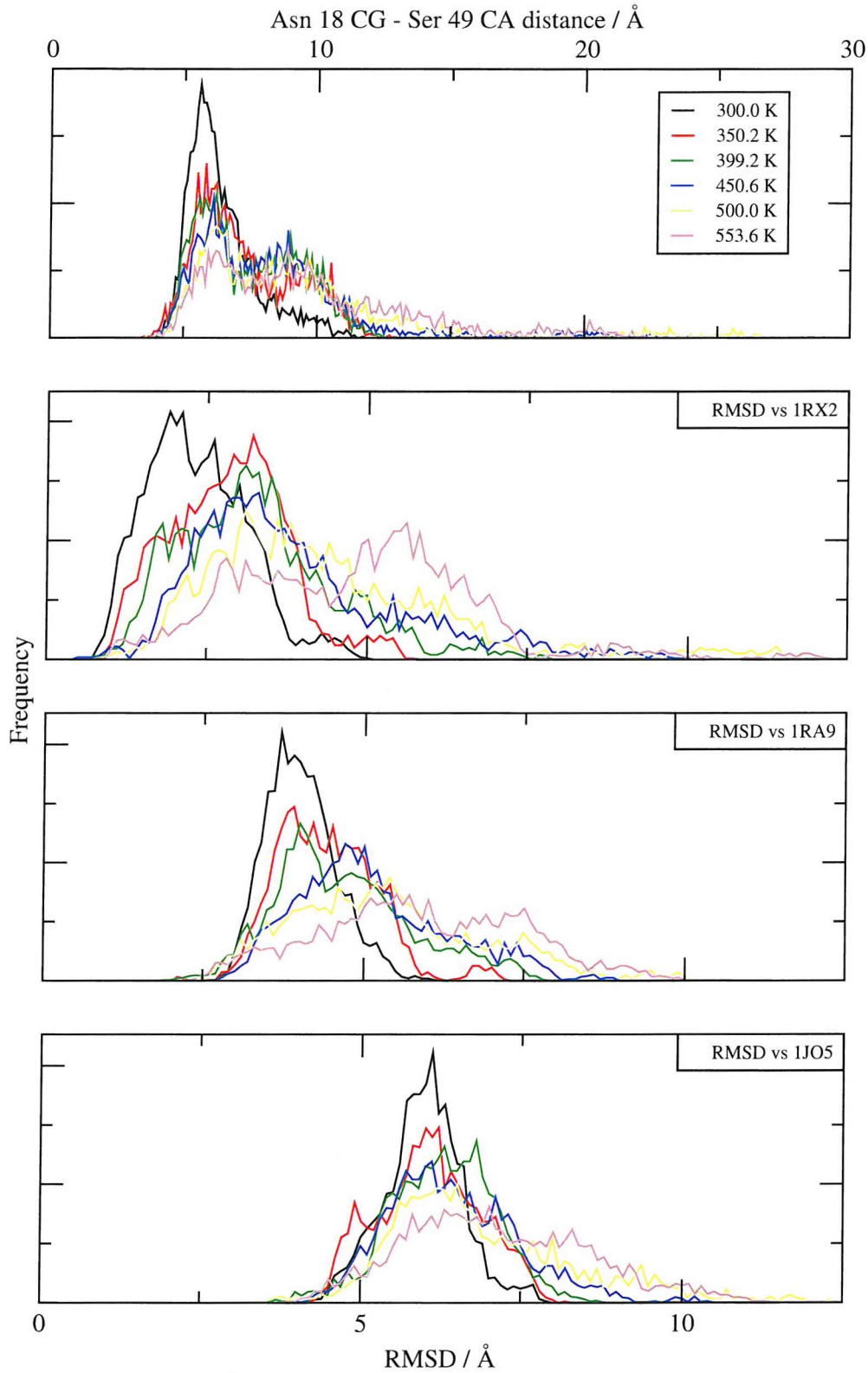



Figure 8.15: Analysis of replicas from 2 ns of EcDHFR PT simulation. Top: Asparagine 18 CG - serine 49 α-carbon distance. RMSD of residues 15 to 20 against the closed 1RX2 structure (upper middle), the open 1RA9 structure (lower middle) and the occluded 1JOM structure is shown.

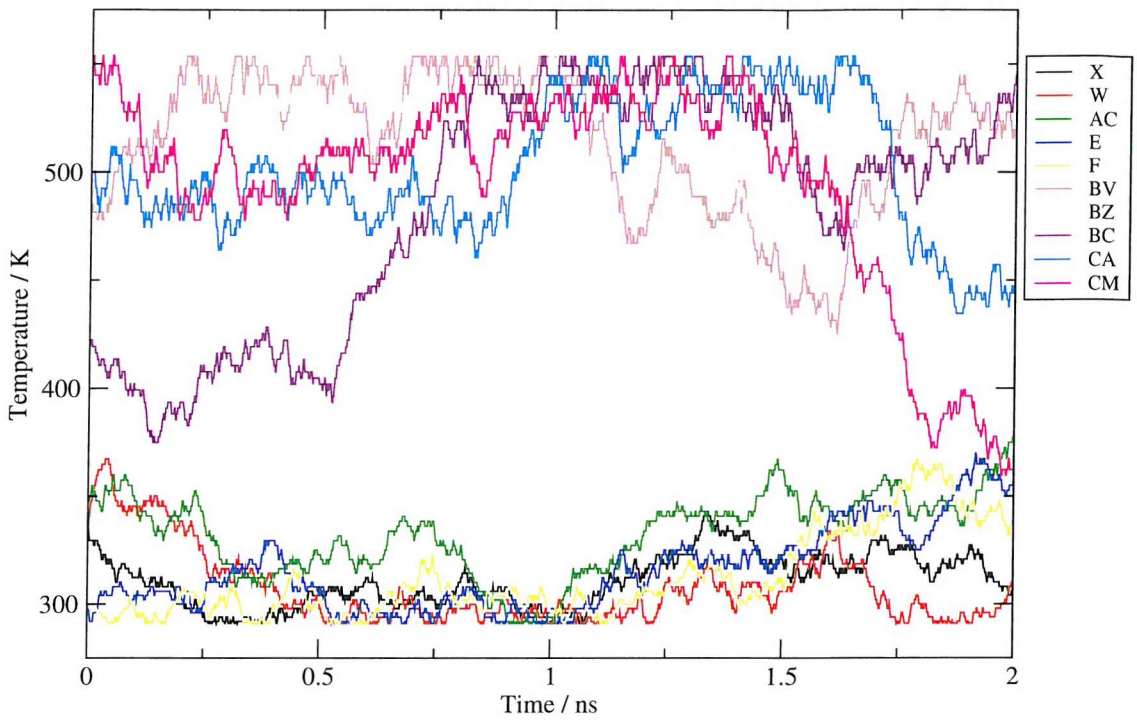



Figure 8.16: Temperature mobility of replicas that occupied the five highest and lowest temperatures after 1 ns of PT simulation.

Figure 8.17. At high temperatures, significant disruption of the helical structure is observed, however this reforms as the replica drops in temperature. This is a positive result for parallel tempering, but the unfolding of the protein is only partial, and more significant denaturing events may not be reformed in the same manner.

### 8.4.3 Summary

The results of the parallel tempering simulation performed so far are in agreement with the molecular dynamics simulations presented earlier in this chapter. The dominant, closed conformation inter-converts with a more open state that is not similar to the open crystal structure. The large temperature interval encompassed by the general ensemble requires further simulation to converge, but there is no suggestion that conformations similar to the occluded or open crystal structures will be sampled.

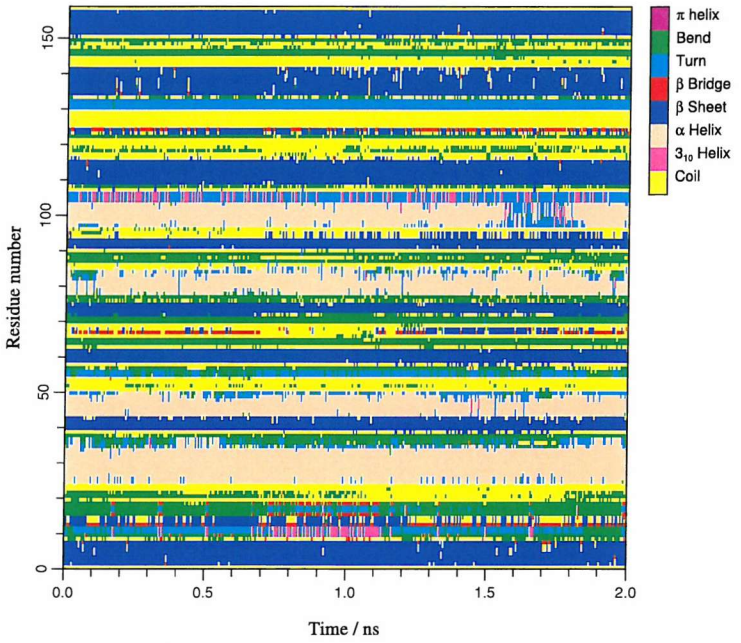



Figure 8.17: Secondary structure analysis of replica CM over 2 ns of PT simulation.

## 8.5 Reversible digitally filtered molecular dynamics

The transition to open and occluded conformations that are abundantly sampled by X-ray crystallography may occur only in the presence of bound substrates. However, it is likely that the conformations are still accessible to the apoenzyme, and RDFMD has been applied to the apoenzyme to investigate this possibility.

### 8.5.1 Selection of a residue target for EcDHFR

A DynDom analysis of the closed 1RX2 and open 1RA9 structures has been performed using all default options, except for the minimum domain size which was decreased to 4 residues. Residues 16 to 19 are located as a dynamic domain, with an effective hinge axis preforming a rotation of 59.1 degrees between the two structures. Residues 15, 16, 19 and 20 are identified as bending residues, and therefore RDFMD has been applied to residues 15 to 20.

Plots of the cumulative amplitude of spectral components for $\psi$, $\phi$ and $\omega$ angles for the target residues are shown in Figure 8.18. Little frequency separation is seen of the $\omega$ angles, and all trajectories presented in this section have been checked to

maintain the all-trans configuration of the M20 loop.

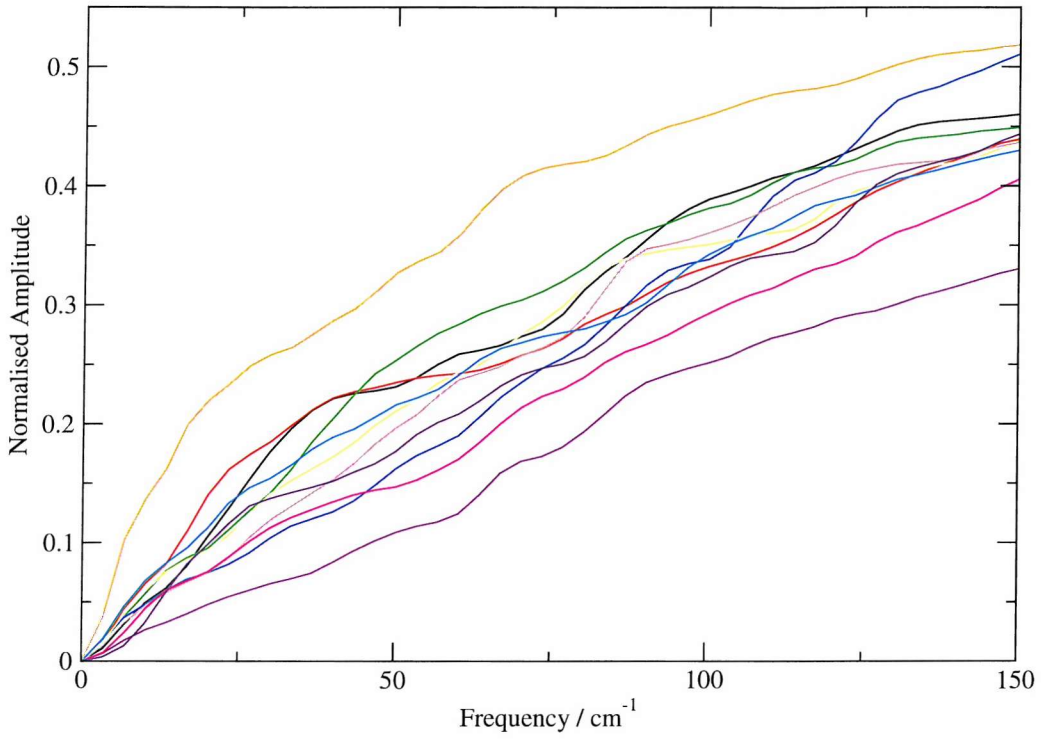
### 8.5.2 Optimised protocol results

RDFMD simulations using the protocol developed in Chapter 5 have been performed on the apoenzyme starting from the closed, 1RX2 structure. The protocol includes the use of a filter delay of 50 or 100 steps, an amplification factor of 2, 4 ps of NPT simulation between each set of filter applications, and a 201 coefficient, 0–100 $cm^{-1}$ digital filter. Simulations have been performed using internal temperature caps of 900, 1100 and 1500 K. The asparagine 18 side chain to serine 49 distance is shown in Figure 8.19 for trajectories that maintain all-trans configurations.

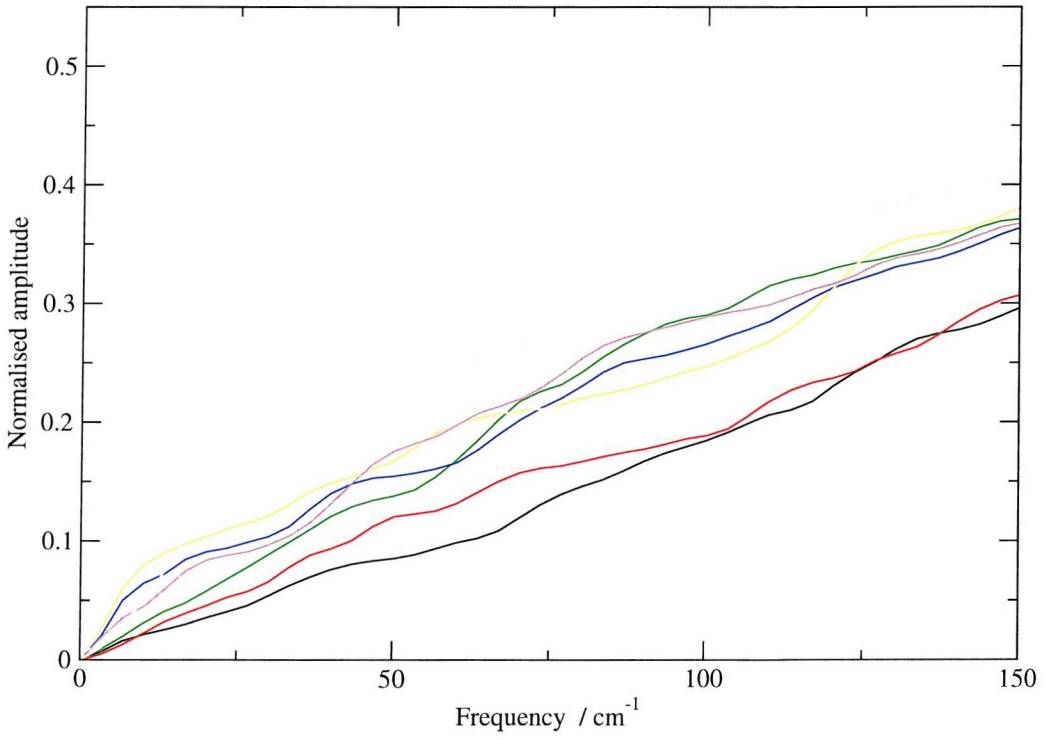
Significantly increased sampling of the M20 loop opening event is seen for all internal temperature caps. Every opening event samples a corresponding closure and interestingly, the internal temperature cap appears to have little affect on the mobility of the system. This suggests that the opening of the M20 loop has a lower energy barrier than the domain motion of T4 lysozyme, that was sampled in a limited fashion with a 900 K internal temperature cap. This is an unsurprising conclusion, considering the molecular scale of the T4 lysozyme motion.

The largest amplitude motion is that generated with a filter delay of 100 steps, and an internal temperature cap of 1100 K. This motion is presented in greater detail in Figure 8.20. The RMSD analysis against known structures indicates the transition to the occluded conformation at 250 ps, sampled for the first time in this investigation. This conformation is characterised by a low RMSD, for residues 15 to 20, against the occluded 1JOM structure, which falls to nearly 2 Å, and by the formation of a turn in residues 16 to 19. These features are similar to those seen throughout the EcDHFR - folate simulation, as shown in Figures 8.4 and 8.5.

Interestingly, the occluded conformation then undergoes a transition to the open state. This conformer has only been previously sampled when used as a starting structure. The transitions from the closed to the occluded and then into the




(a) ψ and ϕ angles




(b) ω angles

Figure 8.18: Cumulative amplitude plots of backbone torsions for residues targeted by RDFMD, and their neighbours. Each colour represents a different backbone torsion.

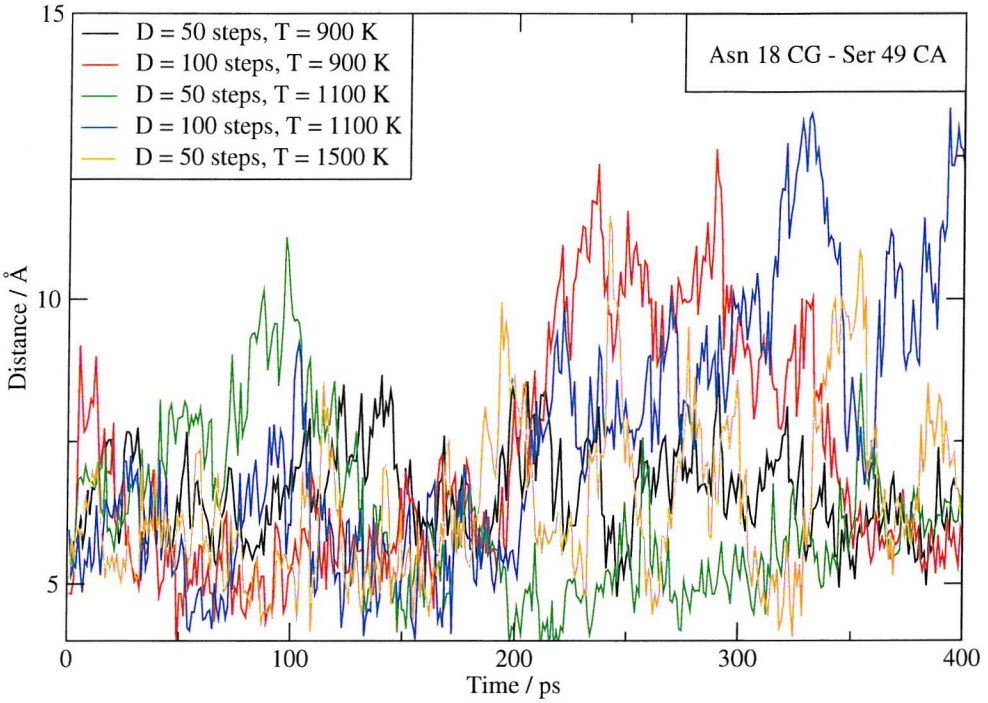
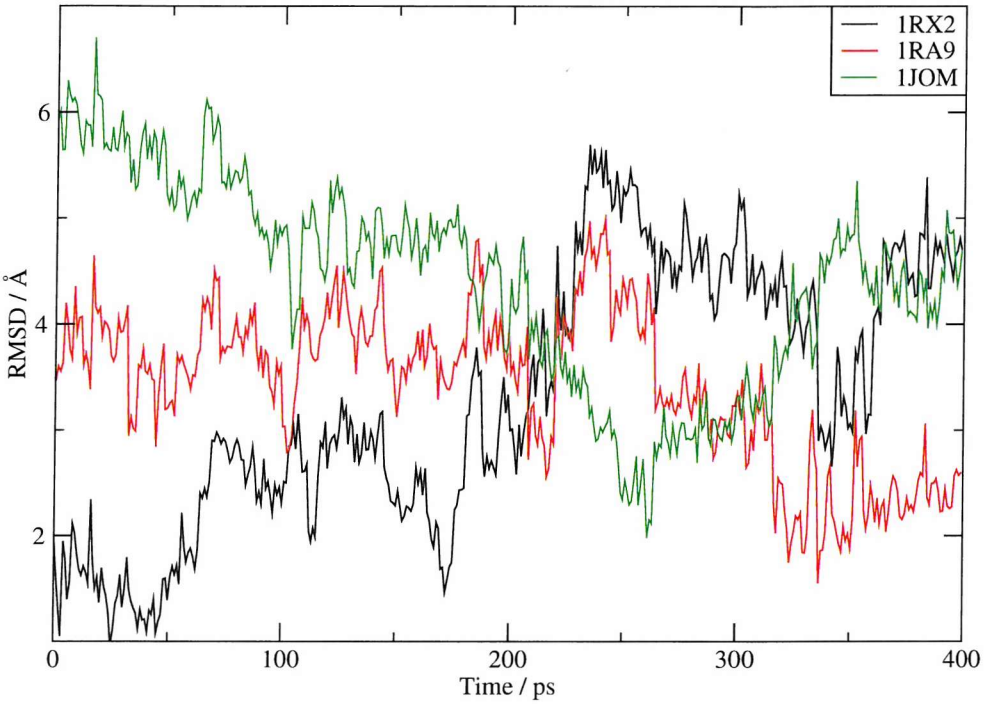


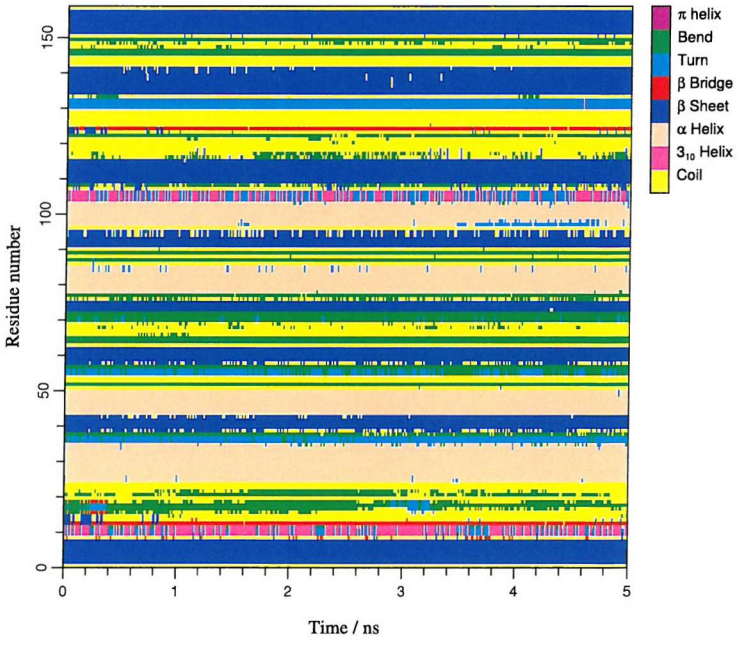

Figure 8.19: Asparagine 18 to serine 49 distance for 400 ps of RDFMD simulation.

open conformation suggests that the proposal of Sawaya *et al.* (that the occluded and closed conformations inter-convert via the open state) is incorrect. Further investigation is needed before this conclusion can be confirmed, but the occluded and closed conformations clearly do not require an open intermediate, as suggested by the study of Lei *et al.* It is also possible that the open conformation is more accessible from the occluded state, than from the closed conformation. RDFMD, or MD simulation, of the apoenzyme, starting from the occluded conformation, could be performed to address these issues.

