Evaluating experimental and theoretical measures of protein conformational dynamics

Benjamin Robert Stone

Doctor of Philosophy

ASTON UNIVERSITY

August 2016

ASTON UNIVERSITY

Evaluating experimental and theoretical measures of protein conformational dynamics

Ben Stone
Doctor of Philosophy

August 2016

Molecular biologists have traditionally interpreted the B-factor data of a protein crystal structure as a reflection of the protein's conformational flexibility. Crystallographers, in contrast, are wary of assigning too much significance to B-factors since they can also be attributed to processes unrelated to conformational dynamics such as experimental imprecision; crystal imperfections; or rigid body motion. In this study, the usefulness of both isotropic and anisotropic B-factors as measures of conformational dynamics were evaluated using high resolution structures. Alpha-carbon B-factor values were analysed in relation to structural properties generally accepted to be correlates of conformational variability. The influence of secondary structure, amino acid type, surface exposure, distance to the centre of mass and packing density were investigated. The results support the argument that B-factors measure conformational variability by demonstrating that atoms with the highest B-factors are typically located in regions expected to have a high degree of conformational freedom. Nevertheless, the results also highlight some of the limitations of crystallographic data. Despite using high quality crystal structures, only very general qualitative trends between B-factors values and the properties investigated could be established. Thus, B-factors appear to be influenced, to a significant degree, by the numerous sources of error in a crystallographic experiment.

By considering proteins with multiple published crystal structures, the existence of consensus B-factor profiles were identified. These consensus profiles were hypothesised to represent the dynamics within the crystal with a high degree of accuracy since much of the variation between individual experiments would be eliminated. However, when compared against measurements derived from molecular dynamic simulations, these consensus profiles only weakly correlated with the predictions of the computer models. Therefore, although there is some evidence to suggest that B-factors reflect conformational variability, B-factors cannot be assumed to be reliable descriptors of the internal dynamics of a protein within a crystal.

Keywords: protein conformational dynamics, X-ray crystallography, molecular dynamics, B-factors, atomic displacement parameters

Acknowledgements

The research presented in this thesis was part-funded by UCB Pharma.

Contents

List of figures

List of tables

Glossary

alpha-carbon

The "central" carbon atom of an amino acid that is bonded to the amine, carboxylic acid and side chain functional groups. In a protein, the alpha-carbons are part of the protein's backbone, positioned on either side of a peptide bond.

Amber99SB-ILDN

A version of the Amber99SB molecular dynamics force field with corrections for isoleucine, leucine, aspartate and asparagine (ILDN) torsion angles.

anisotropic atomic displacement parameter (AADP)

ADPs defining a trivariate Gaussian probability density function to model atomic fluctuations. Unlike refinement with isotropic B-factors, the probabilities of atomic displacements are dependent on both magnitude and direction.

anisotropic displacement covariance matrix (U^C)

Cartesian covariance matrix describing the anisotropic fluctuations of an atom.

asymmetric unit

The smallest repeating structural element of a crystal lattice. In protein crystals, the structure of the whole crystal lattice can be reconstructed from the asymmetric unit by the repeated application of *direct symmetries* i.e., rotations and translations. The choice of asymmetric unit is not necessarily unique.

atomic displacement parameter (ADP)

A measure that quantifies the uncertainty associated with determining the location of an atom within a crystal structure. ADPs define probability density functions that model the atomic fluctuations of a crystal structure.

CHARMM27

Version 27 of the Chemistry at HARvard Macromolecular Mechanics (CHARMM) force field for molecular dynamics simulations.

delta-carbon

A carbon atom of an amino acid side chain that is three chemical bonds removed from the alpha-carbon. Depending on the amino acid type, there may be zero, one or more delta-carbons.

eigenvalues of the matrix U^C(λ_{max}^{aniso} , λ_{mid}^{aniso} and λ_{min}^{aniso})

The three eigenvalues give the mean-square displacements of the anisotropic fluctuations of an atom in the directions of the respective eigenvectors. The eigenvalues λ_{max}^{aniso} and λ_{min}^{aniso} define the maximum and minimum mean-square displacement respectively.

equivalent isotropic B-factor

An "equivalent" metric to the isotropic B-factor derived from anisotropic atomic displacement parameters.