The RDFMD simulation discussed (with an internal temperature cap of 1100 K and a filter delay of 100 steps) has therefore sampled the three major conformations of the M20 loop. For a comparison to Figure 8.1, conformations closest to the known pdb structures are shown in Figure 8.21. The most closed state occurs at 3 ps, the most occluded at 257 ps, and the most similar to the open pdb, at 342 ps (Figure 8.20).

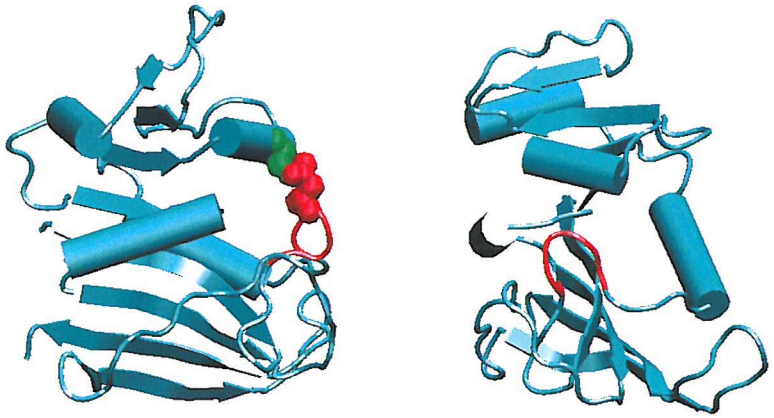



(a) RMSD analysis.

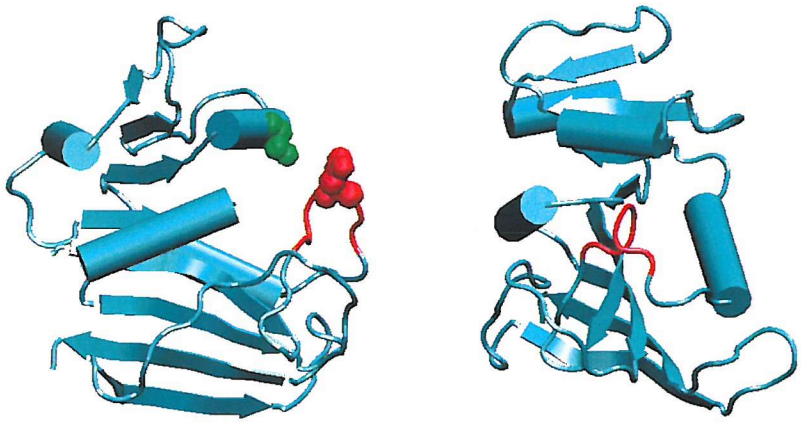



(b) Secondary structure analysis.

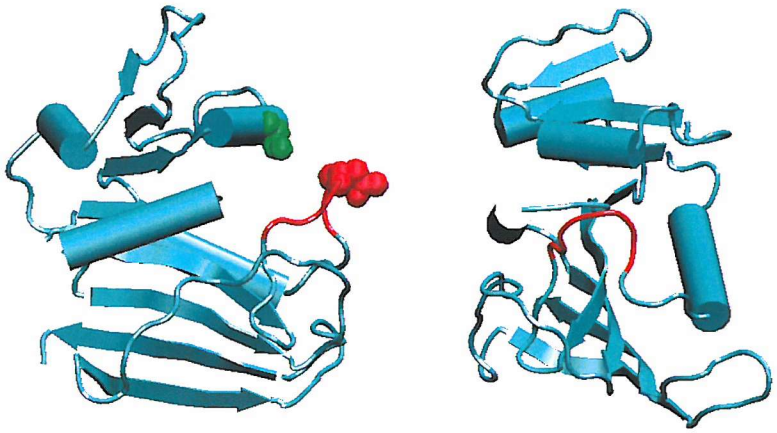
Figure 8.20: Analysis of the WT RDFMD simulation using an internal temperature cap of 1100 K and a filter delay of 100 steps.



(a) Structure most similar to 1RX2 (closed)



(b) Structure most similar to 1JOM (occluded)



(c) Structure most similar to 1RA9 (open)

Figure 8.21: EcDHFR conformations sampled during a 400 ps RDFMD simulation using an internal temperature cap of 1100 K and a filter delay of 100 steps. Representations are as described for Figure 8.1.

### 8.5.3 Summary

RDFMD has clearly shown a significant increase of sampling the conformational states accessible to the apoenzyme, EcDHFR. A transition has been observed from the closed to the occluded state, without an open intermediate. This has not been observed in previous MD studies, and is in agreement with the NMR study of Osborne *et al.*[156] The assumption of Sawaya *et al.*,[146] that an open intermediate is required to move between the closed and occluded conformations, is questioned by this result.

A transition between the occluded and open structures has also been observed using RDFMD. Further work is suggested to study the occluded and open structures, performing RDFMD on the EcDHFR complexes, or on the apoenzyme, starting from the occluded and open conformations.

## 8.6 Conclusions

The investigation of the M20 loop in EcDHFR has included molecular dynamics simulations at a range of temperatures, parallel tempering and RDFMD. Analysis of MD simulations with various bound substrates show well-characterised closed and occluded conformations, although within the 5 ns timescale, apparent instabilities of these conformations are seen. It is not clear whether these are due to the substrate parameters, or if the motions are reversible outside the simulation timescale, and are characteristic of the EcDHFR NADP+ and folate complexes.

MD simulations of the apoenzyme clearly show the closed 1RX2 structure to be dominant in solution. This inter-converts with a less populated, more open state, that is not similar to the 1RA9 structure. Parallel tempering results support this conclusion, but more simulation time is required to fully explore the conformations accessible to the PT ensemble. The more open conformation may be similar to that reported in the simulations of Radkiewicz *et al.*

RDFMD simulations have shown a significant increase in conformational sampling. A transition from the closed to the occluded state was observed, without

requiring a more open intermediate. This is in agreement with the work of Osborne *et al.*, who conclude that the open structure is not seen in a solution of closed and occluded conformers. These results are the first to report the closed to occluded transition in a molecular dynamics simulation, without driving the trajectory towards a known conformation.

A transition to the open state from the occluded conformer was also observed by RDFMD. This open state was similar to that of the 1RA9 structure, as was well characterised by RMSD analysis, and by the wide separation between the asparagine 18 and serine 49 residues. Further RDFMD simulations are proposed to investigate the accessibility of the open and occluded conformations.

# Chapter 9

# Human Immunodeficiency Virus-1 Protease

## 9.1 Introduction

Human immunodeficiency virus-1 Protease (HIV-1 PR) performs a vital step in the lifecycle of HIV, and is therefore an important drug target for the treatment of acquired immunodeficiency syndrome (AIDS).[169] HIV-1 PR catalyses the cleavage of the polyprotein produced by the HIV genome, separating the units that then form the reverse transcriptase, the integrase, and the protease proteins.

HIV-1 PR is a homodimeric enzyme of 198 residues, containing two mobile flaps that cover the active site. For convenience, the residues of the two monomers are numbered 1 to 99, and 101 to 199. The HIV-1 PR active site consists of spatially-neighbouring aspartic acid residues, 25 and 125, at the base of a hydrophobic cleft. Each flap contains two β-sheets (residues 43 to 49 and 52 to 66) connected by a flexible Gly-Gly-Ile-Gly-Gly sequence (residues 48 to 52). Access to the active site requires substantial conformational change in the flap residues and the nature of this motion has been the topic of much research.

In this chapter, recent experimental and theoretical studies of the flap motions of HIV-1 PR are discussed. Molecular dynamics simulations of the apo enzyme at a range of temperatures are presented, and an investigation into flap mobility using

reversible digitally fiiltered molecular dynamics is presented. The computationally expensive parallel tempering method has not been performed on this system, due to the limitation of resources.

### 9.1.1 Experimental studies of HIV-1 PR

The HIV-1 PR system has been extensively studied by experimental methods. Abundant structural data of the apoenzyme and the complex formed with a number of substrate analogues and inhibitors is available from the protein data bank (pdb)[137] and HIV structural reference database (HIVSDB).[170] A recent review of the X-ray data was performed using 73 differing structures,[171] locating rigid regions of the protein (such as the residues that surround the active site) and more flexible areas (such as the "flap elbows" around residue 40). The flap region is seen to exhibit a semi-open conformation in the free protein, and a more ordered, closed conformation in the presence of a ligand. The semi-open and closed conformations are shown in Figure 9.1.

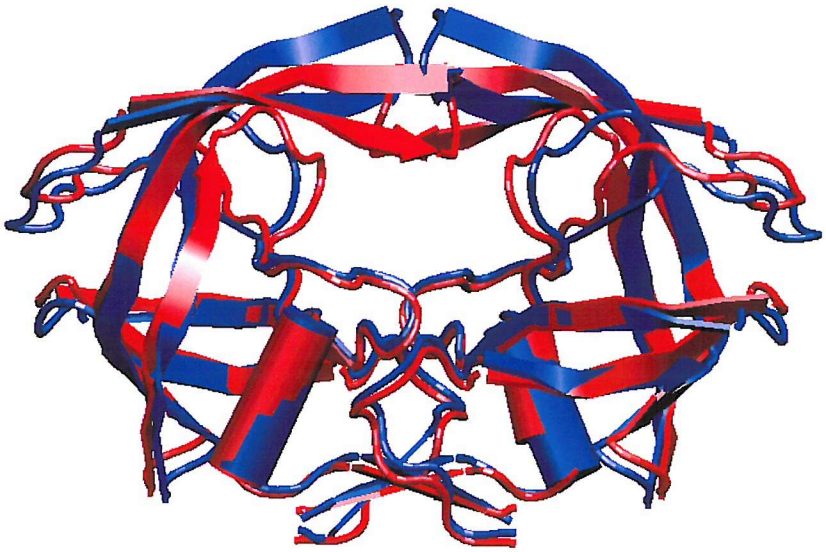


Figure 9.1: Cartoon representation of HIV-1 PR in closed (red, pdb structure 1G6L[172]) and semi-open (blue, pdb structure 1HHP[173]) conformations.

In 1999, a model of the flap mobility of HIV-1 PR was proposed by Ishima *et al.*[174] This was based on a detailed NMR investigation of a mutant that exhibits the same catalytic activity as the wild type protein, but is stable in the concentrations

required for NMR. The free enzyme was proposed to exist mainly in a semi-open state in solution, and two flap motions were detected. One occurs on a timescale well within 10 ns, and involves the flap tip residues, 49 to 52, and the other exists on a timescale of approximately 100 $\mu$s, with less ordered flap residues proposed to be in an open conformation that allows access to the active site. It was also suggested that an approaching substrate could facilitate the opening of the flaps via interation with phenylalanine 53.

An extension of the study by Ishima *et al.*, that includes the effects of a bound substrate, was published in 2003.[175] It was proposed that the association rate of substrates was not controlled by diffusion, but by a rare event, such as a flap opening motion. It is noted that the motions previously observed in the free protein are of limited amplitude in the substrate complex.

### 9.1.2 Theoretical studies of flap mobility in HIV-1 PR

A large number of theoretical studies have been performed on the HIV-1 PR system. One of the earliest molecular dynamics simulations of the free enzyme showed the relaxation of the flap residues from a starting crystal structure over 100 ps using the GROMOS 86 force field in explicit solvent with periodic boundary conditions.[176] The relaxed conformation shows hydrogen bonding connecting the flap tips of the two monomers, and is believed to represent a preferred aqueous structure.

The transition between the closed and semi-open states in the apoenzyme was modelled using a single reaction path method by Rick *et al.* in 1998.[177] Although neither state allows access to the active site, the transition between semi-open and closed conformations could be enzymatically important, possibly describing a closing motion over bound substrates. A loss of β-sheet structure was noted during the simulations, and significant flexibility was reported in the flap tips. Results suggest that the semi-open state exists at a higher potential energy but is stabilised by entropy.

An interesting investigation into implicit solvent models was performed on the HIV-1 PR system by David *et al.* in 2000.[178] Three 1 ns simulations were performed, using the finite-difference Poisson equation (FDPE), the generalised Born (GB) and the distance-dependent dielectric (DDD) methods to describe solute-solvent interactions. The three methods produced very different results, including a collapse of the flap residues into the active site using the DDD method with the apoenzyme.

In 2000, Scott *et al.* reported a significant molecular dynamics simulation of 10 ns length,[143] using explicit SPC/E solvent and the GROMOS96[18] force field. Within 3 ns, a substantially opened conformation is achieved, with isoleucine 50 interacting with the P1 loop, residues 79 to 81, and with isoleucine 47 and 54. This motion is stabilised by the clustering of hydrophobic sidechains, including those of valine 32, proline 79 and proline 81. Within this hydrophobic cluster is a hydroxyl group of threonine 80, proposed to keep the flap regions mobile by destabilising the open conformation. The "flap curling" is asymmetric and exists until the end of the 10 ns simulation. The motion is not sampled in the presence of an inhibitor. Scott *et al.* claim that the simulation results are in agreement with the presence of a motion occuring in a timescale less than 10 ns, reported in the NMR study by Ishima *et al.*. However, the motion described by Ishima *et al.* should be reversibly sampled well within 10 ns, and should include only residues 49 to 52. It seems more likely that the second motion described by Ishima *et al.*, has been sampled in this study. This motion exists on the 100 $\mu$s timescale, and moves to a less ordered, more open state which is considered to be ligand accessible. It is possible that Scott *et al.* have simulated this rare event, but the lack of a reverse path, and the extent of the opening, which has not been previously reported, limits the conclusions that can be taken from this study. The use of the GROMOS force field has also been suggested as a possible cause of the opening motion.[144]

Perryman *et al.* published details of a 22 ns MD simulation of the HIV-1 PR in 2004.[144] The AMBER force field was used with TIP3P explicit solvent. An interesting method of measuring the curling of the flap tips is used, measuring

the "TriCa" angle for the flap tips; that is the angle made between the α-carbon atoms of residues 48, 49 and 50. This angle was sampled from a clearly bimodal distribution suggesting two discrete, inter-converting conformations. The extent of flap opening was measured by the distance between residues isoleucine 50 and the catalytic aspartic acid 25. Simulations began from a closed structure, and sampled the semi-open form. However, simulations of the wild-type protein, and that of a more flexible HIV-1 PR mutant, did not sample the flap opening event reported by Scott *et al.*

A set of simulations were published by Zhu *et al.* in 2003, with a constraint placed on the HIV-1 PR inter-flap distance so that dynamics sample open conformations of the flaps. The free energy profile calculated for the apoenzyme shows an initial barrier to opening, and after an opening of approximately 3 Å between the α-carbon residues, the profile is flat, suggesting a significant flexibility of the open conformation.

Many other theoretical studies of HIV- PR have been performed that are of less interest to an investigation into the flap mobility. These incude studies into the role of hydrogen bonding in catalysis,[179] the folding and stabilisation of the protease monomer,[180] analysis into the first steps of the catalysed reaction,[181] mutation effects,[182,183] and the stability of the dimer formation.[184]

### 9.1.3 Protonation state of the aspartic acid dyad

At the HIV-1 PR active site, aspartic acids 25 and 125 are in close proximity. At pH 7, both acids would normally be modelled in the deprotonated state, but the majority of early studies use a mono-protonated dyad.

A theoretical study by Wang *et al.*,[185] published in 2000, suggests a dianionic state is preferred. The inter-residue distance between the aspartic acid residues was closest to that in the crystal structure when the apoenzyme was simulated in this state using molecular dynamics with an implicit solvent model. However, the MD simulation from which the inter-residue distance was sampled was only

120 ps, and showed little deviation from starting structures for any protonation state. A previous NMR study by Smith *et al* also concluded that the dianionic form existed in solution.[186] The simulations of Scott *et al.* have been confirmed as using a dianionic dyad in the active site (personal communications) and the reported six counterions used by Perryman *et al.* also suggests an unprotonated active site.

Piana *et al.* reported a combination of classical (using an AMBER force field and TIP3P model) and *ab initio* calculations on HIV-1 PR in 2000 and 2002.[187,188] In the free enzyme, the dyad was considered to be almost ionic, with the aspartic acid residues sharing one proton. It is suggested that the ionic state of the proton may account for the interpretation of the NMR data by Smith *et al.*[186] Also, at each step of the proposed catalytic cycle, the aspartic acid dyad exists either in the mono-protonated form, or in a transition state, sharing protons with the substrate or with water molecules. This is in agreement with other ab initio calculations in the presence of inhibitors, that suggest the monoprotonated state is most probable.[189] It is worth noting however, that an *ab initio* study with an alternative inhibitor concludes that the dyad is most stable in a diprotonated form.[190]

There is therefore no clear choice of protonation state of the apoenzyme, which, although is clearly at least monoprotonated in the catalytic cycle, may be dianionic in the absense of a substrate.

### 9.1.4 Summary

The HIV-1 PR apoenzyme is believed to exist in a semi-open structure in solution, and is known to undergo two conformational motions; a limited motion on a timescale well within 10 ns that involves few residues on the flap tips, and an opening event on a 100 $\mu$s timescale (Ishima *et al.*). Scott *et al.* report an opening similar to that occuring on the 100 $\mu$s timescale, within 10 ns of simulation, but it has been suggested that this could be due to use of the GROMOS force field. Perryman *et al.* report 22 ns of simulation which did not sample the large opening event seen by Scott *et al.*, and which seem to be in better agreement with the experimental study of Ishima *et al.*

The protonation state of the apoenzyme is not clear, with short implicit solvent simulations and NMR data suggesting a dianionic aspartic acid dyad. However *ab initio* calculations propose that at least a monoprotonated dyad is required, and it is suggested that the NMR data may have been misinterpreted due to the ionic state of the shared proton in solution.

## 9.2 Molecular dynamics

It is likely that the protonation state of the aspartic acid dyad is important for determining the forces in the cavity of the protein, and these will almost certainly influence the conformational motions of the protein flaps. Initial simulations have therefore been performed with the monoprotonated and unprotonated dyad states. Equilibrated starting systems were produced in collaboration, but all simulations and analysis presented here are the author's work. The set up and equilibration protocol were based upon that used for the T4 lysozyme system, and are presented here in full.

The results included in this section are considered to be representative of the protein's dynamics. However, it is intended to extend the simulations to 20 ns of length, and to duplicate the results, thus increasing confidence in the events sampled. This is beyond the timescale of the current project, and 10 ns of MD simulation is presented for each system.

### 9.2.1 Computational details

The 1HHP pdb structure[173] was used as a starting point for both the monoprotonated and unprotonated active site systems. This structure consists of a single monomer, and a transformation of (x, y, z) to (y, x, -z) is used to generate the symmetrical dimer. The crystal structure was checked, and polar hydrogen atoms were placed, using WHAT IF.[142] All other hydrogen atoms were placed and the structures solvated within the XLEAP utility of the AMBER package. Solvent was

added to give a minimum distance of 12 Å from the protein surface to the cell boundary. The +6 (unprotonated dyad) and +7 (monoprotonated dyad) charges were neutralised by the addition of chloride ions within XLEAP. The CHARMM27 force field has been used with TIPS3P explicit solvent.

Unless otherwise stated, all simulations presented in this chapter have been performed using the NAMD package with cuboid periodic boundary conditions, a particle mesh Ewald treatment of electrostatics (with an interpolation order of 6) and a switching function applied to Lennard-Jones interactions between 9 Å, and the 10.5 Å cutoff. All bonds containing hydrogen atoms have been constrained to equilibrium lengths using SHAKE with a tolerance of $10^{-8}$ Å. Production simulations have been performed in the NVT ensemble (to be consistent with thermal simulations presented in previous chapters), using a Langevin thermostat, with an associated damping parameter. A Nosé-Hoover Langevin barostat (with piston period and decay parameters) has been used during equilibration to control with pressure of the system. A 2 fs timestep is used throughout the equilibration and production stages.

Minimisation of the solvated system was performed, using 30 000 steps on the solvent only, 1 000 steps on the counter ions, 20 000 steps on the solvent and ions, 5 000 steps on the protein, and 40 000 steps on the entire system. The system was then heated in the NVT ensemble with 20 000 steps at temperatures from 50 to 300 K, at 50 K intervals. A 10 $ps^{-1}$ thermostat damping parameter was used to control the system temperature.

NPT simulation was then used to equilibrate the system pressure to a target of 1 atm. 50 000 steps were performed using a decay parameter of 100 fs and a piston period of 200 fs. A further 100 000 steps were then performed using a decay parameter of 300 fs and a piston period of 500 fs.

Production simulations used a 5 $ps^{-1}$ thermostat damping parameter. The equilibrated monoprotonated dyad system contained 9370 water molecules, with cell dimensions of 61.08, 60.98 and 82.37 Å. The unprotonated dyad system

contains 9371 water molecules with cell dimensions of 61.15, 61.05 and 82.47 Å.

### 9.2.2 Molecular dynamics results

It is difficult to describe the flap motions of HIV-1 PR in graphical form, and many analysis methods have been tested. Three conformations are identified in this study, which can be loosely defined by the location of the flap tip isoleucine residues 50 and 150, with respect to important hydrophobic sidechains. How open the structure is can be described using the distance between the two flap tips, and the distances between the flap tips and the catalytic residues (in a simular fashion to that reported by Perryman *et al.*).

The starting semi-open structure, 1HHP, requires the side chain of the isoleucine flap tips to be closely associated with the side chain of the phenylalanine residue on the opposite flap (for example, an interaction between residues 50 and 153). The closed conformation, seen in crystal structures with inhibitors or substrate analogues, is characterised by the two flaps reaching 'past' each other, interacting with the P1 loop (residues 79 to 81) of the opposite monomer. This can be most clearly seen when looking at the distance between the side chains of the isoleucine flap tips and a proline residue in the P1 loop (residue 81 or 181). A third conformation, with the flap tips curled back towards the P1 loop of the same monomer is frequently observed, and can be described using the isoleucine 50 to proline 81 (or residue 150 to 181) sidechain distance.

The results of 10 ns of NVT molecular dynamics are therefore presented for the unprotonated and monoprotonated aspartic acid dyad systems using five graphs. The first graph describes the openess of the structure, displaying the α-carbon distances between the flap tips, and between the flap tips and catalytic aspartic acid residues. The second and third graphs describe the environment of the isoleucine flap tip side chains, generally in a semi-open (associating with the phenylalanine residue in the opposite flap), a curled (with close proximity to the proline in the P1 loop of the same monomer), or a closed (with the flap reaching across the cavity to the proline in the opposite P1 loop) conformation. The second and third graphs

desribe these distances for the flap of each monomer. The fourth and fifth graphs diplay the RMSD of the flap residues 44 to 54 (and 144 to 154) with superposition on the non-flap residues of both monomers. RMSD against the semi-open starting struture (1HHP) and against a known closed structure (crystallised in the presence of an inhibitor, 1G6L[172]), are shown.