GROMOS54a7

Version 54a7 of the GROningen MOlecular Simulation (GROMOS) united atom molecular dynamics force field.

isotropic B-factor (B_{iso})

The ADP of the simplest model to account for the uncertainty in the positions of atoms in a crystal structure. Atomic fluctuations are modelled as Gaussian probability density functions under the constraint that, irrespective of direction, displacements of equal magnitude are equally likely.

median absolute deviation

A robust statistic to measure the spread across a set of values. Unlike variance and standard deviation, the mean absolute deviation is not distorted by atypical "outlier" data.

NVT

The canonical ensemble of statistical thermodynamics where the temperature, volume and number of particles remain constant.

screw-axis symmetry

A direct symmetry combing a translation with a rotation. The rotation is in a plane orthogonal to the direction of the translation so screw-axis symmetries describe helical patterns.

simple point charge (SPC)

A three site model of a water molecule used in molecular dynamics simulations. The molecular topology uses the ideal tetrahedral bond angle for water.

space group

The complete set of symmetries describing the regular arrangement of molecules within a crystal's unit cell. Space groups of protein crystals are comprised of rotations and screw-axis symmetries.

TIP3P

A three site model of a water molecule used in molecular dynamics simulations. The molecular topology uses the experimentally determined bond angle for water.

unit cell

The complete structure of a crystal lattice can be constructed from the unit cell by translations in three dimensional space. A helpful analogy is the unit cell as "building block" whereby the crystal lattice is built by stacking copies of the unit cell one on top of another. The number and arrangement of molecules within the unit cell is defined by the space group.

Acronyms

AADP Anisotropic Atomic Displacement Parameter.

ADP Atomic Displacement Parameter.

ANM Anisotropic Network Model.

COM Centre of Mass.

DSSP Define Secondary Structure of Proteins.

ENM Elastic Network Model.

FRET Fluorescence Resonance Energy Transfer.

FX Femtosecond X-ray Crystallography.

GNM Gaussian Network Model.

GROMACS GROningen MAchine for Chemical Simulation.

IUPAC International Union of Pure and Applied Chemistry.

MAD Median Absolute Deviation.

MD Molecular Dynamics.

MSF Mean Square Fluctuation.

NMA Normal Mode Analysis.

NMR Nuclear Magnetic Resonance.

OPLS-AA All Atom Optimised Potentials for Liquid Simulations.

PDB Protein Data Bank.

PME Particle Mesh Ewald.

RCSB Research Collaboratory for Structural Bioinformatics.

SASA Solvent Accessible Surface Area.

SFX Serial Femtosecond X-ray Crystallography.

SPC Simple Point Charge.

SVM Support Vector Machines.

TLS Translation Libration Screw.

VdW Van der Waals.

XFEL X-ray Free-Electron Laser.

Chapter 1

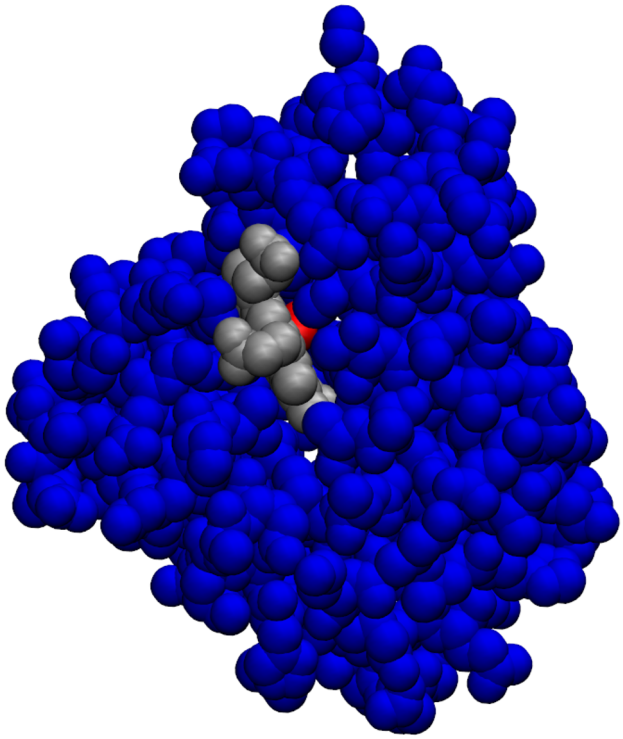
Introduction

1.1 The dynamic character of proteins

Although X-ray crystallography had been used to determine the structures of small organic and inorganic molecules since the 1920s, it was not until the 1950s that the complexities in applying the technique to biological macromolecules had been overcome (Schwarzenbach 2011). It is not an overstatement to say that the publication of the first protein structure at near atomic resolution, that of sperm whale myoglobin (Kendrew *et al.* 1958), heralded a revolution in the biological sciences. From that point onwards, biologist were able to interpret biological processes in terms of molecular interactions and formulate hypotheses on how the structures of protein molecules might relate to function. The application of X-ray crystallography to solve macromolecular structures has been so successful that, to date, over one hundred thousand structures have been deposited in the Protein Data Bank (PDB). In tandem, *structural biology*, the study of macromolecular structures, has itself become a whole new field of investigation within molecular biology. Nonetheless, although X-ray crystallography has laid the foundations of structural biology, it is not the only, or necessarily the most appropriate experimental technique for the study biological processes at the molecular level. Despite the ability of X-ray crystallography to describe proteins at the atomic level, it has one major limitation: it presents proteins as *static* molecules.

The suspicion that X-ray crystallography can only give a partial description of a protein has been present since the very first structures were examined. Continuing the work of Kendrew *et al.*, the structure of myoglobin bound to oxygen was determined by Phillips (1980) revealing two important features of the protein. Firstly, when bound to oxygen, the conformation of myoglobin is subtly altered. Secondly, and somewhat unexpectedly, the oxygen molecule is buried within the interior of the protein with no obvious route of entry or exit (figure 1.1). Thus, myoglobin cannot maintain a fixed conformation, but must instead be a flexible molecule than can adapt its conformation to accommodate the binding and release of oxy-

Figure 1.1: Van der Waals representation of the crystal structure of oxymyoglobin 1MBO (Phillips 1980). The protein atoms are coloured blue and the haem atoms grey. The oxygen molecule is coloured red and is only just visible when viewed from outside the molecule.



gen. The dynamic character of proteins was also implied by early crystal structures of the family of alcohol dehydrogenases. In both dogfish lactate dehydrogenase (Adams *et al.* 1973) and horse liver alcohol dehydrogenase (Eklund and Brändén 1979) marked differences are observed in the conformation of the enzyme when bound to its substrate in comparison to its free state. Consistent with the induced-fit model of enzyme catalysis (Koshland 1958), the active site encloses around the substrate to reorientate catalytic residues and to shield the reaction from the surrounding water molecules. The crystal structures suggest a conformational plasticity that allows the enzyme to mould itself around the substrate to facilitate catalysis. Nevertheless, the crystal structures are just static snapshots of the enzyme in two distinct conformational states, and provides no information about how the conformational transitions are achieved. A crystal structure gives the impression that proteins are rigid molecules that adopt one or more fixed conformations depending on the binding of certain ligands or physicochemical conditions. However, experiments of proteins *in vitro* suggested that the opposite is true. Proteins have a high degree of structural flexibility and, rather than being fixed, protein conformation is more accurately described as a fluid-like state that is continually changing.