Where possible, data is presented on the same scales throughout this chapter for comparison purposes. Alternative inter-atomic distances are shown occasionally to better describe the dynamics of the system.

**Results using the unprotonated aspartic acid dyad**

The results of 10 ns of NVT simulation at 300 K, are shown in Figure 9.2. The symmetry of the starting 1HHP conformation is lost during the equilibration period, and the distances between the aspartic acid residues and the flap tips differ by approximately 5 Å for the first 5 ns of simulation. During this time, residues 44 to 54 (hereafter referred to as the first flap) remain in a semi-open conformation, with the sidechains of residues 50 and 153 in close proximity. Meanwhile, the second flap (residues 144 to 154) begin in a more loosely defined semi-open conformation, and quickly move to a conformation intermediate to the semi-open and curled forms. The closed conformation is not seen for this flap, and the residue 50 to 181 distance is excluded for clarity.

Just after 5 ns, a significant conformational change begins in the second flap, with the distance between the flap tips increasing by approximately 3 Å, and the distance between the second flap tip, and the aspartic acid of the same monomer, decreasing sharply by over 5 Å. This change is caused by the second flap tip reaching into the cavity of the protein, and the third (middle) graph of Figure 9.2 shows several important distances. Initially, the tip of the second flap (residue 150) curls towards an isoleucine residue on the second monomer, but, at 7.3 ns, residue 150 reaches across the cavity, interacting closely with glycine 27. At this point, the second flap occupies much of the protein cavity, appearing in a conformation that further hinders the accessibility of the active site. The distance between flap tips is

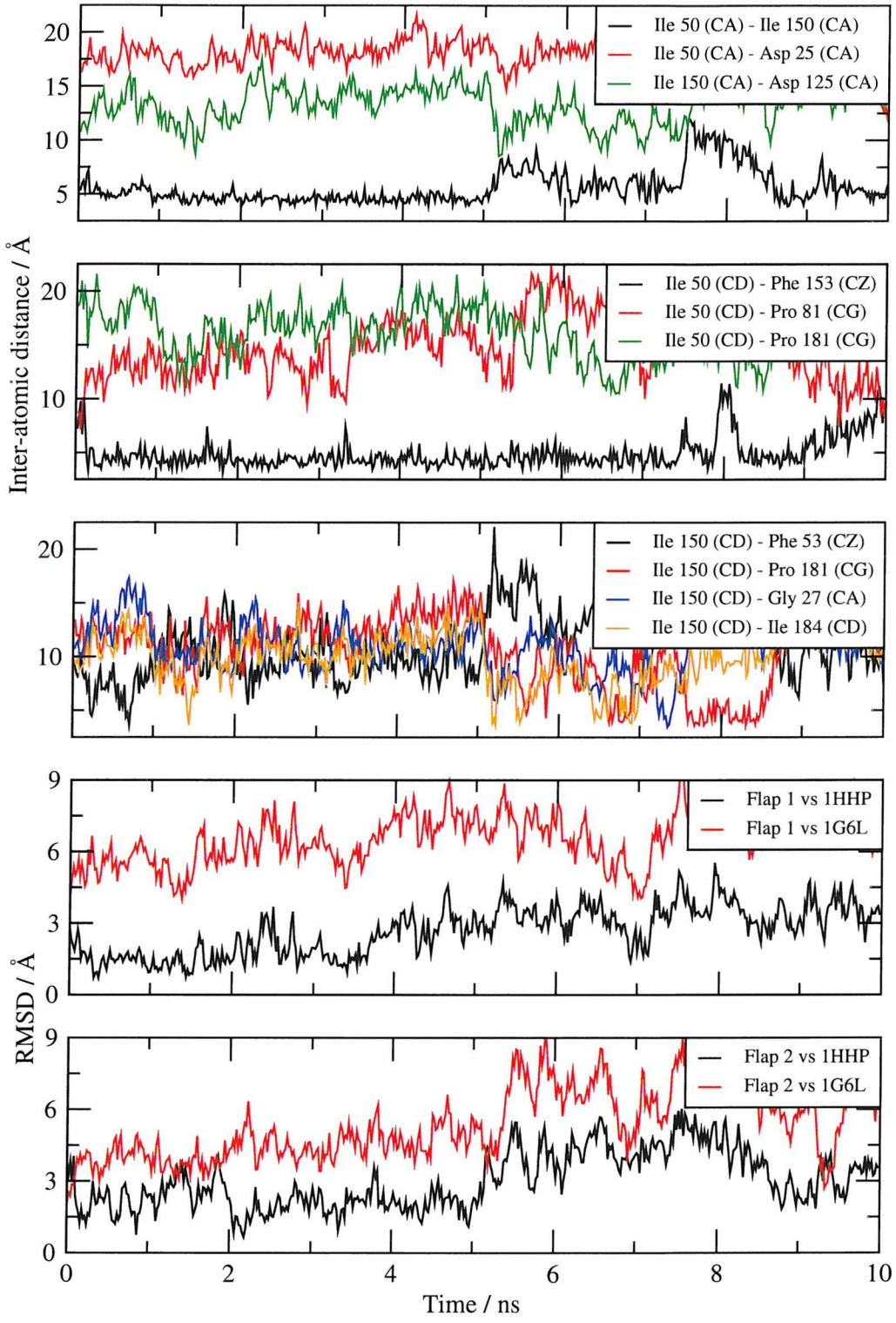



Figure 9.2: Analysis of the 10 ns NVT MD simulation of HIV-1 PR with a unprotonated aspartic acid dyad. Top: α-carbon distance between flap tips. Upper middle: location of flap 1 tip with respect to nearby hydrophobic residues (semi-open, curled and closed conformations are indicated by black, red and green respectively). Middle: location of flap 2 tip with respect to nearby hydrophobic residues. Lower middle: RMSD of flap 1 (residues 46 to 54) against semi-open (1HHP) and closed (1G6L) structures. Bottom: RMSD of flap 2 (residues 146 to 154) against semi-open (1HHP) and closed (1G6L) structures.

at a maximum at 7.3 ns, 6 Å further apart than in the semi-open conformation. The RMSDs against known structures for the second flap indicate a deviation from the semi-open and closed forms of approximately 3 Å.

After 7.3 ns, the second flap moves into the curled conformation, and the distance between the flap tips returns to the starting value of 5 Å. This lasts until 9.4 Å, at which point a conformational change begins in the two flaps, moving to a state intermeditate to the semi-open and curled forms. This motion is continues until the end of the 10 ns simulation, and is not fully described by this time.

The conformations sampled by the system with an unprotonated aspartic acid dyad do not appear to suggest a motion of the flap tips occuring well within the 10 ns timescale, as suggested by NMR data. The crossing of the cavity by the second flap occurs over more residues than just the flap tips, and the first flap remains in a curled conformation for almost the entire simulation. At no point can the active site of the protein be considered ligand accessible, and the flaps remain in contact.

Figure 9.3 shows a secondary structure analysis of the discussed trajectory. There is little disruption of the inital structure, but several β-sheets experience the intermittent formation of coil and bend residues within sheet sections. This is also seen in simulations with the monoprotonated dyad, and is a feature of the flexible regions of the HIV-1 PR system.[177]

**Results using the monoprotonated aspartic acid dyad**

Figure 9.4 shows the results of the 10 ns NVT simulation performed using the monoprotonated aspartic acid dyad system. Throughout the simulation, the distance between the flap tips and the catalytic residues is maintained at approximately 15 Å, indicating that no motion occurs similar to that seen in with the unprotonated dyad system, in which one flap desended into the protein cavity.

The first flap begins in a semi-open conformation, and moves briefly into the curled state at around 2 ns. The flap returns to the semi-open form until 4.6 ns, at

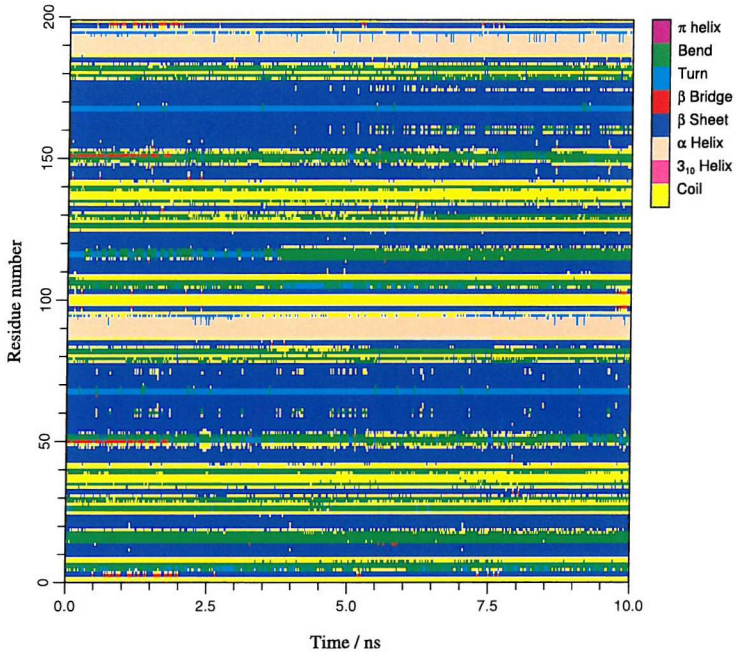



Figure 9.3: Secondary structure analysis for the 10 ns 300 K NVT MD simulation of the unprotonated aspartic acid dyad form of HIV-1 PR.

which point the curled conformation is once again achieved, remaining until the end of the simulation.

The second flap also begins in a state similar to the semi-open form, however after 0.7 ns, a clearly defined closed conformation is obtained. The RMSD against the semi-open structure increases, and the RMSD against the known closed structure falls to a lower level. This change is accompanied the movement of flap tip residue 150 into close proximity with proline 81. This motion requires the flap to reach across the protein, further closing across the cavity as seen in crystal structures when an inhibitor is present.

After 1.7 ns, the second flap returns to a semi-open conformation. After this point, the flap interchanges between the closed and curled forms, in a similar manner to that seen in the first flap.

Conformations adopted by the HIV-1 PR flaps during this simulation are representative of those referred to throughout this chapter as semi-open, curled, and closed. Figures 9.5 and 9.6 show examples of each of these states. Only the

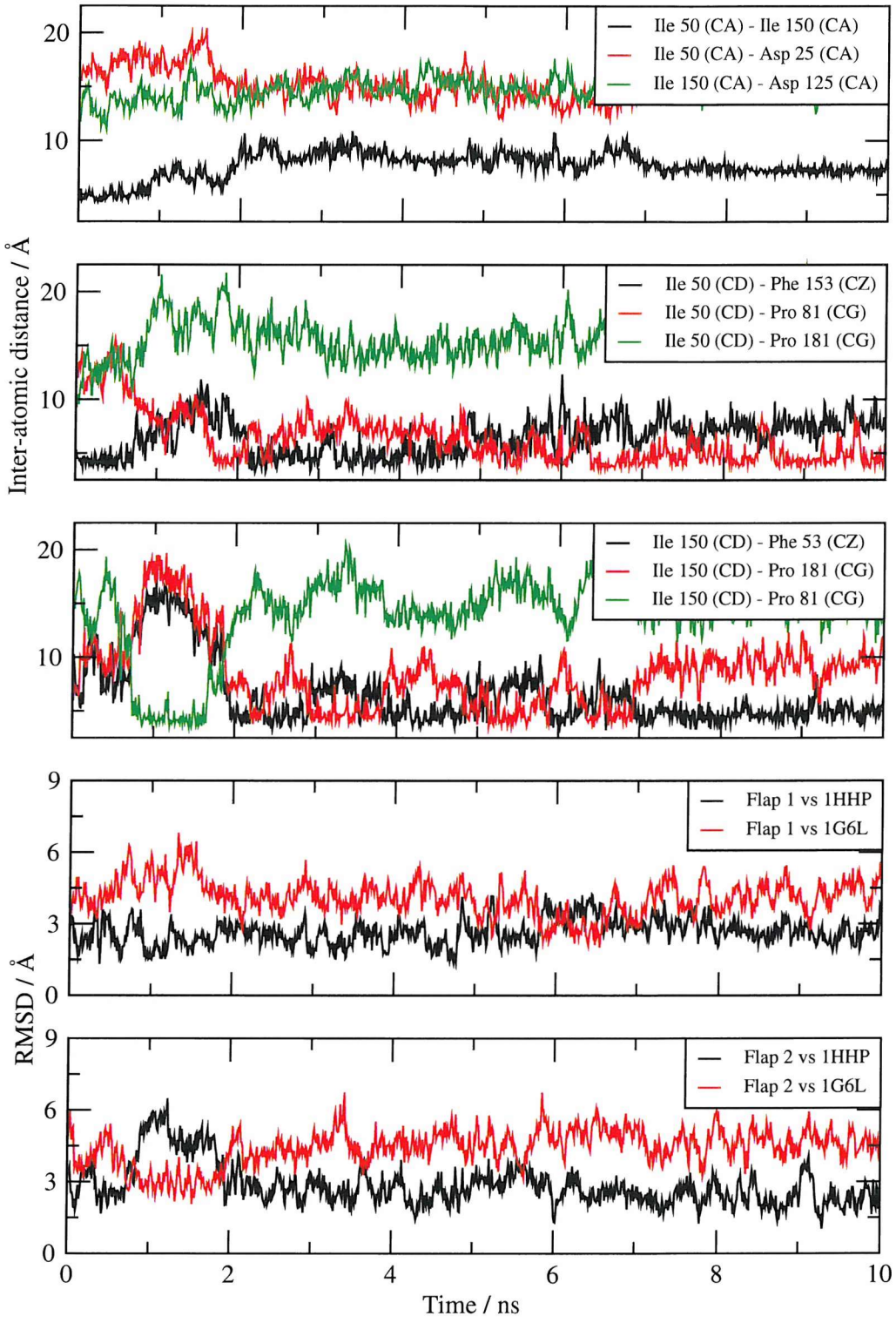



Figure 9.4: Analysis of the 10 ns NVT MD simulation of HIV-1 PR with a monoprotonated aspartic acid dyad. Top: α-carbon distance between flap tips. Upper middle: location of flap 1 tip with respect to nearby hydrophobic residues (semi-open, curled and closed conformations are indicated by black, red and green respectively). Middle: location of flap 2 tip with respect to nearby hydrophobic residues. Lower middle: RMSD of flap 1 (residues 46 to 54) against semi-open (1HHP) and closed (1G6L) structures. Bottom: RMSD of flap 2 (residues 146 to 154) against semi-open (1HHP) and closed (1G6L) structures.

upper part of the protein as shown in Figure 9.1 is displayed.

It is worth noting that the RMSD analysis does necessarily indicate a closed conformation when the structure is more similar to the 1G6L state. The RMSD of the first flap is more similar to 1G6L between 5.8 and 6.6 ns, but residue 50 does not show an interaction with the opposite P1 loop as seen in the closed form. This is considered to indicate the similarities of the three conformations sampled. Owing to this, and the lack of a curled structure for comparison, RMSD analysis will not be shown for later simulations, and the isoleucine 50 and 150 residue environments will be used exclusively to characterise the conformational state.

The secondary structure analysis of the 10 ns simulation show little differences to that seen with the unprotonated dyad system and will be discussed later.

**Protonation state of the aspartic acid dyad**

The analysis of Wang *et al.*[185] performed on a short implicit solvent simulation has been replicated for the 10 ns simluations of the two systems presented in this chapter. This involves monitoring the inter-atomic distance between the two catalytic aspartic acid residues, and the results of this are shown in Figure 9.7. The distance observed in the crystal structure of 1HHP is shown with a dotted line.

The monoprotonated dyad sees two distinct states. The first has an α-carbon separation between residues 25 and 125 of over 7 Å, and the second has a distance less than 7 Å. Visualisation of the simulation trajectories (not shown) indicates that this change, which occurs rapidly at 5 ns, involves the removal of a water molecule from between the aspartic acid sidechains. This water molecule entered the region during the equilibration stage. This is not unexpected, as a water molecule is involved in the reaction catalysed by HIV-1 PR. The second state has a similar inter-residue distance to that seen in the crystal structure.

The unprotonated dyad begins with an aspartic acid separation similar to that seen in the crystal structure (as observed by Wang *et al.*), but this distance quickly increases due to the insertion of several water molecules between the aspartic acid

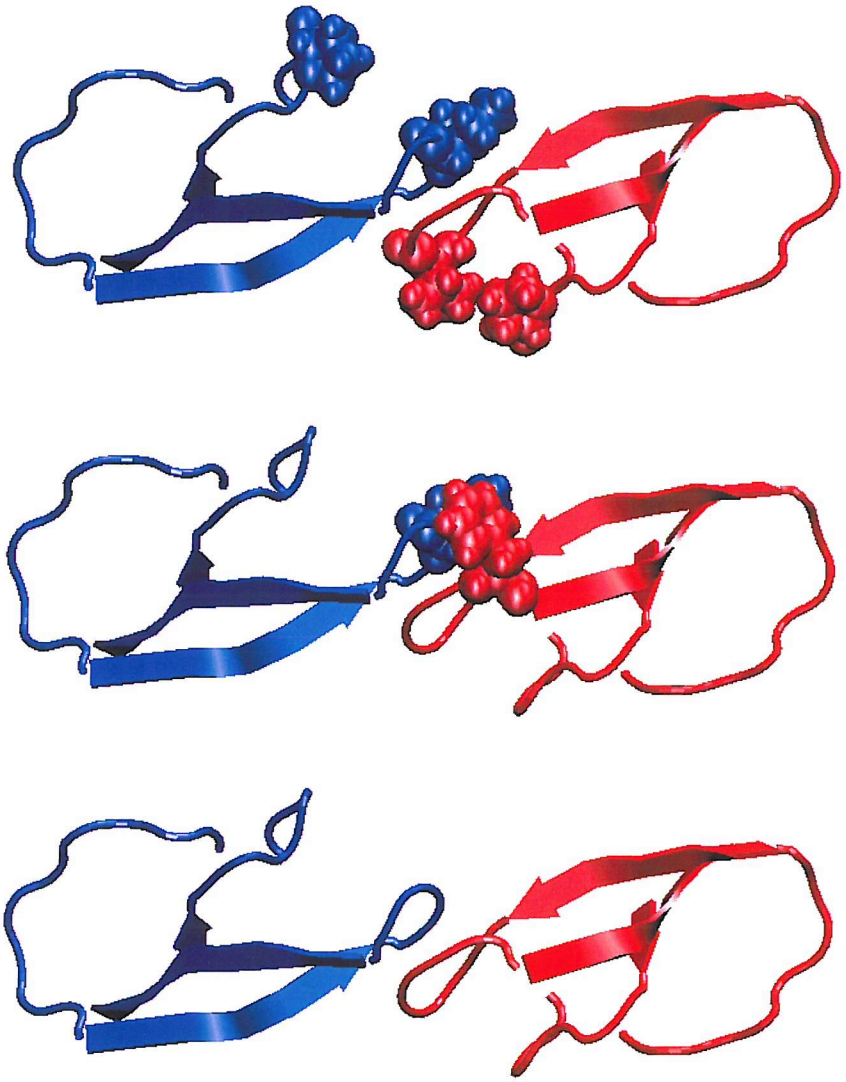


Figure 9.5: Semi-open (blue) and curled (red) conformations, sampled after 10 ns of MD simulation at 300 K with a monoprotonated aspartic acid dyad. Residues 34 to 59 and 76 to 83 are shown in blue, and residues 134 to 159 and 176 to 183 are shown in red. Characteristic distances are indicated with residues shown by their van der Waals radii. The view is equivalent to looking down on Figure 9.1. Top: residues in red show close proximity between the flap tip (residue 150) and the P1 loop of the same monomer (residue 181), characteristic of the curled conformation. Middle: the flap shown in blue is in the semi-open state, shown by interactions between the flap tip (residue 50) and a phenylalanine on the opposite flap (residue 153). Bottom: Cartoon representation only is shown for clarity.

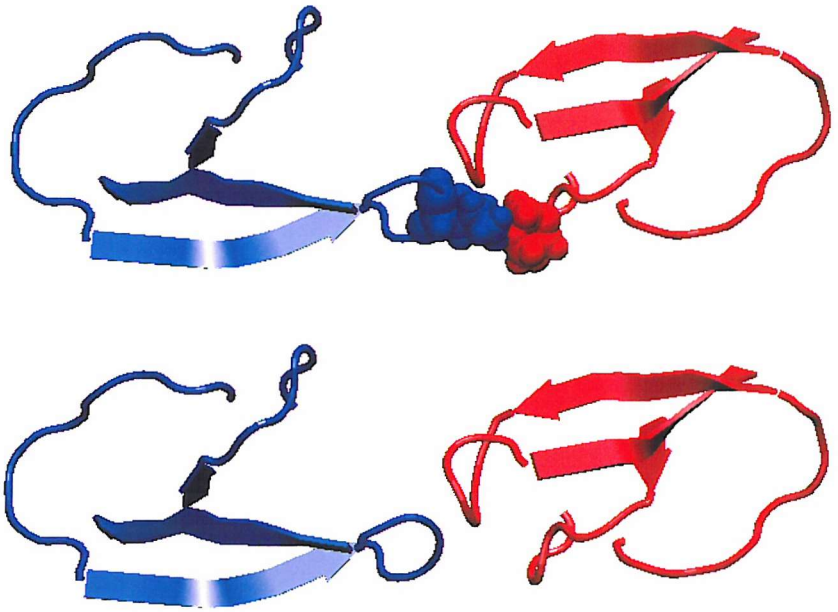


Figure 9.6: Closed conformation shown by the flap in blue, sampled after 1.3 ns of MD simulation at 300 K with a monoprotonated aspartic acid dyad. Residues 34 to 59 and 76 to 83 are shown in blue and residues 134 to 159 and 176 to 183 are shown in red. Characteristic distances are indicated with residues shown by their van der Waals radii. The view is equivalent to looking down on Figure 9.1. Top: the flap shown in blue reaches across the protein, forming an interaction between the flap tip (residue 50) and the P1 loop (residue 181) of the opposite monomer. Bottom: Cartoon representation only is shown for clarity.

side chains. The resulting structure indicates a distance approximately 2 Å larger than that seen in the crystal structure. The one of the widest separations, at 7.3 ns, accompanies the interaction of the second flap tip to residue 27, as previously discussed.

It would therefore appear that the dimer interface is more stable when using a monoprotonated dyad. The results of Wang *et al.* are drawn from a simulation that is clearly too short (120 ps), and the implicit solvent used is unlikely to have represented the important solvation effects seen at the active site.