Evidence from X-ray crystallography that proteins can exhibit a high degree of conformational variability posed the inevitable question of how to study and quantify the internal

motion of a protein molecule. There was an accumulating body of evidence to suggest that X-ray crystallography gives a somewhat misleading representation of the structure of macromolecules. The first insights into protein conformational dynamics came from experiments measuring the rates of hydrogen-deuterium atom exchange between protein molecules and heavy water (deuterium oxide) (Hvidt and Linderstrøm-Lang 1955; Englander *et al.* 1997). In these experiments, the hydrogen atoms of polar amino acid side chains and the peptide bond undergo exchange through chemical reactions with water. What makes this phenomenon of particular interest to biologists is that the rate of exchange depends on the degree to which the hydrogen atom is exposed to the solvent. Thus, the rate of hydrogen exchange gives an indication of the proximity of amino acids to the surface of a protein. Unexpectedly, even hydrogen atoms buried deep within the core of a protein undergo exchange reactions with water, albeit more slowly than those at the surface. Analysis of the kinetics of hydrogen exchange by Rosenberg and Chakravarti (1968) and Rosenberg and Enberg (1969) led to a dual mechanism model to explain how the solvent can penetrate the protein's interior. Rosenberg and co-workers proposed that regions of the protein may become exposed to the solvent through spontaneous local unfolding and refolding. Simultaneously, through rigid-body "segmental motions" of the protein, now commonly referred to as "breathing" movements, the protein will transiently open up to allow the solvent to diffuse into the interior. The relative contributions of these two mechanisms to the overall rate of hydrogen exchange will depend on physicochemical conditions with unfolding becoming dominant under structurally destabilising conditions such as higher temperatures. Work by Eftink and Ghiron (1975) on the dynamics of ribonuclease, probed using the alternative technique of fluorescence spectroscopy, supported the findings of the hydrogen exchange experiments. Eftink and Ghiron argue that protein structure is far removed from the "pseudo-static" models suggested by X-ray crystallography. Instead, protein conformation is in continual flux, rapidly interchanging between similar folded forms, which has the side-effect of creating short-lived channels into the protein's interior.

The concept that proteins can be porous, malleable structures answers many of the questions raised by the early crystallographic structures. The model is particularly appealing in the case of the myoglobin structures because it suggests a mechanism by which oxygen molecules can permeate the protein to gain access to the buried haem. Furthermore, the model is also consistent with current opinions concerning the processes that drive protein folding (Anfinsen 1973; Dill 1990; Leopold *et al.* 1992; Sali *et al.* 1994; Wolynes 2005; Baldwin 2007). In their fully folded states, many proteins are only marginally thermodynamically stable (Kamerzell *et al.* 2008) with overall Gibbs' free energies of folding that are typically within the range -20 to $-100\,\mathrm{kJ\,mol^{-1}}$ (Privalov and Khechinashvili 1974; Pace 1975; Dill 1990). The network of hydrogen bonds, Van der Waals (VdW) interactions and salt bridges that are established when a protein folds are enthalpically favoured. In addition, protein folding is also driven by the hydrophobic effect: an increase in the entropy of the surrounding water molecules that arises when bulky hydrophobic side chains are buried within the interior of

the protein. However, in achieving its compact fully folded form, the protein's conformational freedom is significantly reduced in comparison to its prior unfolded state. Thus, folding imposes a severe entropic penalty which is only just offset by the combination of the hydrophobic effect and stabilising non-covalent interactions. Therefore, under physiological conditions, the opposing forces that drive protein folding and unfolding are almost balanced. From a purely thermodynamic perspective, folding does not appear to favour proteins tightly constricting into one specific conformation.

The discovery that the conformations of folded proteins are only moderately stable prompts the logical question to ask why there appears to be no evolutionary pressure to increase stability. The answer could be that high thermodynamic stability may be undesirable because it would impose conformational rigidity which could have a detrimental effect on the protein's ability to function. Supporting evidence comes from studies on enzymes where mutations that increase structural stability typically reduce the enzyme's catalytic activity (Shoichet *et al.* 1995; Beadle and Shoichet 2002). The function versus stability hypothesis is also supported by studies comparing the flexibilities, stabilities and activities of homologous enzymes from bacteria adapted to extremes of temperature (Fields 2001; Jaenicke 1991). Homologous enzymes from thermophilic bacteria are generally more thermally stable than their mesophilic equivalents (Razvi and Scholtz 2006). However, improved stability through greater conformational rigidity exacts a price at mesophilic temperatures. Thermostable enzymes are typically less active at low temperatures due to the loss of conformational flexibility. Conversely, enzymes from cold adapted bacteria are less thermally stable than those from mesophilic bacteria due to the enzymes having increased flexibilities in order to function at low temperatures (Georlette *et al.* 2004). Therefore, there appears to be a three-way trade-off between a protein's structural stability, conformational flexibility and its biological activity. Evolution will favour proteins that are thermodynamically stable and fold quickly and spontaneously into a single specific shape. At the same time, proteins need a certain degree of flexibility in order to function, so the native structure cannot be so stable as to prohibit small fluctuations in conformation. Furthermore, it has been argued that marginal protein stability may be a necessary requirement for protein evolution (Taverna and Goldstein 2002; Tomatis *et al.* 2008). If proteins were too stable, it would be highly unlikely that any natural mutation would ever perturb a protein's structure and dynamics to such an extent as to alter the protein's functionality.

The dynamic character of proteins is now widely accepted as being essential for life. Protein conformational variability facilitates most, if not all, biological processes (Karplus and McCammon 1983; Teilum *et al.* 2009). Enzymes (Henzler-Wildman *et al.* 2007), receptors (Brzozowski *et al.* 1997; Prade *et al.* 1997) and transporter proteins (Hollenstein, Dawson *et al.* 2007; Hollenstein, Frei *et al.* 2007) are all examples of proteins whose conformations need to be flexible in order to achieve their biological function. Protein flexibility and its relationship to stability is also an important consideration in the study of the molecular

basis of diseases such as Alzheimer's, Parkison's, Creutzfeldt-Jakob and type II diabetes. The pathologies of these diseases are often associated with abnormal protein conformational dynamics that lead to protein misfolding and aggregation (Dobson 2004; Chiti and Dobson 2006; Herczenik and Gebbink 2008).

1.2 Defining protein conformational dynamics

Despite the central importance of protein conformational dynamics in biology, it is very difficult to quantify the movements of protein molecules by experimental methods. Proteins undergo conformational rearrangements at the atomic scale over time frames that can be as brief as a few picoseconds or as long several hours. Part of the difficulty in quantifying protein dynamics is that, in a sense, protein dynamics is a blanket term covering a wide range of different types of molecular movements. This thesis is only concerned with the *equilibrium dynamics* of proteins; that is the "steady state" conformational fluctuations of proteins at thermal equilibrium with the environment. In contrast, many biologically interesting phenomena arise through *non-equilibrium* dynamics where protein conformational change is induced through the action of some external stimulus. The stimuli can be physicochemical such as changes in temperature, pressure, pH, ionic strength or viscosity. In addition, a protein may undergo conformational change as a result of covalent modification, binding with a ligand or association with another protein. Non-equilibrium dynamics, therefore, can explain the molecular mechanisms involved in protein denaturation; enzymatic catalysis; the opening or closing of a membrane channel; or signal transduction by a receptor. For this reason, non-equilibrium dynamics are sometimes referred to as "activated processes" (Chandler 1986; Henzler-Wildman and Kern 2007) that describe the changes between two stable states. However, unlike equilibrium dynamics, non-equilibrium dynamics do not reveal anything about the inherent flexibility of protein molecules. Equilibrium dynamics are important because they can provides insights into a protein's structural stability and how a protein is able to facilitate the conformation changes induced under non-equilibrium conditions.

Equilibrium dynamics are generally divided into two categories: "fast" and "slow" dynamics. Fast dynamics are typically localised conformational changes that encompass atomic fluctuations; bond rotations that flex amino acid side chains; and the rigid body movements of elements of secondary structure. Slow dynamics usually describe global conformational rearrangements and include the concerted movements of extended structural motifs, domains and subunits. Slow dynamics also encompass the unfolding and refolding of extended regions of the polypeptide chain. Table 1.1 summarises the different types of conformational change that comprise equilibrium dynamics.