### 9.2.3 Summary

The simulation with the unprotonated dyad (previously used by Perryman *et al.* and Scott *et al.*) shows a motion in which one flap reaches into the protein cavity, interacting with an area close to the active site. This motion is accompanied by an

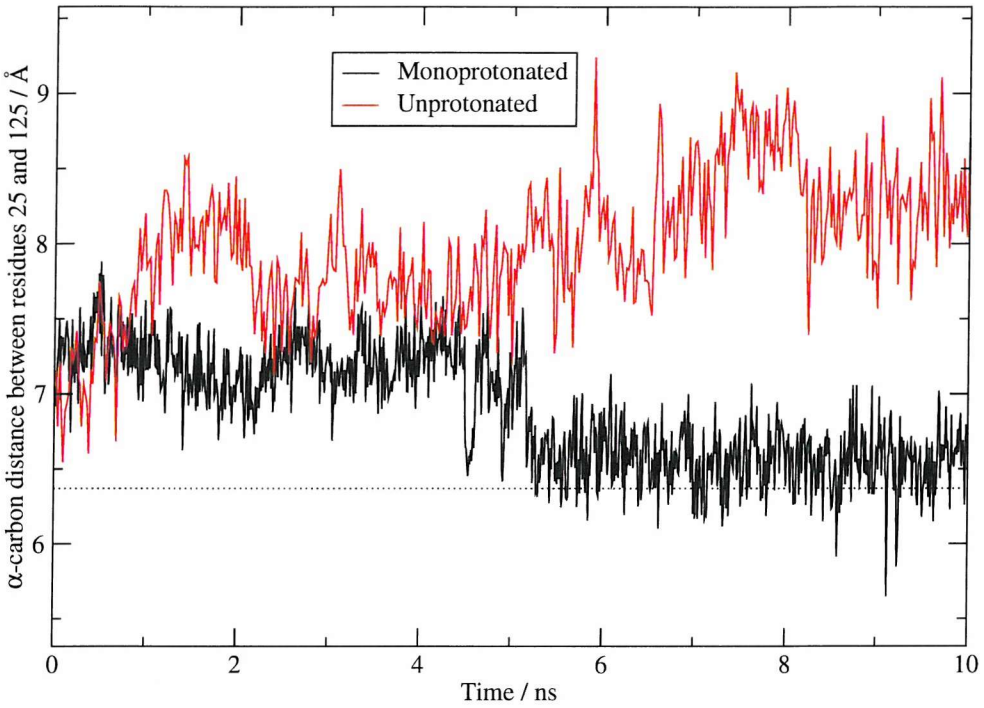



Figure 9.7: Inter-residue distance between catalytic aspartic acids 25 and 125. Results from 10 ns of MD simulation using the unprotonated and monoprotonated conformations are shown. The dotted line indicates the inter-residue distance in the 1HHP crystal structure.

unexpected separation of the catalytic aspartic acid residues.

Results from the simulation of the monoprotonated dyad system indicate a stable cavity region, with the flap tips remaining a constant distance from the active site. A conformational change to the closed state occurs, and the semi-open and curled conformations inter-convert at a timescale well within that of the simulation.

These two motions seen in the monoprotonated dyad system are likely to correspond to the fast event described in the NMR study by Ishima *et al.* The flaps do not separate in either simulation as seen by Scott *et al.*, but it is unlikely that the separation event was due to the use of an unprotonated dyad, since Perryman *et al.* did not sample such an event. The opening was therefore either a random occurrence, or due to the use of the GROMOS force field. The simulations presented here agree with those of Perryman *et al.*, which also sampled a closed protein cavity and curling of the flap tips.

The opening event described by Ishima *et al.* as occuring on a 100 $\mu$s timescale

has not been sampled in the molecular dynamics simulation presented so far. Investigation is therefore continued at elevated temperatures using the monoprotonated dyad system only.

## 9.3 Thermal simulations

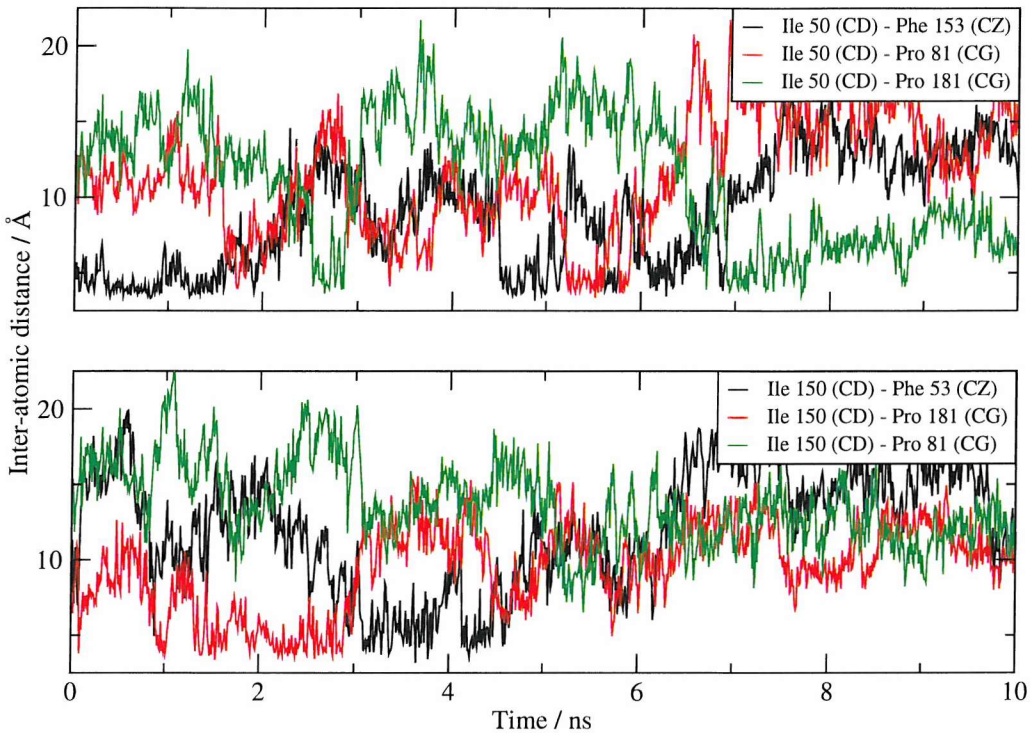
Simulations have been performed at 350, 400, 450 and 500 K for the monoprotonated dyad system. Simulation velocities have not been randomly reassigned at the target temperature as previously described for the T4 lysozyme and EcDHFR systems, due to concerns that the random velocities may produce undesired results before equilibrating. Instead, the velocities equilibrated at 300 K are used, and the Langevin thermostat adjusts these to the required temperature.

Owing to the complexities of describing the conformation of HIV-1 PR, results at each temperature are individually displayed. Figures 9.8 (a) and (b) show the results at 350 and 400 K respectively.

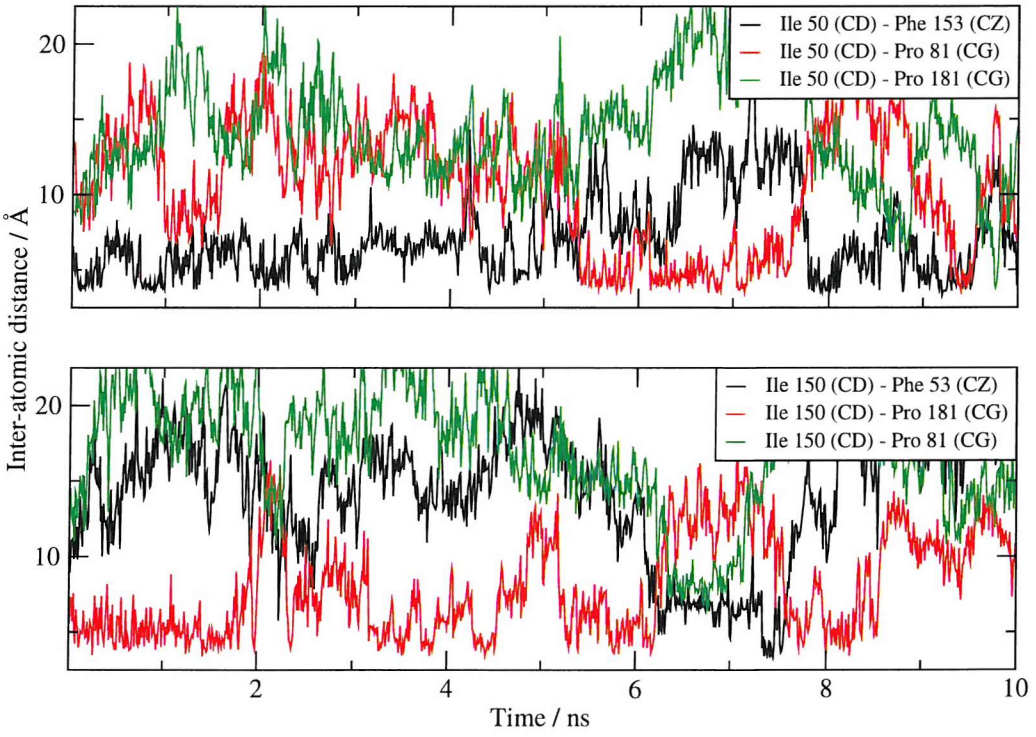
The first flap in the 350 K simulation samples the closed conformation on two occasions (from 2.4 to 2.8 ns and from 6.9 ns to the end of the simulation), and otherwise inter-converts between the semi-open and curled states. The second flap samples the curled and semi-open forms, but when the first flap is closed, the second exists in an undefined conformation. Inspection of the closed flap in the 300 K simulation shows the same results. In all simulations the flap opposite to a closed flap exists in this intermediate state.

The 400 K simulation shows only a briefly formed closed conformation in the first flap at 9.8 ns, and otherwise indicates interconversion between the semi-open and curled conformations. Both the 350 and 400 K simulations show similar results to those seen at 300 K, with the two protein flaps maintaining contact at all times.

The 450 K simulation is shown in Figure 9.9. The first flap inter-converts between the semi-open and curled forms, but the conformation of the second flap is less clearly defined throughout the simulation. After 6 ns, the first flap reaches into,




(a) 350 K




(b) 400 K

Figure 9.8: Analysis of the 350 and 400 K 10 ns NVT MD simulation of HIV-1 PR with a monoprotonated aspartic acid dyad. Top: location of flap 1 tip with respect to nearby hydrophobic residues. Bottom: location of flap 2 tip with respect to nearby hydrophobic residues.

and across, the protein cavity, corresponding to a close distance between the flap tip and catalytic aspartic acid of the opposite monomer. Contact with the second flap is lost, which then has nothing to stabilise the semi-open conformation, and the second flap opens away from the active site, showing a 15 Å separation of the flap tips. At 7.4 ns the first flap adopts a tightly curled conformation, maintaining a significant flap separation. As the end of the simulation approaches, the second flap also moves into a curled conformation. The flaps do not meet at this point and the active site is accessible. The trajectory sampled in this simulation appears to be similar to that reported by Scott *et al.*, although Scott *et al.* sample an immediate opening. Even at 450 K, results presented here indicate some stability of conformations with touching flap tips.

The 500 K simulation shows a remarkably different event to any previously sampled, as shown in Figure 9.10. The two flaps both drop into the protein cavity, indicated by a decrease in the distance between the catalytic aspartic acid residues and the flap tips. Water is completely excluded from the cavity in this collapsed conformation, which exists from 1.2 ns to the end of the simulation. There is no experimental data to suggest the existance of this conformation, although a similar event was sampled during a simualtion performed by David *et al.* with an implicit solvent distance-dependent dielectric model. It is therefore likely that the accessibility of this conformation indicates a deficit of the simulation model for this system using such high temperatures.

The stability of the collapsed conformation seen at 500 K has been investigated by reducing the temperature at the end of 10 ns to 300 K. 2 ns of simulation has then been performed, during which the collapased conformation is maintained (not shown). This result indicates potential difficulties of a parallel tempering simulation of this system. Temperatures of 450 K appear to be required to sample an opening event on the simulation timescale, but at 500 K, a collapsed conformation is accessible, which is unlikely to reopen.

Figure 9.11 shows the secondary stucture analysis of all simulations for the monoprotonated dyad system presented so far. As the temperature increases, the

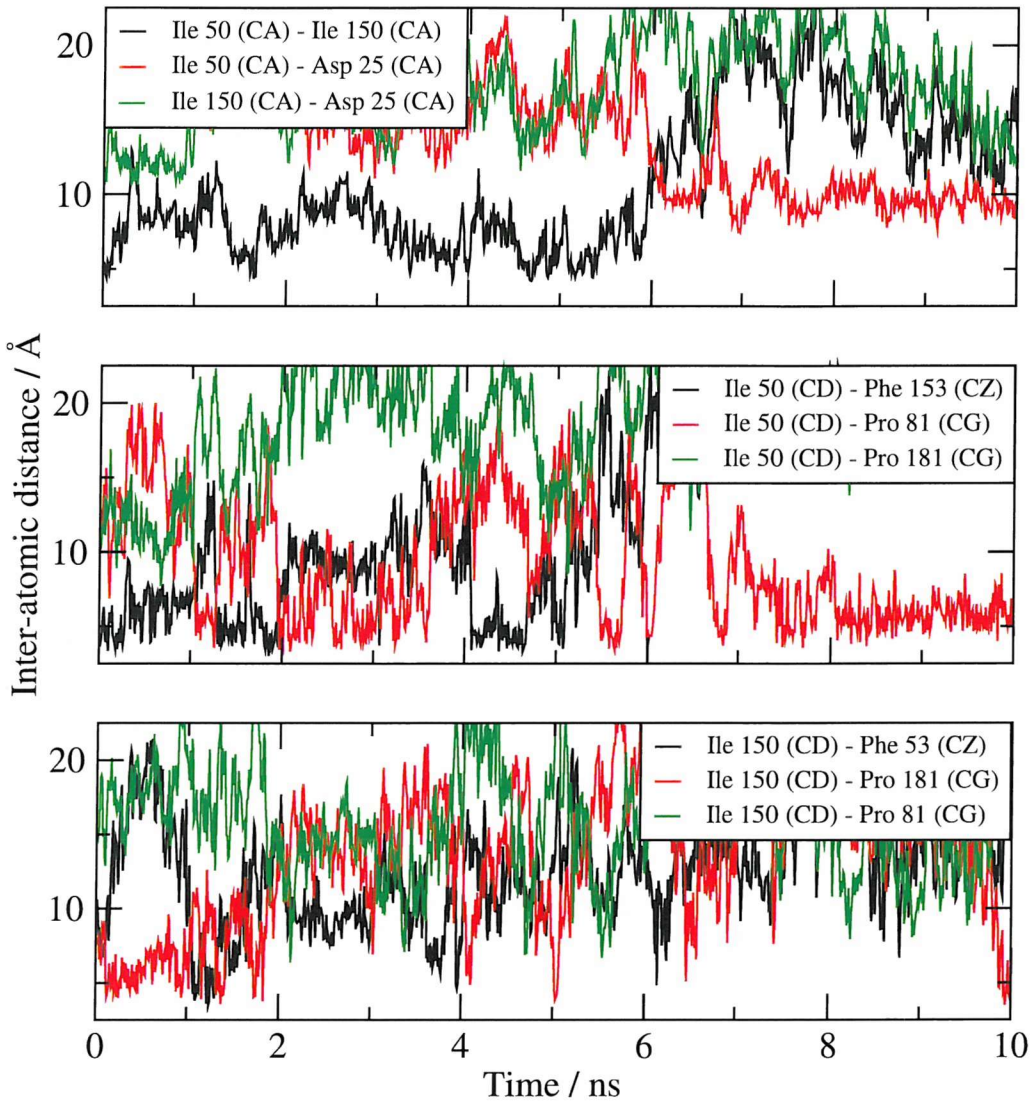



Figure 9.9: Analysis of the 10 ns 450 K NVT MD simulation of HIV-1 PR with a monoprotonated aspartic acid dyad. Top: location of flap 1 tip with respect to nearby hydrophobic residues. Bottom: location of flap 2 tip with respect to nearby hydrophobic residues.

β-sheet structure of the protein is dramatically reduced, particularly in the vicinity of the protein flaps. HIV-1 PR is clearly not stable at high temperatures, with significantly increased β-sheet loss at only 350 K. It is likely that the opening event at 450 K, and the collapse of the protein flaps at 500 K, are due to the breakdown of the protein's structure.

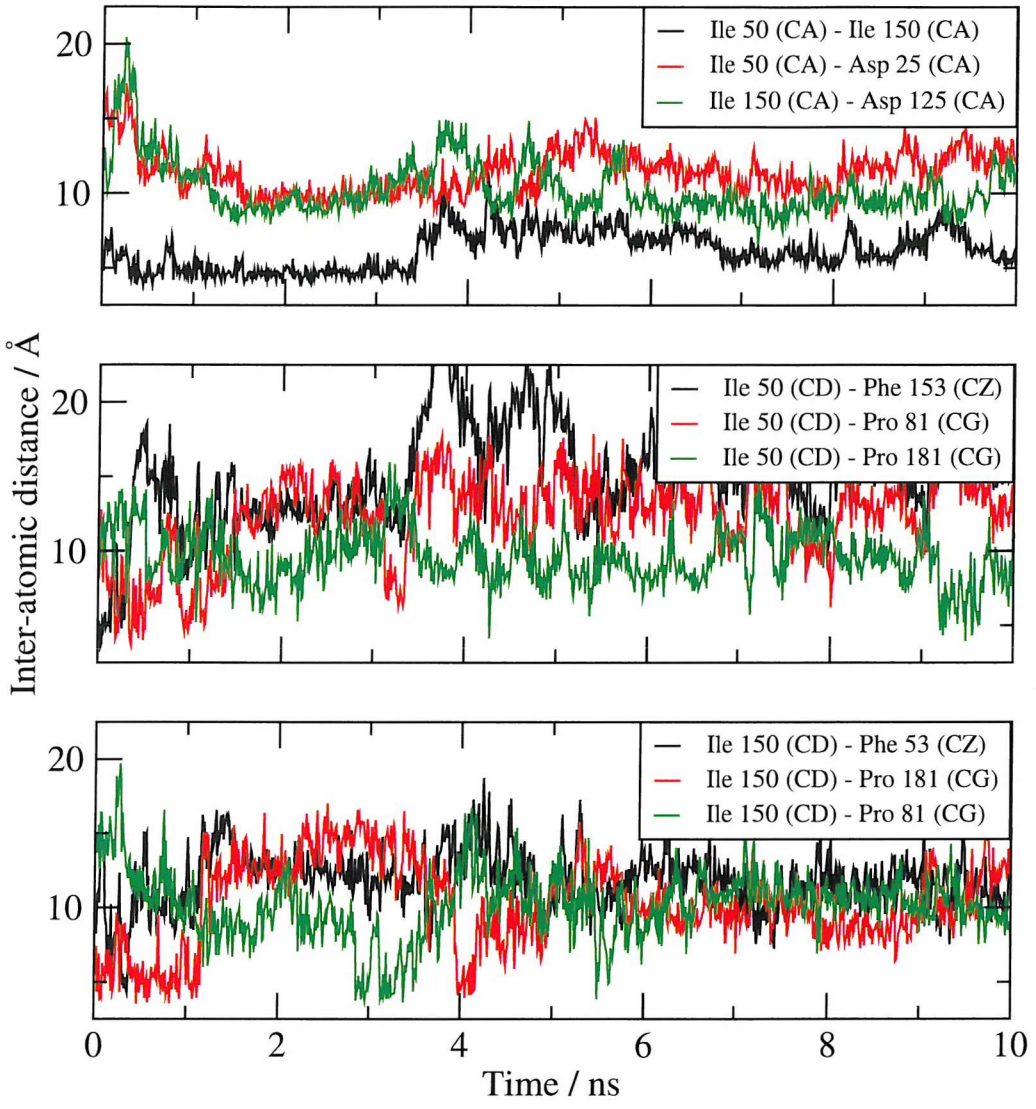
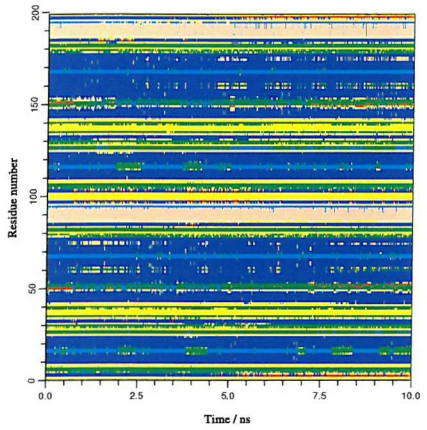


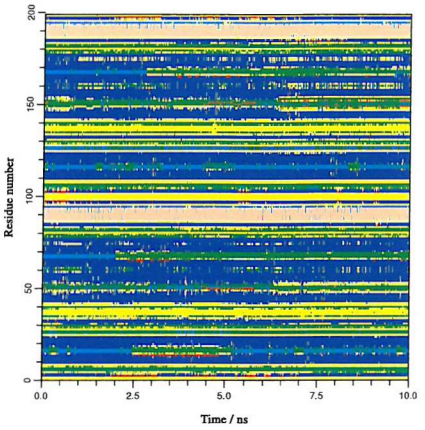

Figure 9.10: Analysis of the 10 ns 500 K NVT MD simulation of HIV-1 PR with a monoprotonated aspartic acid dyad. Top: α-carbon distance between flap tips. Middle: location of flap 1 tip with respect to nearby hydrophobic residues. Bottom: location of flap 2 tip with respect to nearby hydrophobic residues.

### 9.3.1 Protonation state of the aspartic acid dyad

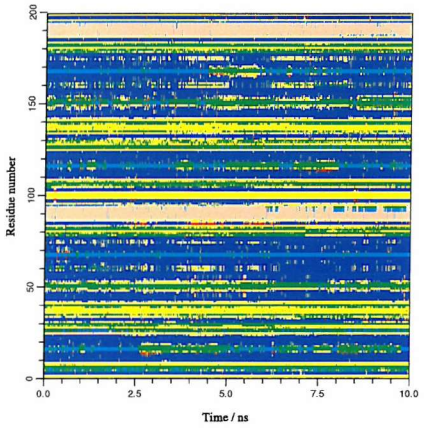
The stability of the aspartic acid dyad has been monitored for the thermal simulations of the monoprotonated dyad system, and for a 10 ns simulation with the unprotonated dyad at 500 K, which has not been discussed. Table 9.1 shows the mean and standard deviations of the α-carbon distances between residues 25 and 125. The conclusions previously drawn are confirmed; the monoprotonated dyad exhibits shorter distances between the catalytic residues that are more similar to the crystal structure.



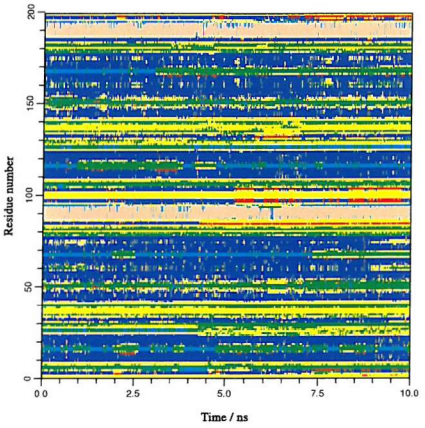

(a) 300 K



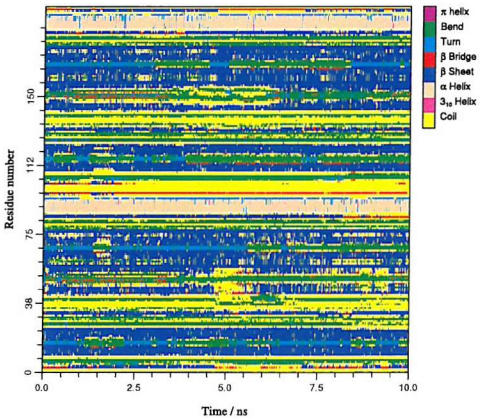

(b) 350 K




(c) 400 K




(d) 450 K




(e) 500 K

Figure 9.11: Secondary structure analysis for the 10 ns NVT MD simulations of the monoprotonated aspartic acid dyad form of HIV-1 PR at a range of temperatures.

| Protonation state of dyad | Temperature / K | Average α-carbon distance / Å | Standard deviation |
|---|---|---|---|
| -1 | 300 | 6.89 | 0.38 |
| -1 | 350 | 6.50 | 0.25 |
| -1 | 400 | 6.42 | 0.38 |
| -1 | 450 | 7.55 | 1.13 |
| -1 | 500 | 6.28 | 0.29 |
| -2 | 300 | 7.99 | 0.47 |
| -2 | 500 | 8.79 | 1.07 |

Table 9.1: Inter-residue distance for aspartic acids 25 and 125.