Fast and slow dynamics can also be described in terms of the energetics of protein folding. The most widely accepted model for the folding process visualises folding as a random de-

Table 1.1: Summary of the types of movement that are collectively described as equilibrium protein dynamics. The time scales and magnitudes of the displacements are only approximate and serve to give an indication of how the different types of movement compare to one another. The information in the table is adapted from Karplus and McCammon (1983), Petsko and Ringe (1984) and Henzler-Wildman and Kern (2007).

Type of dynamics	**Time scale**	**Extent (Å)**	**Description**
Fast fluctuations	fs	0.01–1	Bond vibrations, bending and rotations.
Fast collective motion	ps	0.01–10	Ring flipping and side chain flexing.
	ns		Secondary structure reorganisation and rigid body motion.
Slow collective motion	µs	$\gtrsim 10$	Global concerted movements. Domain and subunit rigid body motion.
Slow rearrangements	$\gtrsim$ms	$\gtrsim 100$	Unfolding and refolding.

cent down a funnel shaped conformational energy landscape (Leopold *et al.* 1992; Wolynes 2005). Folding does not occur through a sequence of prescribed steps. Instead, by a principal referred to as "minimal frustration" (Bryngelson and Wolynes 1987), the native folded state can be achieved via many alternative routes. Conceptually, the free energy landscape for folding is a massively high dimensional rugged funnel. The huge numbers of random unfolded conformations lie around the lip of the funnel while the native fold, the global free energy minimum, sits at the base. A folding protein descends the landscape by making small reversible conformational adjustments, favoured by a gradual lowering of free energy, that bring the conformation ever closer to the native state. Once at the bottom of the funnel, random conformational perturbations drive the protein's equilibrium dynamics rather than folding. Thus, equilibrium dynamics can be viewed in terms of a subset of the overall conformational free energy landscape; that is, an exploration of the area in the vicinity of the global minimum. Equilibrium dynamics are the transitions between the metastable conformational states (local free energy minima) that surround the global minimum. The protein never truly achieves the minimal energy "native" folded state, but fluctuates around it, adopting many near-native conformations. Fast dynamics and slow dynamics are differentiated by their positions on the free energy landscape. Fast dynamics are conformational fluctuations between free energy minima close to the global minimum separated by small energy barriers. Slow dynamics, in contrast, are represented by conformational transitions over higher energy barriers that take the protein farther from the global minimum. Figure 1.2 illustrates the concept of the folding landscape and conformational dynamics about the free energy minimum.

Figure 1.2: Stylised representation of the protein folding free energy landscape. The folding process is illustrated on the left where unfolded proteins descend the free energy "funnel" to achieve the native conformation. Fast conformational dynamics are illustrated on the right as rapid transitions exploring the free energy landscape about the global minimum.

Slow dynamics have usually been regarded as more functionally significant than fast dynamics with rapid conformational fluctuations considered to be an "unimportant" and "uninteresting" (Berendsen and Hayward 2000) aspect of protein flexibility. However, this view is not universally shared, and in their review of the study of conformational dynamics, Karplus and McCammon (1983) describe small scale fluctuations as the "lubricant" that facilitates larger scale transitions. Similarly, Teilum *et al.* (2009) view protein flexibility at all scales as being a fundamental characteristic of proteins since it allows for conformational adaptability. It should also be noted that, in many ways, the terms "slow" and "fast dynamics" are inaccurate descriptors of protein conformational flexibility. Although, generally true, there is an assumption that large scale conformational changes occur over longer time scales than smaller displacements. One counter-example is the flipping of aromatic side chains within the interior of a protein, which despite being a rapid small-scale fluctuation, is an infrequent event due to the high energy barrier associated with rotating such a bulky functional group within a conformationally restricted space (Petsko and Ringe 1984). Henzler-Wildman and Kern (2007) avoid such ambiguity by describing protein dynamics in terms of a tiered hierarchy of conformational fluctuations classified on the relative sizes of the energy barriers. Equilibrium dynamics are divided into a discrete tiers with the highest tier of dynamics (tier 2) encompassing all the small amplitude oscillations about the global energy minimum. The lower tiered dynamics (tiers 1 and 0) are associated with progressively larger energy barriers and describe collective motions or infrequent conformational rearrangements.