### 9.3.2 Summary

The HIV-1 PR system has been shown to be unstable at high temperatures, possibly due to the high β-sheet content of the protein, with a significant loss of secondary structure at only 350 K. At 400 K and below, the protein flaps inter-convert between clearly defined semi-open, closed and curled conformations. No opening between the flaps is observed at these temperatures. At 450 K, an opening event was sampled in which one flap reaches into the protein cavity, leaving the other free to move away from the active site. A flap moving into the protein cavity was also seen using the unprotonated dyad, and an increase in the distance between the catalytic aspartic acid residues occurs in both instances (as shown in Table 9.1). Towards the end of the 450 K simulation, a conformation is adopted in which the flaps are both in a curled state, but they do not touch and the protein cavity does not close. At 500 K, both flaps are seen to move into the protein cavity, adopting a collapsed conformation. Significant disruption of the protein's secondary structure could be responsible for sampling of the open and collapsed states.

## 9.4 Parallel tempering

A parallel tempering simulation with temperatures up to 400 K could be of interest in investigating the distribution of the semi-open, closed and curled flap conformations. However, it is unlikely that an opening event will occur within a feasible timescale, and resources have therefore been concentrated on RDFMD simulations.

## 9.5 Reversible digitally filtered molecular dynamics

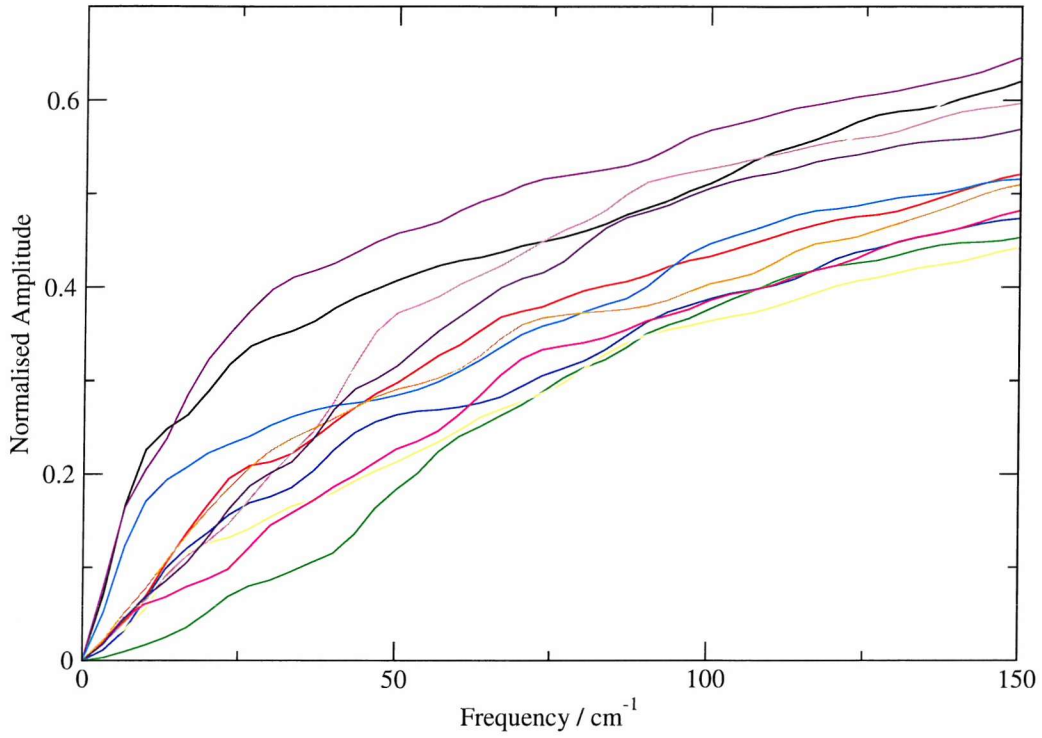
RDFMD has the ability to selectively enhance motions in certain regions of the protein, without increasing the energy in all degrees of freedom. This is ideal for the HIV-1 PR system, which is highly flexible and unstable at elevated temperatures. The flap tip residues are known to be mobile and have been shown to adopt distinct conformers. RDFMD is therefore applied to the flap tip residues 49 to 51 and 149 to 151.

Cumulative amplitude plots of the spectral components of $\psi$, $\phi$ and $\omega$ angles in the targeted residues, and their neighbours, are shown in Figure 9.12. Results have been taken from a 10 ps NVE simulation from the starting equilibrated system, with sampling every 2 fs. Similar to the results for T4 lysozyme and EcDHFR, it appears that a frequency target cannot be used to avoid amplifying $\omega$ angle motions, which also contain significant motions with frequencies of 0–100 $cm^{-1}$. RDFMD simulations are therefore performed at a range of internal temperature caps, excluding any trajectories that show isomerisation of the peptide bond.

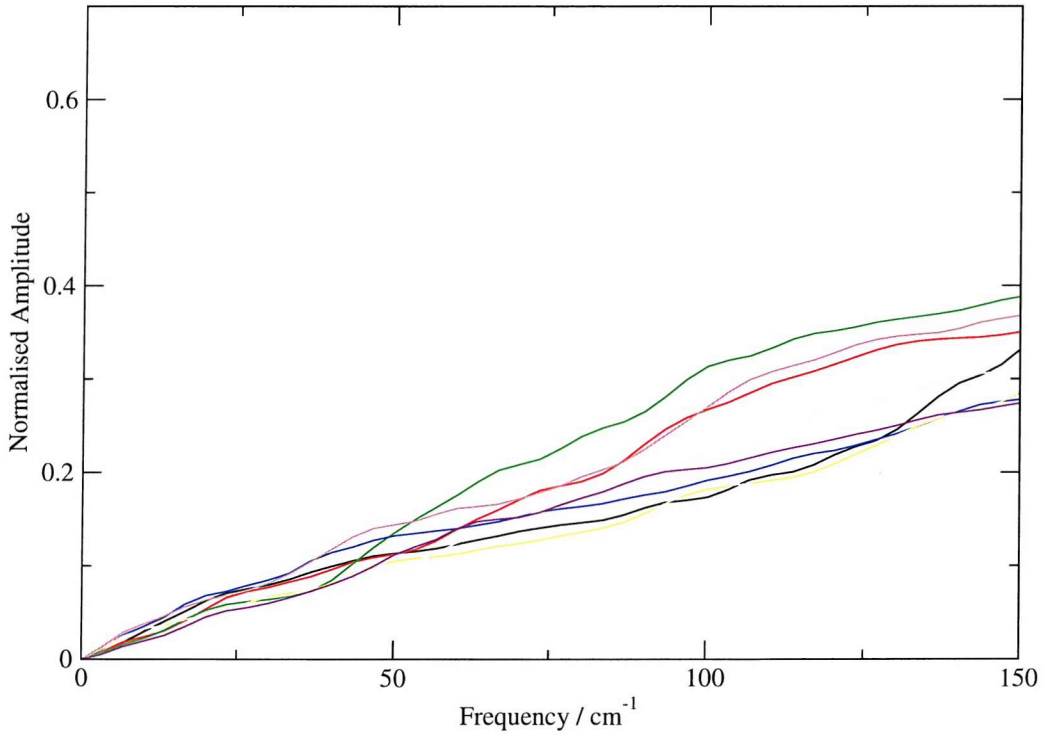
### 9.5.1 Optimised protocol results

RDFMD simulations have been performed using the protocol optimised in Chapter 5. This includes a filter delay of 50 or 100 steps, the use of a 201 coefficient, 0–100 $cm^{-1}$ digital filter, an amplification factor of 2, and 4 ps of NPT simulation between filter sequences to allow the system temperature to equilibrate. 100 filter pulses have been performed, and conformations sampled from the resulting 400 ps of NPT simulation have been analysed. Internal temperature caps of 900, 1100, 1500 and 2000 K have been tested, but all simulations with temperature caps of 1500 K or above see isomerisation of peptide bonds, and are therefore excluded. Once again, owing to the complexity of describing the HIV-1 PR conformation, each simulation is shown separately.

Figure 9.13 shows the results of the RDFMD simulation using a filter delay of 50 steps and an internal temperature cap of 900 K. The first flap inter-converts




(a) ψ and ϕ angles




(b) ω angles

Figure 9.12: Cumulative amplitude plots of backbone torsions for residues targeted by RDFMD, and their neighbours. Each colour represents a different backbone torsion.

between the semi-open and curled conformations, and the second flap exhibits a conformation intermediate to the semi-open, curled and closed states, until 300 ps, at which point a semi-open form is briefly sampled, occuring again at 350 ps. At the very end of the simulation, from 350 ps onwards, both flaps twist away from each other, into curled conformations. The extent of this motion is further than curling motions previously described, indicated by the increasing distance between the α-carbons of the isoleucine flap tip residues, which opens by approximately 10 Å. The flaps separate entirely, leaving an open path to the catalytic residues of approximately 9 Å width. The final conformation is similar to the final state of the 450 K thermal simulation, but the movement into this open state does not go via a flap moving into the protein cavity, or via an opened flap with an increased flap tip to active site distance. The curling event could therefore be sufficient to allow substrate binding, without an opening similar to that described by Scott *et al.* Importantly, the motion that leads to an increased accessibility of the active site is similar to the interconversion between the semi-open and curled forms, which is rapidly sampled by room temperature simulations. The final conformation is shown in Figure 9.14.

The RDFMD simulation with a filter delay of 100 steps and an internal temperature cap of 900 K is shown in Figure 9.15. A similar opening event is sampled to that seen in the previous RDFMD simulation. The opening occurs from 176 ps onwards, and the flaps make contact intermittently, at around 230 and 366 ps. Again, the opening is achieved by both flaps adopting tightly curled conformations. The behaviour of the flap tips differs from that seen in the last RDFMD simulation, with a reduced distance between the flap tips and the catalytic residues. This is not similar to the collapsed state previously observed at 500 K, and an accessible cleft above the active site is maintained at all times.

The third RDFMD simulation (not shown), with a filter delay of 50 steps and an internal temperature cap of 1100 K, samples a lesser opening to the RDFMD simulations previously discussed. The two flaps lose contact with each other, but do not separate beyond a few angstroms, and repeatedly touch and reopen throughout the simulation. Once again the semi-open and curled conformations

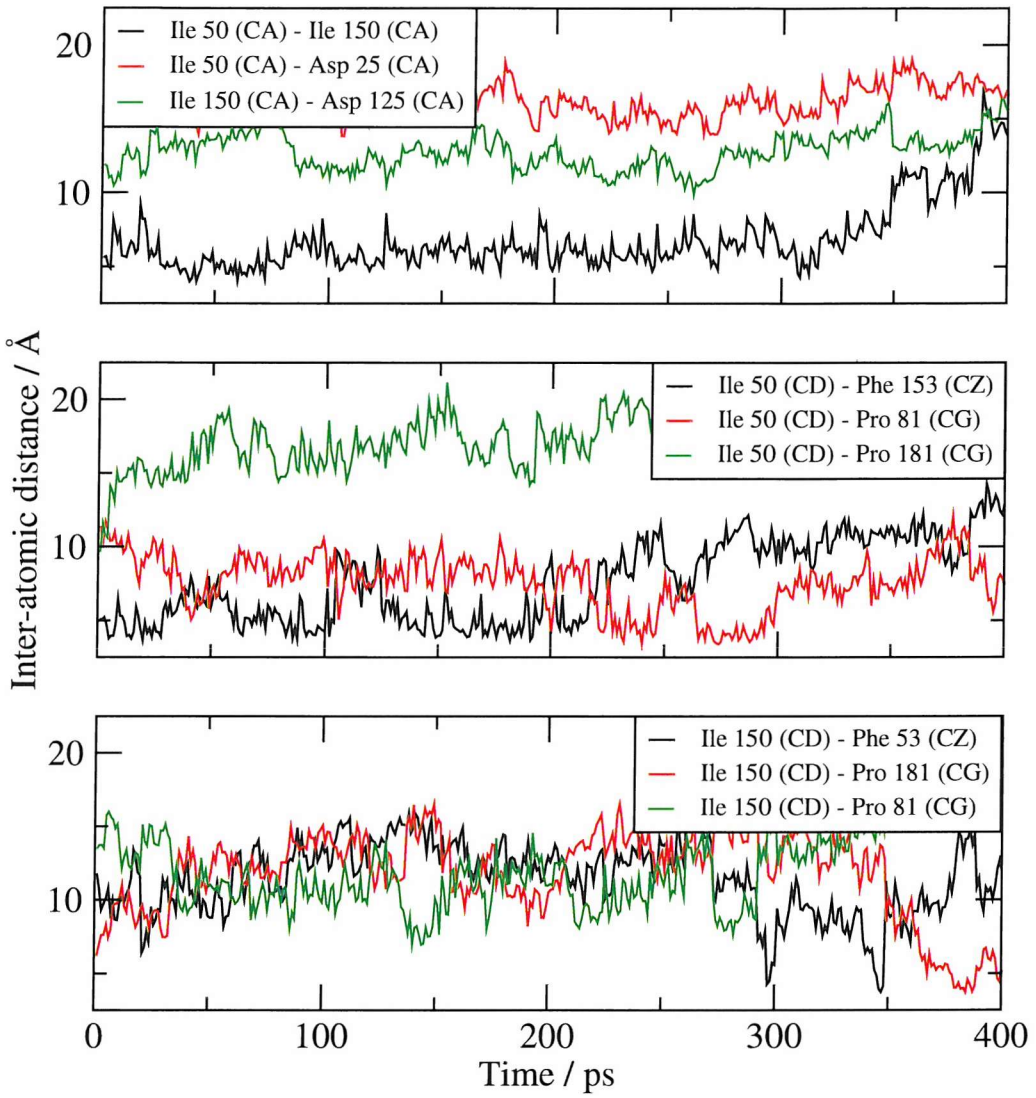
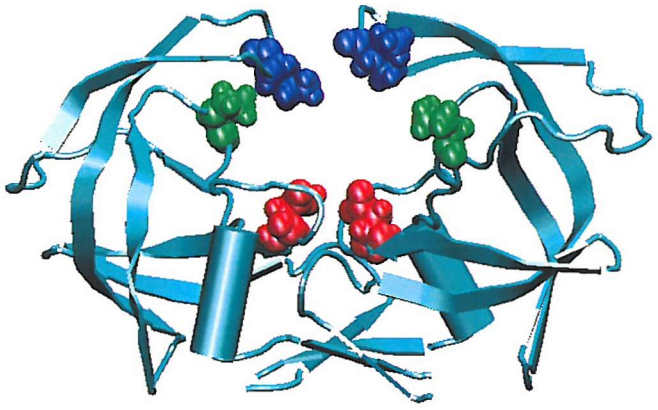


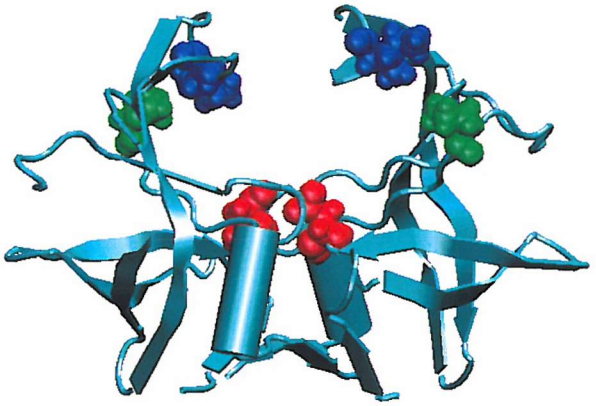

Figure 9.13: Analysis of the 400 ps RDFMD simulation of HIV-1 PR using a filter delay of 50 steps and an internal temperature cap of 900 K. Top: α-carbon distances that can describe opening events. Middle: location of flap 1 tip with respect to nearby hydrophobic residues. Bottom: location of flap 2 tip with respect to nearby hydrophobic residues.

are seen, with both flaps in curled conformations when the cavity is open.

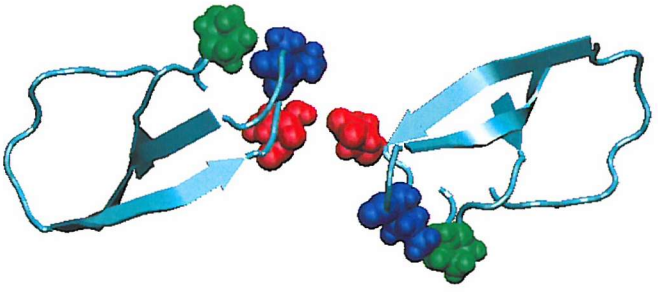
Figure 9.16 shows the results of an RDFMD simulation with a filter delay of 100 steps, and an internal temperature cap of 1100 K. During this simulation, a significant opening event occurs that reaches a conformation similar to that reported by Scott *et al.* After 168 ps, the second flap adopts a curled conformation, and the first flap is surrounded only by solvent, and slowly increases in distance from the active site until 280 ps. This prerequisite to opening is similar to that seen with the 450 K thermal simulation, although the flap left in the semi-open state



(a) Front view of protein, as used in Figure 9.1.



(b) Angled view showing separation of flap residues and accessibility of the active site.



(c) View from above with residues 1 to 24, 36 to 33, 60 to 75 and 84 to 99 (and the same for the second monomer) excluded for clarity.

Figure 9.14: The final conformation obtained from 400 ps of RDFMD simulation of HIV-1 PR. A filter delay of 50 steps and an internal temperature cap of 900 K have been used. The flap tips (residues 50 and 150) are shown in blue, the proline of the P1 loop (residues 81 and 181) are shown in green, and the clearly accessible active site is shown in red.

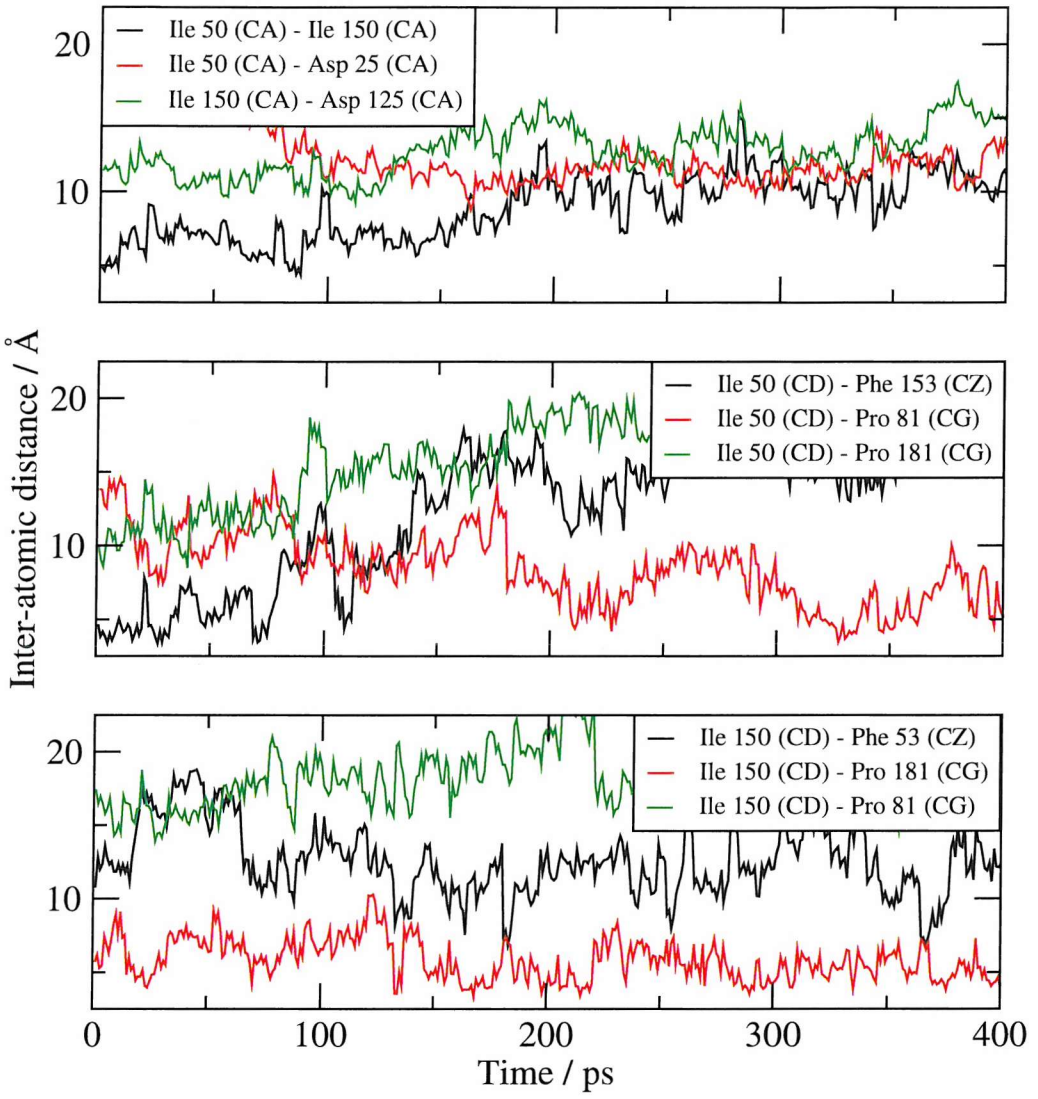



Figure 9.15: Analysis of the 400 ps RDFMD simulation of HIV-1 PR using a filter delay of 100 steps and an internal temperature cap of 900 K. Top: α-carbon distances that can describe opening events. Middle: location of flap 1 tip with respect to nearby hydrophobic residues. Bottom: location of flap 2 tip with respect to nearby hydrophobic residues.

without any hydrophobic connections to other regions of the protein in the thermal simulation did so due to the opposite flap having dropped towards the active site. In this case the cavity is maintained, and the flap which does not open remains in a curled conformation. After 280 ps, the opening is reversed, and the structure returns to one in which the active site is inaccessible. The first flap approaches a closed conformation at the end of the simulation. This simulation appears to more completely sample the motion reported by Scott *et al.*, with both an opening motion away from the active site, and a closing that Scott *et al.* did not observe.

The most open conformation, sampled at 280 ps, is shown in Figure 9.17.

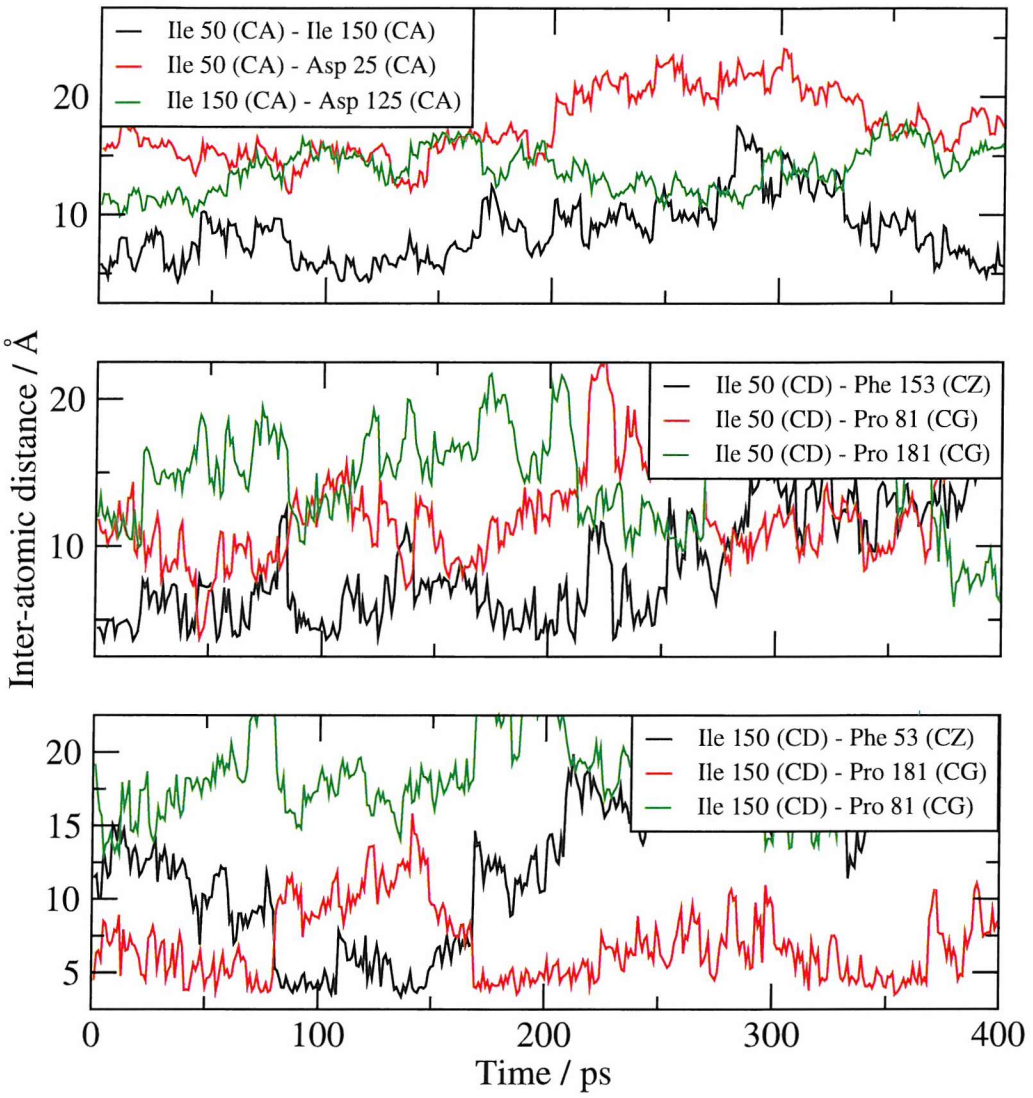
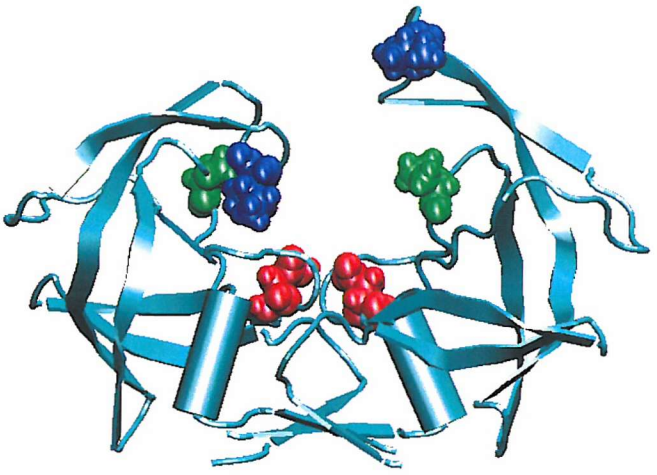
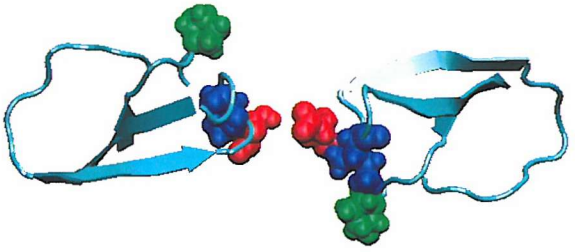



Figure 9.16: Analysis of the 400 ps RDFMD simulation of HIV-1 PR using a filter delay of 100 steps and an internal temperature cap of 1100 K. Top: α-carbon distances that can describe opening events. Middle: location of flap 1 tip with respect to nearby hydrophobic residues. Bottom: location of flap 2 tip with respect to nearby hydrophobic residues.

It is worth noting the sudden changes in the lower graph in Figure 9.16. At 80 ps, the second flap moves from the curled to the semi-open state, with no intermediate conformations sampled during the NPT simulation. The reverse occurs at 168 ps, returning to the curled state. The lack of intermediate states indicates that the conformational change is sampled entirely during one set of digital filter applications. This occurs in all the RDFMD simulations (although is clearest in this example) and shows that RDFMD is efficient at moving between



(a) Front view of protein, as used in Figure 9.1.

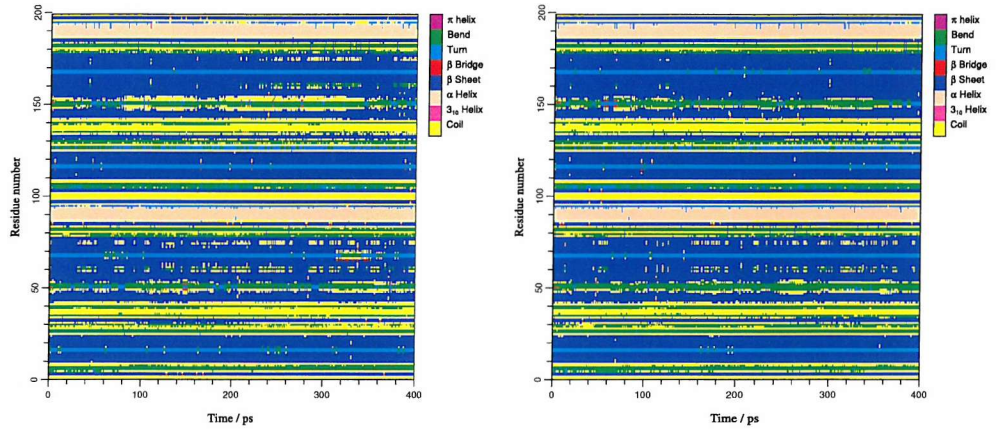


(b) View from above with residues 1 to 24, 36 to 33, 60 to 75 and 84 to 99 (and the same for the second monomer) excluded for clarity.

Figure 9.17: The conformation sampled after 280 ps of RDFMD simulation of HIV-1 PR. A filter delay of 100 steps and an internal temperature cap of 1100 K have been used. The flap tips (residues 50 and 150) are shown in blue, the proline of the P1 loop (residues 81 and 181) are shown in green, and the active site is shown in red.

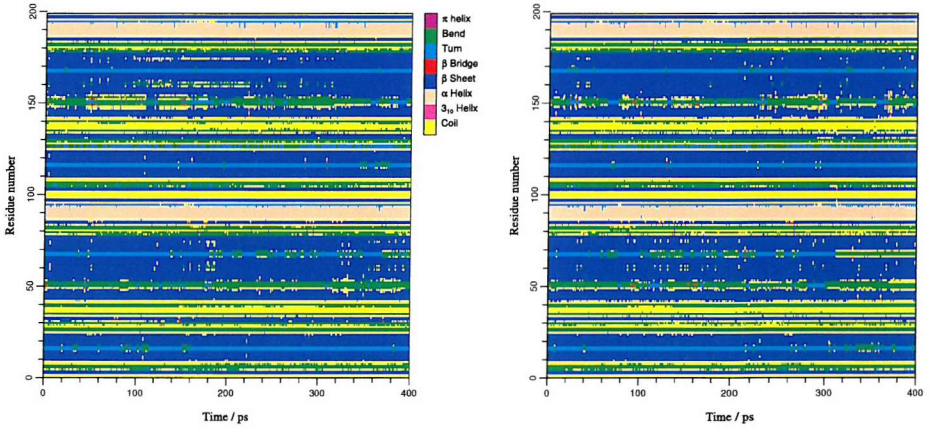
conformations.

Figure 9.18 shows the secondary stucture analysis of the four RDFMD simulations discussed in this chapter. The abundance of β-sheet structure is similar to that of the 300 K simulation shown in Figure 9.11, and the disruption due to the application of RDFMD is far less than that seen when increasing the energy in all degrees of freedom.




(a) D = 50, T = 900 K

(b) D = 100, T = 900 K




(c) D = 50, T = 1100 K

(d) D = 100, T = 1100 K

Figure 9.18: Secondary structure analysis for the four RDFMD simulations of HIV-1 PR. The filter delay, D, and internal temperature cap, T, are shown for each plot.

## 9.6 Summary

Four RDFMD simulations have been described, and in each case, opening events occur that increase the accessibility of the active site. One simulation shows an opening similar to that reported by Scott *et al.*, with one flap lifting away from the active site. The corresponding closure is not sampled by Scott *et al.*, but is seen within the RDFMD simulation.

The other three RDFMD simulations observe opening events via the curled conformations of both flap tips. Opening in this manner appears to occur as an extension of the curling motion seen in MD simulations presented earlier in the chapter. Accessibility of the active site by this mechanism has not been previously disscussed in literature, although many simulations report a motion similar to that defined in this project as the conformational change from the semi-open to curled state.

## 9.7 Conclusions

This chapter has involved a detailed intestigation of the conformational motions of the HIV-1 PR flaps in solution. The protonation state of the catalytic residues have been studied, with the monoprotonated aspartic acid dyad producing the most stable dimer interface. Simulations with the unprotonated dyad saw one flap reaching into the protein cavity, near to the active site. This corresponded to a seperation of the catalytic residues well above that seen in the crystal structure.

The NMR study by Ishima *et al.* observes two motions; one involving the flap tips on a timescale well within 10 ns, and the other on a timescale of approximately 100 $\mu$s. The MD simulations presented in this chapter show interconversions between closed, semi-open and a curled state on a timescale within 10 ns, that fit the description of the fast motion observed by NMR.

A 10 ns simulation by Scott *et al.* sampled an opening motion to a ligand accessible state, possibly similar to the slower motion discussed by Ishima *et al.* Such a motion was not observed during 300 K MD simulations. Thermal

simulations above 400 K showed both an opening motion, and a conformation with collapsed flap tips. These could have been produced by the breakdown of secondary structure that occured at elevated temperatures. However, an RDFMD simulation, which amplified only low frequency degrees of freedom in the flap tip regions, sampled a motion similar to that seen by Scott *et al.*, and also reports the corresponding closure of the open state.

An alternative mechanism of making the active site accessible for ligand binding was sampled by three RDFMD simulations. In these, both monomers exhibit curled flap conformations, and a channel is opened between the flaps of up to 9 Å. An important feature of the flaps opening in this fashion, is that the motion appears to be an extension of the transition between the semi-open and curled conformation, rather than a lifting of one flap that is not indicated in the low temperature MD simulations, or the extensive simulations of Perryman *et al.* Further investigation should be performed into the accessibility of the active site via this opening mechanism. If this mechanism is one by which substrates are able to reach the HIV-1 PR active site, this result could be of significant importance to mutant analysis and drug design.

# Chapter 10

# Conclusions

Computational methods are able to provide atomic detail of protein motions. The timescales of many conformational changes are beyond those accessible using traditional molecular dynamics simulations, which are typically performed for tens of nanoseconds. A range of methods exist that increase the sampling of conformational events, including techniques that provide energy to help overcome potential energy barriers. Reversible digitally filtered molecular dynamics (RDFMD) and parallel tempering (PT) are two such methods, and these have been developed and compared in this thesis.

Frequency analysis is an integral part of analysing and optimising the RDFMD procedure, and this project has relied heavily on digital signal processing techniques. In Chapter 4, the Hilbert-Huang transform (HHT) has been compared to established Fourier methods, and applied to various signals extracted from MD simulations. HHT is a combination of the recently developed empirical mode decomposition (EMD) and the Hilbert transform (HT), and is shown to surpass the Fourier transform in the analysis of nonlinear and nonstationary signals.

RDFMD can be used to amplify low frequency motions associated with conformational changes. The method requires significant parameterisation, in particular of the choice of frequencies to be targeted. In Chapter 5, a procedure has been described to systematically optimise each parameter, given a measurable conformational target. This has been used to generate a protocol that efficiently

increases dihedral angle motions, without overheating the system. This has been applied to the YPGDV pentapeptide, inducing greater sampling of phase space than seen using MD of RDFMD simulation with parameters prior to optimisation.

PT is a computationally expensive equilibrium method that has been widely used to address the protein folding problem, and to test different force field parameters. By constructing an ensemble of inter-converting replicas, simulated in parallel across a range of temperatures, conformational sampling can be increased. Energy barriers are overcome more rapidly at higher temperatures, and the correct distribution of conformers is maintained by a test applied to move replicas within the ensemble. In Chapter 6, a method for determining a temperature distribution that produces a uniform mobility of replicas has been described and applied to the YPGDV system. The isomerisation of peptide bonds has been investigated, and a large proportion of *cis*-proline is seen at 300 K, in agreement with experimental data. The distributions of all-*trans* YPGDV conformations generated by long MD, PT and RDFMD have been compared, assisting in the parameterisation of the RDFMD internal temperature cap.

The methods described in this thesis have been applied to three protein systems of significant size and biological interest. T4 lysozyme, a two domain protein that exhibits a hinge motion, has been studied in Chapter 7. A distinct, closed conformation is seen to be dominant in solution, with infrequent opening events that have been characterised in detail. This is in agreement with experimental evidence of an open, ligand-accessible state occupying a few percent of solution conformations. Previously reported simulations did not observe a two-state system, and appear not to have sampled a stable, closed conformer. A parallel tempering simulation of the T4 lysozyme system has been performed, and results support those of long MD and thermal simulations. More simulation time is required for the distribution of conformers to converge.

The RDFMD protocol developed on the YPGDV system has been applied to bridging residues of T4 lysozyme that link the protein domains and inhibit the hinge opening motion. Conformational changes have been induced over far shorter

time periods than those required using molecular dynamics, and the opening and closing of the T4 lysozyme protein has been abundantly sampled. A method for the further development of RDFMD has been presented, and promising results are obtained on an example protocol in which all residues that are not involved in secondary structure units are targeted.

E. coli dihydrofolate reductase (EcDHFR) has a flexible loop region (the M20 loop) that is known to be important in the enzyme's catalytic cycle. Previously reported studies identify closed and occluded conformations of this loop, depending on the substrates bound. It has widely been assumed that the transition between these states occurs via an open conformation, that has been observed using X-ray crystallography. However, recent experimental evidence suggests that the open conformation is not present in a solution of EcDHFR and various substrates. Analysis of MD trajectories with EcDHFR complexed to folate, and NADP+ - folate has been used to explore and characterise the closed and occluded states.

A range of MD and thermal simulations of the EcDHFR apoenzyme are presented in Chapter 8. These suggest that the closed conformation is dominant in solution, with transitions to a more open state. This state is not similar to the open conformation seen in crystal structures. The EcDHFR system has been used to investigate the response of the parallel tempering algorithm to subtle changes in the temperature distribution, with mathematically predictable results. Analysis of the PT results supports the findings of the long MD and thermal simulations, but more PT simulation time is required before conformer distributions can be determined. RDFMD of the EcDHFR system has produced dramatically increased conformational sampling, observing a transition from the closed to the occluded and from the occluded to the open conformations. An open intermediate was not required to move between the closed and occluded forms. The transition between the three conformations has not been previously reported without biasing simulation towards known structures.

The HIV-1 protease dimer (HIV-1 PR) is a major drug target for the treatment

of AIDS. Much research has been performed on this system, focusing particularly on the two protein flaps that form a ligand inaccessible cavity in obtained crystal structures. These flaps must undergo a large conformational change prior to substrate binding, and there has been much debate about the mechanism by which this occurs. Two motions involving the flap residues have been identified by experimental studies, one with a timescale within 10 ns, and the other with a timescale of 100 $\mu$s that is proposed to lead to a conformation with an accessible active site.

In Chapter 9, MD simulations of HIV-1 PR are presented, that show a rapid inter-conversion between closed, semi-open and curled conformations of the flaps tips. These occur on a timescale comparable to the faster motion described using experimental techniques. Simulations at raised temperatures appear to sample a collapsed conformation that is stable when reduced in temperature, which may indicate a breakdown of the system model. This could introduce problems for a PT simulation, and PT has not been performed on this system.

RDFMD has been applied to the flap tips of HIV-1 PR, showing a significant increase in conformational sampling, often moving between the curled and closed conformations within a single set of filter applications. An opening motion reported in a previous MD study is reproduced by one of four RDFMD simulations. The corresponding closure of the open conformation is also sampled. The other three RDFMD simulations reported on HIV-1 PR also see flap separation, but this is produced via a different mechanism in which both flaps adopt a curled conformation, separating a channel, of up to 9 Åwidth, to the active site. Two of these simulations observe the closing of this open state. Further analysis of the flap movements induced by RDFMD could prove to be of great value to drug design.

Molecular dynamics simulations included in this project have been of significant length, exceeding the timescale of many recently published studies on similar sized systems. Some of the longest trajectories presented in this thesis are those of T4 lysozyme, in which a dominant, closed conformation is clearly distinguishable from infrequent opening events. Having observed and characterised these events,

motions obtained using RDFMD can be compared and validated.

Parallel tempering has been used to investigate the isomerisation of peptide bonds, producing converged populations of *cis*-proline at a range of temperatures. Such a result could not have been produced using molecular dynamics alone due to the significant potential energy barrier between isomers. In this study, almost 3000 isomerisation events have been sampled in a 20 ns PT simulation involving 1.22 $\mu$s of MD simulation.

The computational expense of the PT algorithm has limited its use with the larger protein systems investigated in this study. Initial results on both T4 lysozyme and EcDHFR support conclusions drawn from MD simulations, but convergence of the PT ensembles is outside the timescale of this project. Since PT increases the rate of conformational change by increasing temperature, all degrees of freedom are amplified. Conformational changes are induced, but these cannot always be separated from disruption of the secondary structure, or limitations of the protein force field at high temperatures.

RDFMD removes the requirement to target all degrees of freedom, and is capable of amplifying low frequency motions produced by the simulation. The protocol generated in this project was designed to maximise dihedral angle motion, and has induced significant conformational changes in each of the systems to which it has been applied. A target of this project has been the reduction of the energy that must be put into the system, by RDFMD, to induce conformational motions. Prior to optimisation, an internal temperature cap of 2000 K was required to produce results similar to those generated on YPGDV with the new parameter set and an internal temperature cap of 900 K.

The portability of the optimised RDFMD parameters is shown by the conformational motions that have been induced in a range of systems. It is very important to consider exactly what the RDFMD protocol has been designed to do; increase the backbone motions of a short peptide chain. With the T4 lysozyme system, RDFMD has been used to disrupt the interactions between residues that inhibit a conformational motion, and the filter applications are therefore not inducing the

resulting opening events directly. In contrast, the application of RDFMD to the EcDHFR M20 loop region directly induces the transition between the closed and occluded states. To investigate whether RDFMD can be used to directly amplify molecular-scale motions, all residues of T4 lysozyme that are not involved in secondary structure units were targeted. A domain opening motion was sampled, but the significant loss of β-sheet structure suggests that the protocol is not suitable for application to such a large portion of the protein. To pursue the use of RDFMD on entire protein molecules, it is likely that a lower frequency target, a reduced internal temperature cap, and longer periods of molecular dynamics between sets of filter applications, will be required to ensure the protocol is sufficiently gentle.

In summary, PT has been used to generate a converged distribution of conformations that could not have been sampled within a feasible timescale using traditional MD simulation. The computational expense of maintaining an equilibrium limits the application of the algorithm to larger systems.

The RDFMD method is unable to generate equilibrium distributions, but carries several particular strengths that answer the deficits of many other conformational analysis methods. The motions that are amplified have been produced during molecular dynamics simulation with no simplification of the force field. The method of amplification does not rely on pre-calculated information, and evolves with the trajectory. RDFMD has been shown to move between known conformations, without the use of a driving force that biases simulation towards a desired result. RDFMD simulations presented here have shown reproducible sampling of reversible events that are not accessible to long MD simulations, and that are in agreement with available experimental data.

The investigations of the conformational changes available to T4 lysozyme, EcDHFR, and HIV-1 PR provide information of significant interest that enhances the current understanding of these systems. The observation of a two-state T4 lysozyme system, the transition from the closed to occluded conformation of EcDHFR, and the accessibility of the HIV-1 PR active site via flap motions are all significant results in the field of computational chemistry.

# Appendix A

# Mathematical relationships used with the Hilbert transform

## A.1 The convolution integral

The convolution of a function, $x(t)$, and its impulse response (the function created when a unit impulse is passed through $x(t)$), $h(t)$, is represented by a star, as shown in Equation A.1.

$$y(t) = x(t) * h(t) \tag{A.1}$$

This notation indicates the evaluation of $y(t)$ is performed using the convolution integral shown in Equation A.2, in which the impulse response function is reversed, slid across the input function, and the product of the two integrated.

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \tag{A.2}$$

## A.2 Evaluation of an integral using the Cauchy principal value

The equation for the Hilbert transform contains a discontinuity at $t = u$ as shown in Equation A.3. The $P$ denotes the Cauchy principal value and indicates that to evaluate the integral it is split into two that are allowed to approach the discontinuity of the function from either side as shown in Equation A.4. It is not possible to calculate the values of each integral independently due to the discontinuity involved. The calculation of the Cauchy principal value converges however, as the values of the integrals close to, and on either side of the discontinuity, cancel out.

$$h(t) = P\int_{-\infty}^{\infty} \frac{x(u)}{\pi(t-u)}du \tag{A.3}$$

$$h(t) = \lim_{\eta \to 0}\left[\int_{-\infty}^{t-\eta} \frac{x(u)}{\pi(t-u)}du + \int_{t+\eta}^{\infty} \frac{x(u)}{\pi(t-u)}du\right] \tag{A.4}$$

# Appendix B

# Energy of a harmonic oscillator

The force experienced by a particle of mass, $m$, and position, $x$, moving with simple harmonic motion, is shown in Equation B.1, where $k$ is a force constant.

$$F(x) = -kx \tag{B.1}$$

The angular frequency (in rad $s^{-1}$), $\omega$, can be defined as in Equation B.2.

$$\omega^2 = \frac{k}{m} \tag{B.2}$$

The angular frequency is related to the frequency of the harmonic oscillator's motion (in Hz) by the expression $\omega = 2\pi\nu$. Thus the relationships shown in Equations B.3 and B.4 can be derived.

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{B.3}$$

$$k = m(2\pi\nu)^2 \tag{B.4}$$

By applying Newton's second law of motion to Equation B.1, and including the definition of angular frequency shown in Equation B.2, a differential equation can be derived as shown in B.5.

$$-kx = m\frac{d^2x}{dt}$$

$$\frac{d^2x}{dt} + \omega x = 0 \tag{B.5}$$

A solution to Equation B.5 is shown in Equation B.6, where $A$ is the amplitude of the sine wave (i.e. the length of maximum displacement) and $\phi$ is the phase. From this an expression for the particle's velocity, $v(t)$ or $\frac{dx(t)}{dt}$, can be calculated, as given in Equation B.7.

$$x(t) = A\sin(\omega t - \phi) \tag{B.6}$$

$$\frac{dx(t)}{dt} = A\omega\cos(\omega t - \phi) \tag{B.7}$$

The total energy, $E$, of the particle is the sum of the kinetic energy, $E_K$, and potential energy, $E_U$. The kinetic energy can be calculated from the velocity:

$$\begin{aligned} E_K(t) &= \frac{1}{2}mv(t)^2 \\ &= \frac{1}{2}m(A\omega\cos(\omega t - \phi))^2 \\ &= \frac{1}{2}mA^2\omega^2\cos^2(\omega t - \phi) \end{aligned} \tag{B.8}$$

The force can be expressed as the negative of the gradient of potential energy, and integration of this shows the potential energy to be equal to the sum of forces required to move the particle from rest to its current displacement:

$$E_P = -\int_0^x F dx \tag{B.9}$$

Substituting Equation B.1 into B.9 gives:

$$\begin{aligned} E_P &= \int_0^x kx dx \\ &= \left[\frac{1}{2}kx^2\right]_0^x \\ &= \frac{1}{2}mA^2\omega^2 \sin^2(\omega t - \phi) \end{aligned} \tag{B.10}$$

An expression for total energy can then be determined:

$$\begin{aligned} E &= \frac{1}{2}mA^2\omega^2 \left(\sin^2(\omega t - \phi) + \cos^2(\omega t - \phi)\right) \\ &= \frac{1}{2}mA^2\omega^2 \\ &= \frac{1}{2}A^2 k \end{aligned} \tag{B.11}$$

$$= 2m\pi^2 A^2 \nu^2 \tag{B.12}$$

The general case of Equation B.12, in which a distribution of frequencies contribute to the total energy, can be evaluated using the Fourier transform:

$$\begin{aligned} E(\nu) &= \frac{1}{2}A^2(\nu)k(\nu) \\ &= \frac{1}{2}k(\nu)\left(\int_{-\infty}^{\infty} x(t)e^{-i2\pi\nu t}dt\right)^2 \end{aligned} \tag{B.13}$$

Inserting Equation B.4, and using the spectral component, $F(\nu)$, the result of a Fourier transform for given $\nu$ can be calculated:

$$E(\nu) = 2m\pi^2\nu^2 F^2(\nu) \tag{B.14}$$

# Appendix C

# Relationships derived for parallel tempering

## C.1 Energy fluctuations

The canonical ensemble average of any physical property, $M$, from a simulation can be calculated using Equation C.1.

$$\overline{M} = \sum_j M_j P_j \tag{C.1}$$

The probability, $P_j$, of the system having an energy of $E_j$ is therefore given by the Boltzmann factor shown in Equation C.2. $\beta$ is the inverse temperature, $\frac{1}{k_B T}$, and $k_B$ is the Boltzmann constant. $Q$ exists to normalise the probability to between zero and unity, and is a sum over all possible states, termed the partition function.

$$\begin{aligned} P_j &= \frac{e^{-\beta E_j}}{\sum_j e^{-\beta E_j}} \\ &= \frac{e^{-\beta E_j}}{Q} \end{aligned} \tag{C.2}$$

The variance, $\sigma_E^2$, of the distribution of system potential energy can be described as in Equation C.3.

$$\begin{aligned}\sigma_E^2 &= \overline{(E-\overline{E})^2} \\ &= \overline{E^2}-\overline{E}^2 \end{aligned} \tag{C.3}$$

$\overline{E^2}$ is simply an average of the square of the system energy and can be described using Equation C.1. By noting a differential condition, the derivation can then continue:

$$\begin{aligned}\overline{E^2} &= \sum_j E_j^2 P_j \\ &= \frac{1}{Q}\sum_j E_j^2 e^{-\beta E_j} \\ &= -\frac{1}{Q}\frac{\partial}{\partial\beta}\sum_j E_j e^{-\beta E_j} \\ &= -\frac{1}{Q}\frac{\partial}{\partial\beta}(\overline{E}Q) \\ &= -\frac{1}{Q}\left[Q\frac{\partial\overline{E}}{\partial\beta}+\overline{E}\frac{\partial Q}{\partial\beta}\right] \\ &= -\frac{\partial\overline{E}}{\partial\beta}-\overline{E}\frac{1}{Q}\frac{\partial Q}{\partial\beta} \\ &= -\frac{\partial\overline{E}}{\partial\beta}+\overline{E}\frac{1}{Q}\sum_j E_j e^{-\beta E_j} \\ &= -\frac{\partial\overline{E}}{\partial\beta}+\overline{E}^2 \end{aligned} \tag{C.4}$$

Combining Equations C.3 and C.4 and using the chain rule to convert the partial derivative in $\beta$ to one in $T$:

$$\begin{aligned}\sigma_E^2 &= -\frac{\partial\overline{E}}{\partial\beta} \\ &= k_B T^2\left(\frac{\partial E}{\partial T}\right)_{N,V} \\ &= k_B T^2 C_V \end{aligned} \tag{C.5}$$

where $C_V$ is the molar heat capacity (i.e. the energy required to raise a mole of substance by a single Kelvin). For a harmonic oscillator, each degree of freedom contributes $\frac{k_B T}{2}$ towards the potential energy, which is therefore $\frac{Dk_B T}{2}$, where $D$ is the number of degrees of freedom. The specific heat capacity is given by:

$$\frac{\partial E}{\partial T} = \frac{DK_B}{2} \tag{C.6}$$

Inserting this into Equation C.5 gives an approximation of the order of potential energy fluctuations for a non-ideal system.

$$\begin{aligned} \sigma_E^2 &\sim k_B T^2 D k_B \\ \sigma_E &\sim k_B T \sqrt{D} \end{aligned} \tag{C.7}$$

## C.2 The parallel tempering test

A state in a general ensemble of $S$ independent replicas, can be described by the product of the Boltzmann factors for each replica as shown in Equation C.8. The partition function for the general ensemble, $Q_{REM}$, is the product of the partition functions for each replica (Equation C.9).

$$\begin{aligned} W_{REM}(X) &= W_{REM}(X_1 + X_2 + \ldots + X_S) \\ &= \frac{e^{-\beta_1 E_1} e^{-\beta_2 E_2} \ldots e^{-\beta_S E_S}}{Q_{REM}} \\ &= \frac{e^{-\sum_{s=1}^{S} \beta_s E_s}}{Q_{REM}} \end{aligned} \tag{C.8}$$

$$\begin{aligned} Q_{REM} &= \prod_{s=1}^{S} Q_s \\ &= \prod_{s=1}^{S} \sum_j e^{-\beta_s E_{s,j}} \end{aligned} \tag{C.9}$$

To ensure that an equilibrium method is created, the detailed balance condition (that forward and backward transitions are equally likely), shown in C.10, must be satisfied.

$$W_{REM}(X)w(X \rightarrow X') = W_{REM}(X')w(X' \rightarrow X) \tag{C.10}$$

When swapping replicas, it is equivalent to exchange either the coordinates (denoted by a superscript in square brackets) or temperature (denoted by a subscript). However, the nature of this exchange is yet to be defined, and a swapped replica is represented by a prime.

By arranging Equation C.10, and explicitly defining the probabilities of two states to be tested, steps can be made towards the derivation of the parallel tempering test:

$$\begin{aligned}
\frac{w(X \rightarrow X')}{w(X' \rightarrow X)} &= \frac{W_{REM}(X')}{W_{REM}(X)} \\
&= \frac{e^{-\beta_m E'^{[i]}} e^{-\beta_n E'^{[j]}} \left(\prod_{s=1}^{S-2} e^{-\beta_s E_s}\right) Q_{REM}}{e^{-\beta_m E^{[i]}} e^{-\beta_n E^{[j]}} \left(\prod_{s=1}^{S-2} e^{-\beta_s E_s}\right) Q_{REM}} \\
&= e^{-\beta_m E'^{[i]} - \beta_n E'^{[j]} + \beta_m E^{[i]} + \beta_n E^{[j]}}
\end{aligned} \tag{C.11}$$

At this stage the energy, $E$, can be defined as the sum of the kinetic energy (function of velocities), $E_K(\mathbf{v})$, and the potential energy (a function of positions), $E_P(\mathbf{r})$.

$$\begin{aligned}
\frac{w(X \rightarrow X')}{w(X' \rightarrow X)} = e^{-\beta_m\left(E_K(\mathbf{v}'^{[i]}) + E_P(\mathbf{r}'^{[i]})\right) - \beta_n\left(E_K(\mathbf{v}'^{[j]}) + E_P(\mathbf{r}'^{[j]})\right) \ldots} \\
\ldots + \beta_m\left(E_K(\mathbf{v}^{[i]}) + E_P(\mathbf{r}^{[i]})\right) + \beta_n\left(E_K(\mathbf{v}^{[j]}) + E_P(\mathbf{r}^{[j]})\right)
\end{aligned} \tag{C.12}$$

It is necessary to define the transformation that occurs when a swap is performed. This is trivial for the coordinate set, which is independent of the temperature and can be simply transferred from one replica to another. Therefore

the a coordinate set to be swapped, $\mathbf{r}'^{[i]}$, is simply replaced with a new one, $\mathbf{r}^{[j]}$. The velocities can be rescaled to the new temperature using $\mathbf{v}'^{[i]} = \sqrt{\frac{\beta_n}{\beta_m}}\mathbf{v}^{[j]}$. In this way $K(\mathbf{v}'^{[i]})$ becomes $\frac{\beta_n}{\beta_m}K(\mathbf{v}^{[j]})$ (as the kinetic energy depends on the square of the velocities). This has the effect of removing the velocities from consideration:

$$\begin{aligned}\frac{w(X \rightarrow X')}{w(X' \rightarrow X)} &= e^{-\beta_m \frac{\beta_n}{\beta_m} E_K(\mathbf{v}^{[j]}) - \beta_n \frac{\beta_m}{\beta_n} E_K(\mathbf{v}^{[i]}) + \beta_m E_K(\mathbf{v}^{[i]}) + \beta_n E_K(\mathbf{v}^{[j]})} \\ &\quad \ldots - (\beta_n - \beta_m)\left(E_P(\mathbf{r}^{[i]}) - E_P(\mathbf{r}^{[j]})\right) \\ &= e^{-(\beta_n - \beta_m)\left(E_P(\mathbf{r}^{[i]}) - E_P(\mathbf{r}^{[j]})\right)} \\ &= e^{-\Delta} \end{aligned} \tag{C.13}$$

where

$$\Delta = -(\beta_n - \beta_m)\left(E_P(\mathbf{r}^{[i]}) - E_P(\mathbf{r}^{[j]})\right)$$

## C.3 Temperature distribution of replicas

Most degrees of freedom in a molecular dynamics simulation are harmonic in nature. It is therefore not an unreasonable to approximate the total energy in a system as being of the order $Dk_BT$, where D is the number of degrees of freedom. As previously stated in Equation C.7, the energy fluctuations of a system, $\sigma_E$, are of the order $k_BT\sqrt{D}$. For a parallel tempering test to have a reasonable acceptance probability, the energy fluctuations of a system must be of a similar order to the energy difference between two replicas, $E_s - E_{s-1}$.

$$E_s - E_{s-1} \sim \sigma_{E,s} \tag{C.14}$$

More generally:

$$\begin{aligned} \frac{\Delta E_s}{\Delta s} &\sim \sigma_{E,s} \\ \frac{Dk_B\Delta T_s}{\Delta s} &\sim \sqrt{D}k_B T_s \\ \frac{\sqrt{D}\Delta T_s}{\Delta s} &\sim T_s \end{aligned} \tag{C.15}$$

Using the continuum approximation for the replica number $s$:

$$\begin{aligned} \sqrt{D}\frac{\partial T}{\partial s} &\sim T_s \\ \sqrt{D}\int \frac{1}{T_s}\partial T_s &\sim \int \partial s \\ \sqrt{D}\ln T_s &\sim s \qquad \text{(C.16)} \\ T_m &\sim e^{\frac{s}{\sqrt{D}}} \qquad \text{(C.17)} \end{aligned}$$

From Equation C.17, the temperature distribution required to maintain an acceptance proability for varied $s$ is shown to be exponential.

## C.4 The number of replicas required

The number of replicas, $S$, required to cover a temperature range from the lowest temperature in the general ensemble, $T_{MIN}$, to the highest, $T_{MAX}$, can be derived using Equation C.16, where $D$ is the number of degrees of freedom in the system.

$$\begin{aligned} S &\sim \sqrt{D}(\ln T_{MAX} - \ln T_{MIN}) \\ &\sim \sqrt{D}\ln\frac{T_{MAX}}{T_{MIN}} \end{aligned} \tag{C.18}$$

The number of replicas required therefore scales in the order of $\sqrt{D}$.

# Appendix D

# Parallel Tempering Results

Results at temperatures intermediate to those presented in Chapter 6 are shown for the 20 ns PT simulation of YPGDV, for comparison to Figures 6.10 and 6.11. As the temperature is decreased, the clouds of dots representing conformers become more defined, and gradually separate as they drop below temperatures at which transitions occur.

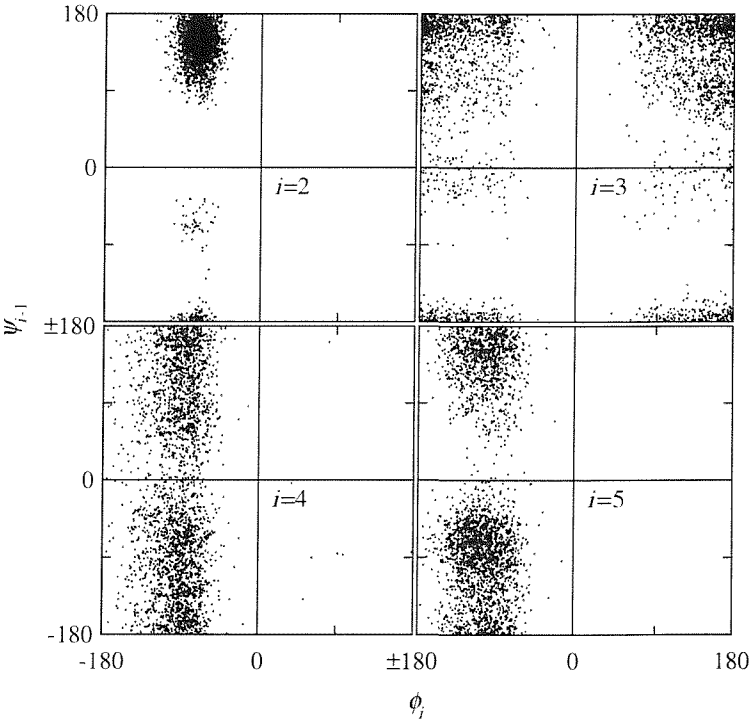



Figure D.1: REM3 498.8 K.

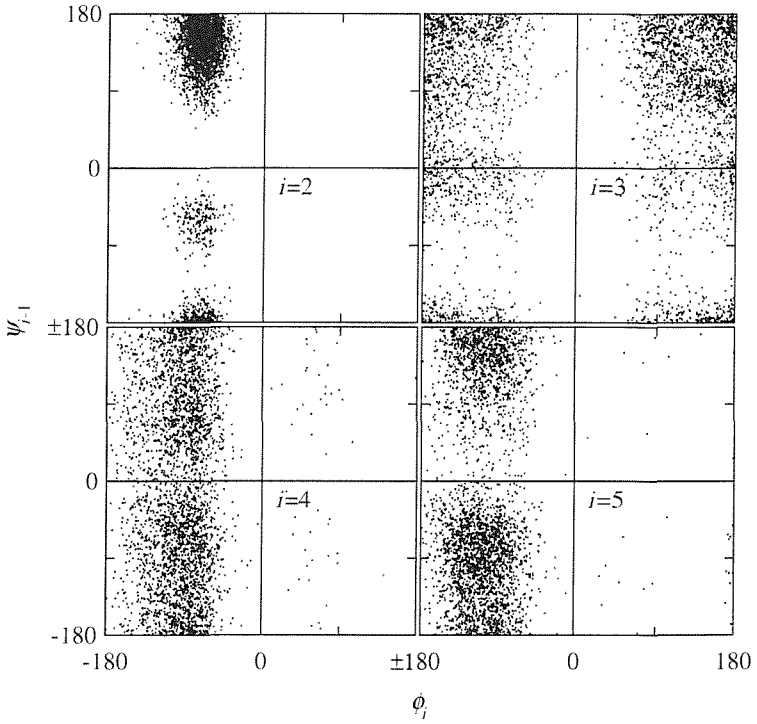



Figure D.2: REM3 753.5 K.

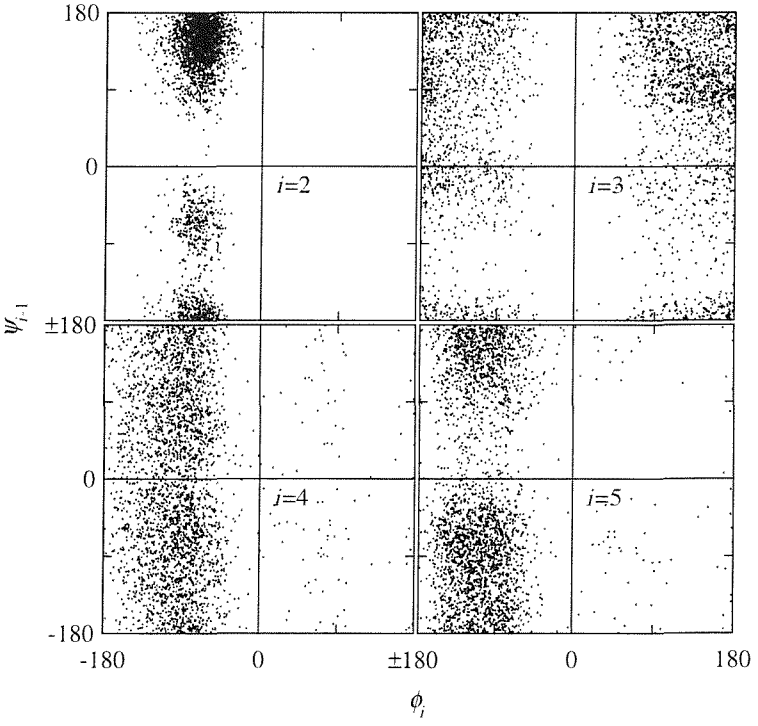



Figure D.3: REM3 1008.3 K.

# Appendix E

# The kinetics of a system returning to equilibrium

Consider a system with a proportion of states represented by $[T]$ and $[C]$ (for example, *trans* and *cis* conformations). At equilibrium, the two states exist in the proportions $T_e$ and $C_e$. If disturbed from equilibrium, the kinetics of the system can be defined using the deviations from the equilibrium proportions, $T$ and $C$, and the rate constants for the forward and backward conversion between conformers, $k_f$ and $k_b$. Thus $[T] = T_e - T$, $[C] = C_e - C$, and $T = -C$.

$$\begin{aligned} \frac{d(T_e - T)}{dt} = \frac{d(T_e)}{dt} - \frac{d(T)}{dt} &= k_f(C_e - C) - k_b(T_e - T) \\ &= k_f C_e + k_f T - k_b T_e + k_b T \\ &= k_f C_e - k_b T_e + (k_f + k_b) T \end{aligned} \tag{E.1}$$

At equilibrium:

$$\frac{d(T_e)}{dt} = k_f C_e - k_b T_e = 0 \tag{E.2}$$

By putting Equation E.2 into Equation E.1:

$$-\frac{d(T)}{dt} = (k_f + k_b) T \tag{E.3}$$

This differential equation can be analytically solved:

$$\int \left(\frac{1}{T}\right) dT = -\int \left(k_f + k_b\right) dt$$

$$\ln T = -\left(k_f + k_b\right) t + \ln T_o \tag{E.4}$$

where $T_o$ is the value of $T$ at $t = 0$. Equation E.4 can be rewritten as:

$$T = T_o e^{-\left(k_f + k_b\right)t} \tag{E.5}$$

If at $t = 0$, some shift is applied to the system that moves it away from equilibrium and entirely into one conformation, then $[T]_o = T_e + C_e = 1$, and therefore, $T_o = C_e$. Inserting this into Equation E.5:

$$[T] = T_e + C_e e^{-\left(k_f + k_b\right)t} \tag{E.6}$$

Since by definition, $C_e = 1 - T_e$:

$$[T] = T_e + (1 - T_e)\, e^{-\left(k_f + k_b\right)t} \tag{E.7}$$

Therefore, $[T]$ will exponentially decay towards its equilibrium value of $T_e$.

# References


[1] M. Karplus and J. A. McCammon, *Nature Structural Biology*, **9**, 646, (2002).

[2] K. Tai, *Biophysical Chemistry*, **107**, 213, (2004).

[3] C. Guilbert, D. Perahia and L. Mouawad, *Comput. Phys. Commun.*, **91**, 263, (1995).

[4] A. Mitsutake, Y. Sugita and Y. Okamoto, *Biopolymers*, **60**, 96, (2001).

[5] U. H. E. Hansmann, *Chem. Phys. Lett.*, **281**, 140, (1997).

[6] S. C. Phillips, M. T. Swain, A. P. Wiley, J. W. Essex and C. M. Edge, *J. Phys. Chem. B*, **107**, 2098, (2003).

[7] J. S. Bendat, *The Hilbert Transform.*, Brüel & Kjær, Nærum, Denmark.

[8] N. E. Huang, Z. Shen, S. R. Long, M. L. C. Wu, H. H. Shih, Q. N. Zheng, N. C. Yen, C. C. Tung and H. H. Liu, *Proceedings of the Royal Society of London Series a- Mathematical Physical and Engineering Sciences*, **454**, 903, (1998).

[9] J. L. Gao and D. G. Truhlar, *Ann. Rev. Phys. Chem.*, **53**, 467, (2002).

[10] W. Wang, O. Donini, C. M. Reyes and P. A. Kollman, *Annual Review of Biophysics and Biomolecular Structure*, **30**, 211, (2001).

[11] D. Reith, H. Meyer and F. Muller-Plathe, *Macromolecules*, **34**, 2335, (2001).

[12] M. Feig and C. L. Brooks, *Current Opinion in Structural Biology*, **14**, 217, (2004).

[13] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, **117**, 5179, (1995).

[14] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta and P. Weiner, *J. Am. Chem. Soc.*, **106**, 765, (1984).

[15] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. M. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. M. Wang and P. Kollman, *J. Comput. Chem.*, **24**, 1999, (2003).

[16] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, **4**, 187, (1983).

[17] A. D. Mackerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-Mccarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, **102**, 3586, (1998).

[18] W. R. P. Scott, P. H. Hunenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kruger and W. F. van Gunsteren, *J. Phys. Chem. A*, **103**, 3596, (1999).

[19] W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, **110**, 1657, (1988).

[20] T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, **98**, 10089, (1993).

[21] P. T. Vanduijnen and J. A. C. Rullmann, *Int. J. Quant. Chem.*, **38**, 181, (1990).

[22] H. Q. Ding, N. Karasawa and W. A. Goddard, *Chem. Phys. Lett.*, **196**, 6, (1992).

[23] K. Tasaki, S. Mcdonald and J. W. Brady, *J. Comput. Chem.*, **14**, 278, (1993).

[24] W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *J. Chem. Phys.*, **76**, 637, (1982).

[25] J. P. Ryckaert, G. Ciccoti and H. J. C. Berendsen, *J. Comput. Phys.*, **23**, 327, (1977).

[26] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, **79**, 926, (1983).

[27] H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *J. Phys. Chem.*, **91**, 6269, (1987).

[28] M. W. Mahoney and W. L. Jorgensen, *J. Chem. Phys.*, **112**, 8910, (2000).

[29] K. A. Sharp and B. Honig, *J. Phys. Chem.*, **94**, 7684, (1990).

[30] J. Tomasi and M. Persico, *Chem. Rev.*, **94**, 2027, (1994).

[31] C. J. Cramer and D. G. Truhlar, *Chem. Rev.*, **99**, 2161, (1999).

[32] M. L. Connolly, *Science*, **221**, 709, (1983).

[33] D. Sitkoff, K. A. Sharp and B. Honig, *J. Phys. Chem.*, **98**, 1978, (1994).

[34] M. Nina, W. Im and B. Roux, *Biophysical Chemistry*, **78**, 89, (1999).

[35] W. C. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson, *J. Am. Chem. Soc.*, **112**, 6127, (1990).

[36] D. Qiu, P. S. Shenkin, F. P. Hollinger and W. C. Still, *J. Phys. Chem. A*, **101**, 3005, (1997).

[37] D. Bashford and D. A. Case, *Ann. Rev. Phys. Chem.*, **51**, 129, (2000).

[38] J. W. Ponder and D. A. Case, *Protein Simulations*, **66**, 27, (2003).

[39] G. Moraitakis, A. G. Purkiss and J. M. Goodfellow, *Reports On Progress in Physics*, **66**, 383, (2003).

[40] R. H. Zhou, *Proteins: Struct. Funct. Genet.*, **53**, 148, (2003).

[41] L. V. Woodcock, *Chem. Phys. Lett.*, **10**, 257, (1971).

[42] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola and J. R. Haak, *J. Chem. Phys.*, **81**, 3684, (1984).

[43] T. Munakata, *Physical Review E*, **59**, 5045, (1999).

[44] *Computer Simulation of Liquids*, Oxford University Press, Oxford, UK, (2001).

[45] S. Hayward, *Proteins: Struct. Funct. Genet.*, **36**, 425, (1999).

[46] M. Gerstein, A. M. Lesk and C. Chothia, *Biochemistry*, **33**, 6739, (1994).

[47] P. Källblad and P. M. Dean, *J. Mol. Biol.*, **326**, 1651, (2003).

[48] F. Cardone, Q. G. Liu, R. Petraroli, A. Ladogana, M. D'Alessandro, C. Arpino, M. Di Bari, G. Macchi and M. Pocchiari, *Brain Research Bulletin*, **49**, 429, (1999).

[49] G. Caliskan, A. Kisliuk, A. M. Tsai, C. L. Soles and A. P. Sokolov, *J. Chem. Phys.*, **118**, 4230, (2003).

[50] F. Rosca, A. T. N. Kumar, D. Ionascu, X. Ye, A. A. Demidov, T. Sjodin, D. Wharton, D. Barrick, S. G. Sligar, T. Yonetani and P. M. Champion, *J. Phys. Chem. A*, **106**, 3540, (2002).

[51] G. Rhodes, *Crystallography Made Crystal Clear.*, 2nd ed. San Diego, California: Academic Press.

[52] X. J. Zhang, J. A. Wozniak and B. W. Matthews, *J. Mol. Biol.*, **250**, 527, (1995).

[53] N. K. Goto, N. R. Skrynnikov, F. W. Dahlquist and L. E. Kay, *J. Mol. Biol.*, **308**, 745, (2001).

[54] J. K. M. Sanders and B. K. Hunter, *Modern NMR Spectroscopy.*, 2nd ed. Oxford: Oxford University Press.

[55] A. T. Brunger and P. D. Adams, *Acc. Chem. Rev.*, **35**, 404, (2002).

[56] L. E. Kay, *Nature Structural Biology*, **5**, 513, (1998).

[57] J. P. Ma, *Current Protein & Peptide Science*, **5**, 119, (2004).

[58] A. Amadei, A. B. M. Linssen and H. J. C. Berendsen, *Proteins: Struct. Funct. Genet.*, **17**, 412, (1993).

[59] W. Wriggers and K. Schulten, *Proteins: Struct. Funct. Genet.*, **29**, 1, (1997).

[60] S. Hayward and H. J. C. Berendsen, *Proteins: Struct. Funct. Genet.*, **30**, 144, (1998).

[61] K. Hinsen, A. Thomas and M. J. Field, *Proteins: Struct. Funct. Genet.*, **34**, 369, (1999).

[62] O. S. Smart, *Chem. Phys. Lett.*, **222**, 503, (1994).

[63] S. C. Harvey and H. A. Gabb, *Biopolymers*, **33**, 1167, (1993).

[64] J. R. Gullingsrud, R. Braun and K. Schulten, *J. Comput. Phys.*, **151**, 190, (1999).

[65] B. Isralewitz, S. Izrailev and K. Schulten, *Biophysical Journal*, **72**, TU410, (1997).

[66] H. Lu and K. Schulten, *J. Mol. Graphics Modell.*, **16**, 290, (1998).

[67] R. E. Bruccoleri and M. Karplus, *Biopolymers*, **29**, 1847, (1990).

[68] R. Elber and M. Karplus, *J. Am. Chem. Soc.*, **112**, 9161, (1990).

[69] P. Dauber-Osguthorpe, C. M. Maunder and D. J. Osguthorpe, *J. Comput.-Aided Mol. Design*, **10**, 177, (1996).

[70] X. W. Wu and S. M. Wang, *J. Phys. Chem. B*, **102**, 7238, (1998).

[71] I. Andricioaei, A. R. Dinner and M. Karplus, *J. Chem. Phys.*, **118**, 1074, (2003).

[72] R. B. Sessions, P. Dauber-Osguthorpe and D. J. Osguthorpe, *J. Mol. Biol.*, **210**, 617, (1989).

[73] P. Dauber-Osguthorpe and D. J. Osguthorpe, *J. Am. Chem. Soc.*, **112**, 7921, (1990).

[74] P. Dauber-Osguthorpe and D. J. Osguthorpe, *Biochemistry*, **29**, 8223, (1990).

[75] D. J. Osguthorpe and P. Dauber-Osguthorpe, *J. Mol. Graphics*, **10**, 178, (1992).

[76] P. Dauber-Osguthorpe and D. J. Osguthorpe, *J. Comput. Chem.*, **14**, 1259, (1993).

[77] A. P. Lemon, P. Dauber-Osguthorpe and D. J. Osguthorpe, *Comput. Phys. Commun.*, **91**, 97, (1995).

[78] S. C. Phillips, J. W. Essex and C. M. Edge, *J. Chem. Phys.*, **112**, 2586, (2000).

[79] T. Nagasima, Y. Sugita, A. Mitsutake and Y. Okamoto, *Comput. Phys. Commun.*, **146**, 69, (2002).

[80] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov and P. N. Vorontsovvelyaminov, *J. Chem. Phys.*, **96**, 1776, (1992).

[81] B. A. Berg and T. Neuhaus, *Physics Letters B*, **267**, 249, (1991).

[82] Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, **329**, 261, (2000).

[83] C. Bartels, M. Schaefer and M. Karplus, *Theo. Chem. Acc.*, **101**, 62, (1999).

[84] Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, **314**, 141, (1999).

[85] A. Mitsutake and Y. Okamoto, *Chem. Phys. Lett.*, **332**, 131, (2000).

[86] N. C. Ekdawi-Sever, P. B. Conrad and J. J. De Pablo, *J. Phys. Chem. A*, **105**, 734, (2001).

[87] R. H. Zhou, B. J. Berne and R. Germain, *P. Natl. Acad. Sci. USA*, **98**, 14931, (2001).

[88] A. E. Garcia and K. Y. Sanbonmatsu, *Proteins: Struct. Funct. Genet.*, **42**, 345, (2001).

[89] A. Suenaga, *Journal of Molecular Structure-Theochem*, **634**, 235, (2003).

[90] F. Rao and A. Caflisch, *J. Chem. Phys.*, **119**, 4035, (2003).

[91] T. E. Creighton, *Proteins structures and molecular properties: Conformational properties of polypeptide chains*, 2nd ed. New York: W. H. Freeman and Company, pp. 192-193.

[92] A. P. Wiley, R. J. Gledhill, S. C. Phillips, M. T. Swain, C. M. Edge and J. W. Essex, *Hilbert-Huang Transform Engineering: The analysis of molecular dynamics simulations by the Hilbert-Huang transform*, Marcel Dekker, Inc., N.Y. In Press.

[93] S. C. Phillips, R. J. Gledhill, J. W. Essex and C. M. Edge, *J. Phys. Chem. A*, **107**, 4869, (2003).

[94] J. Y. Pan, X. H. Yan, Q. N. Zheng, W. T. Liu and V. V. Klemas, *Remote Sensing of Environment*, **84**, 53, (2003).

[95] P. Goupillaud, A. Grossman and J. Morlet, *Geoexploration*, **23**, 85, (1984).

[96] D. J. Tobias, J. E. Mertz and C. L. Brooks, *Biochemistry*, **30**, 6054, (1991).

[97] X. W. Wu and S. M. Wang, *J. Phys. Chem. B*, **104**, 8023, (2000).

[98] H. J. Dyson, M. Rance, R. A. Houghten, R. A. Lerner and P. E. Wright, *J. Mol. Biol.*, **201**, 161, (1988).

[99] W. L. Jorgensen, *MCPRO 1.4, Yale University, New Haven, CT.*, (1996).

[100] L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan and K. Schulten, *J. Comput. Phys.*, **151**, 283, (1999).

[101] M. G. Paterlini and D. M. Ferguson, *Chem. Phys.*, **236**, 243, (1998).

[102] S. E. Feller, Y. H. Zhang, R. W. Pastor and B. R. Brooks, *J. Chem. Phys.*, **103**, 4613, (1995).

[103] A. P. Wiley, M. T. Swain, S. C. Phillips, C. M. Edge and J. W. Essex, *Journal of Chemical Theory and Computation, In press.*

[104] *MATLAB 6.1.0*, The MathWorks Inc., Natick, MA, (2001).

[105] S. Melchionna, A. Luise, M. Venturoli and S. Cozzini, The dlprotein 1.0 user manual, (1998).

[106] W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577, (1983).

[107] R. H. Zhou, *P. Natl. Acad. Sci. USA*, **100**, 13280, (2003).

[108] A. Jabs, M. S. Weiss and R. Hilgenfeld, *J. Mol. Biol.*, **286**, 291, (1999).

[109] A. K. E. Svensson, J. C. O'Neill and C. R. Matthews, *J. Mol. Biol.*, **326**, 569, (2003).

[110] M. S. Weiss, A. Jabs and R. Hilgenfeld, *Nature Structural Biology*, **5**, 676, (1998).

[111] S. Fischer, R. L. Dunbrack and M. Karplus, *J. Am. Chem. Soc.*, **116**, 11931, (1994).

[112] M. G. Wu and M. W. Deem, *Mol. Phys.*, **97**, 559, (1999).

[113] R. N. Riemann and M. Zacharias, *Journal of Peptide Research*, **63**, 354, (2004).

[114] G. Zanotti, M. Saviano, G. Saviano, T. Tancredi, F. Rossi, C. Pedone and E. Benedetti, *Journal of Peptide Research*, **51**, 460, (1998).

[115] M. H. G. Wu and M. W. Deem, *J. Chem. Phys.*, **111**, 6625, (1999).

[116] H. Kessler, M. Klein, A. Müller, K. Wagner, J. W. Bats, K. Ziegler and M. Frimmer, *Angewandte Chemie-International Edition in English*, **25**, 997, (1986).

[117] X. Grabuleda, C. Jaime and P. A. Kollman, *J. Comput. Chem.*, **21**, 901, (2000).

[118] A. J. Kirby, *Nature Structural Biology*, **8**, 737, (2001).

[119] B. W. Matthews and S. J. Remington, *Proc. Nat. Acad. Sci. USA*, **71**, 4178, (1974).

[120] N. Podhipleux, J. McGuire, M. K. Bothwell and T. A. Horbett, *Colloids and Surfaces B-Biointerfaces*, **27**, 277, (2003).

[121] M. Matsumura, W. J. Becktel and B. W. Matthews, *Nature*, **334**, 406, (1988).

[122] S. Daopin, E. Soderlind, W. A. Baase, J. A. Wozniak, U. Sauer and B. W. Matthews, *J. Mol. Biol.*, **221**, 873, (1991).

[123] J. Antosiewicz, J. A. Mccammon and M. K. Gilson, *J. Mol. Biol.*, **238**, 415, (1994).

[124] D. E. Anderson, W. J. Becktel and F. W. Dahlquist, *Biochemistry*, **29**, 2403, (1990).

[125] F. Dong and H. X. Zhou, *Biophysical Journal*, **83**, 1341, (2002).

[126] M. W. F. Fischer, A. Majumdar, F. W. Dahlquist and E. R. P. Zuiderweg, *Journal of Magnetic Resonance Series B*, **108**, 143, (1995).

[127] L. P. Mcintosh, A. J. Wand, D. F. Lowry, A. G. Redfield and F. W. Dahlquist, *Biochemistry*, **29**, 6341, (1990).

[128] H. S. Mchaourab, K. J. Oh, C. J. Fang and W. L. Hubbell, *Biochemistry*, **36**, 307, (1997).

[129] Y. Chen, D. H. Hu, E. R. Vorpagel and H. P. Lu, *J. Phys. Chem. B*, **107**, 7947, (2003).

[130] F. A. A. Mulder, B. Hon, A. Mittermaier, F. W. Dahlquist and L. E. Kay, *J. Am. Chem. Soc.*, **124**, 1443, (2002).

[131] F. A. A. Mulder, A. Mittermaier, B. Hon, F. W. Dahlquist and L. E. Kay, *Nature Structural Biology*, **8**, 932, (2001).

[132] G. E. Arnold and R. L. Ornstein, *Proteins: Struct. Funct. Genet.*, **18**, 19, (1994).

[133] G. E. Arnold and R. L. Ornstein, *Biopolymers*, **41**, 533, (1997).

[134] I. Bahar, B. Erman, T. Haliloglu and R. L. Jernigan, *Biochemistry*, **36**, 13512, (1997).

[135] S. Hayward, B. L. De Groot, D. M. F. Van Aalten, A. Amadei and H. J. C. Berendsen, *Proteins: Struct. Funct. Genet.*, **31**, 116, (1998).

[136] Z. Y. Zhang, Y. Y. Shi and H. Y. Liu, *Biophysical Journal*, **84**, 3583, (2003).

[137] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Research*, **28**, 235, (2000).

[138] M. B. Bass, D. F. Hopkins, W. A. N. Jaquysh and R. L. Ornstein, *Proteins: Struct. Funct. Genet.*, **12**, 266, (1992).

[139] P. Dauber-Osguthorpe, V. A. Roberts, D. J. Osguthorpe, J. Wolff, M. Genest and A. T. Hagler, *Proteins: Struct. Funct. Genet.*, **4**, 31, (1988).

[140] M. Matsumura, J. A. Wozniak, S. Daopin and B. W. Matthews, *Journal of Biological Chemistry*, **264**, 16059, (1989).

[141] X. J. Zhang and B. W. Matthews, *Protein Sci.*, **3**, 1031, (1994).

[142] G. Vriend, *J. Mol. Graphics*, **8**, 52, (1990).

[143] W. R. P. Scott and C. A. Schiffer, *Structure*, **8**, 1259, (2000).

[144] A. L. Perryman, J. H. Lin and J. A. Mccammon, *Protein Sci.*, **13**, 1108, (2004).

[145] C. A. Fierke, K. A. Johnson and S. J. Benkovic, *Biochemistry*, **26**, 4085, (1987).

[146] M. R. Sawaya and J. Kraut, *Biochemistry*, **36**, 586, (1997).

[147] P. K. Agarwal, S. R. Billeter, P. T. R. Rajagopalan, S. J. Benkovic and S. Hammes-Schiffer, *P. Natl. Acad. Sci. USA*, **99**, 2794, (2002).

[148] J. L. Radkiewicz and C. L. Brooks, *J. Am. Chem. Soc.*, **122**, 225, (2000).

[149] Y. Y. Sham, B. Y. Ma, C. J. Tsai and R. Nussinov, *Proteins: Struct. Funct. Genet.*, **46**, 308, (2002).

[150] L. Y. Li, C. J. Falzone, P. E. Wright and S. J. Benkovic, *Biochemistry*, **31**, 7826, (1992).

[151] J. R. Schnell, H. J. Dyson and P. E. Wright, *Biochemistry*, **43**, 374, (2004).

[152] http://chem faculty.ucsd.edu/kraut/dhfr.html.

[153] C. J. Falzone, P. E. Wright and S. J. Benkovic, *Biochemistry*, **33**, 439, (1994).

[154] R. Kitahara, S. Sareth, H. Yamada, E. Ohmae, K. Gekko and K. Akasaka, *Biochemistry*, **39**, 12789, (2000).

[155] M. J. Osborne, J. Schnell, S. J. Benkovic, H. J. Dyson and P. E. Wright, *Biochemistry*, **40**, 9846, (2001).

[156] M. J. Osborne, R. P. Venkitakrishnan, H. J. Dyson and P. E. Wright, *Protein Sci.*, **12**, 2230, (2003).

[157] T. Lazaridis and M. Karplus, *Proteins: Struct. Funct. Genet.*, **35**, 133, (1999).

[158] M. Lei, M. I. Zavodszky, L. A. Kuhn and M. F. Thorpe, *J. Comput. Chem.*, **25**, 1133, (2004).

[159] T. H. Rod and C. L. Brooks, *J. Am. Chem. Soc.*, **125**, 8718, (2003).

[160] S. Ferrer, E. Silla, I. Tunon, S. Marti and V. Moliner, *J. Phys. Chem. B*, **107**, 14036, (2003).

[161] P. L. Cummins, S. P. Greatbanks, A. P. Rendell and J. E. Gready, *J. Phys. Chem. B*, **106**, 9934, (2002).

[162] P. L. Cummins and J. E. Gready, *J. Am. Chem. Soc.*, **123**, 3418, (2001).

[163] J. B. Watney, P. K. Agarwal and S. Hammes-Schiffer, *J. Am. Chem. Soc.*, **125**, 3745, (2003).

[164] P. Shrimpton and R. K. Allemann, *Protein Sci.*, **11**, 1442, (2002).

[165] Y. Y. Sham, B. Y. Ma, C. J. Tsai and R. Nussinov, *Protein Sci.*, **10**, 135, (2001).

[166] M. T. Swain, (personal communications).

[167] H. Lee, V. M. Reyes and J. Kraut, *Biochemistry*, **35**, 7012, (1996).

[168] W. R. Cannon, B. J. Garrison and S. J. Benkovic, *J. Am. Chem. Soc.*, **119**, 2386, (1997).

[169] S. Oroszlan and R. B. Luftig, *Current Topics in Microbiology and Immunology*, **157**, 153, (1990).

[170] http://xpdb.nist.gov/hivsdb/hivsdb.html.

[171] V. Zoete, O. Michielin and M. Karplus, *J. Mol. Biol.*, **315**, 21, (2002).

[172] B. Pillai, K. K. Kannan and M. V. Hosur, *Proteins: Struct. Funct. Genet.*, **43**, 57, (2001).

[173] S. Spinelli, Q. Z. Liu, P. M. Alzari, P. H. Hirel and R. J. Poljak, *Biochimie*, **73**, 1391, (1991).

[174] R. Ishima, D. I. Freedberg, Y. X. Wang, J. M. Louis and D. A. Torchia, *Structure*, **7**, 1047, (1999).

[175] E. Katoh, J. M. Louis, T. Yamazaki, A. M. Gronenborn, D. A. Torchia and R. Ishima, *Protein Sci.*, **12**, 1376, (2003).

[176] C. Debouck, J. G. Gorniak, J. E. Strickler, T. D. Meek, B. W. Metcalf and M. Rosenberg, *P. Natl. Acad. Sci. USA*, **84**, 8903, (1987).

[177] S. W. Rick, J. W. Erickson and S. K. Burt, *Proteins-Structure Function and Bioinformatics*, **32**, 7, (1998).

[178] L. David, R. Luo and M. K. Gilson, *J. Comput. Chem.*, **21**, 295, (2000).

[179] J. Trylska, P. Grochowski and J. A. Mccammon, *Protein Sci.*, **13**, 513, (2004).

[180] Y. Levy, A. Caflisch, J. N. Onuchic and P. G. Wolynes, *J. Mol. Biol.*, **340**, 67, (2004).

[181] J. Trylska, P. Bala, M. Geller and P. Grochowski, *Biophysical Journal*, **83**, 794, (2002).

[182] M. Baca and S. B. H. Kent, *Tetrahedron*, **56**, 9503, (2000).

[183] S. Piana, P. Carloni and U. Rothlisberger, *Protein Sci.*, **11**, 2393, (2002).

[184] Y. Levy and A. Caflisch, *J. Phys. Chem. B*, **107**, 3068, (2003).

[185] W. Wang and P. A. Kollman, *J. Mol. Biol.*, **303**, 567, (2000).

[186] R. Smith, I. M. Brereton, R. Y. Chai and S. B. H. Kent, *Nature Structural Biology*, **3**, 946, (1996).

[187] S. Piana and P. Carloni, *Proteins: Struct. Funct. Genet.*, **39**, 26, (2000).

[188] S. Piana, P. Carloni and M. Parrinello, *J. Mol. Biol.*, **319**, 567, (2002).

[189] K. Y. Nam, B. H. Chang, C. K. Han, S. G. Ahn and K. T. No, *Bulletin of the Korean Chemical Society*, **24**, 817, (2003).

[190] P. Carloni, U. Rothlisberger and M. Parrinello, *Acc. Chem. Rev.*, **35**, 455, (2002).