In describing protein dynamics, it is apparent that there are both *spatial* and *temporal* aspects to conformational flexibility. In their review of the interplay between protein flexibility and stability, Kamerzell and Middaugh (2008) rightly make the distinction between measures of protein flexibility that are dependent and independent of time. A time independent metric of flexibility simply measures the magnitude of the spatial deviations between different conformations with no consideration of the time scales involved. A time dependent metric, on the other hand, will quantify the *rate* of interconversion between the different conformational states. Thus, depending on whether time is factored into the measurements, there can potentially be confusion surrounding the concept of protein flexibility. For example, which are the more "flexible" regions of a protein? Surface amino acid side chains whose rotomers only differ by a few angstrom but exchange in picoseconds? Or domains that can undergo rigid body displacements of the order of tens or hundreds of angstrom but oscillate with periods of milliseconds? Therefore, to avoid any ambiguity, all measures of protein flexibility or rigidity need to be defined precisely. Most experimental techniques for probing protein flexibility provide no or only limited information about the temporal aspects of protein motion. Consequently, this thesis will primarily focus on spatial measures of conformational variability. The most flexible regions of a protein are defined as those regions where the positions of the atoms undergo the greatest displacements.

1.3 Measuring conformational dynamics experimentally

The complex hierarchies of molecular movements that give proteins their intrinsic flexibility makes direct measurement of protein motion extremely difficult. To date, no single experimental technique can be regarded to be universally applicable when attempting to quantify protein equilibrium dynamics. Yet, a huge body of published dynamical data has been generated from one experimental method in particular: X-ray diffraction. This may seem somewhat contradictory, since X-ray diffraction is generally considered to be a technique that reveals the *static* structures of protein molecules. Furthermore, an X-ray diffraction experiment examines proteins in their crystalline form; an environment far removed from the cell and the near physiological conditions of the early hydrogen exchange studies. Conceptually, crystalline materials are visualised as highly regular rigid structures, and are not usually associated with molecular flexibility. However, as will be outlined below, X-ray diffraction can provide insights into the conformational dynamics of protein molecules. Although the crystalline state imposes limitations on the types of dynamics that X-ray diffraction can measure, the technique has the advantage of probing molecular movements at atomic or near-atomic resolution.

1.3.1 Outline of protein X-ray crystallography

The physical principle behind X-ray crystallography is the scattering of X-rays as they pass through matter. The extent to which X-rays are scattered depends on the molecules' electron densities, since scattering arises through the absorption and emission of X-ray photons by electrons. In the case of biological macromolecules, scattering is due to carbon, nitrogen, oxygen, sulphur and phosphorous while smaller atoms, hydrogen in particular, are invisible to X-rays. Theoretically, if it were possible to direct a beam of X-rays at a single molecule and observe the directions, intensities and phases of all the scattered X-rays, then a three-dimensional map of the electron density for that molecule could be deduced. A model for the structure of the molecule could then be proposed by fitting what is known about the molecule's chemistry to the electron density map. In the case of proteins, the process of model fitting would be aided by a knowledge of primary structure, post translational modifications and the presence of bound cofactors.

Unfortunately, measuring X-ray scattering by isolated protein molecules is not currently feasible. Instead, it is more practical to grow crystals of protein molecules and to observe how X-rays are scattered by the molecules *en masse*. In a crystal, there are billions of structurally identical protein molecules aligned regularly with respect to one another across a three-dimensional lattice. When a beam of X-rays passes through a crystal, all the proteins scatter X-rays identically, and, because the molecules are aligned, the scattered X-rays combine to produce a detectable signal. The scattered X-rays can combine constructively

or destructively, depending on how the proteins are arranged across the crystal lattice. The crystal lattice, therefore, acts as a diffraction grating, and the diffracted X-rays, resulting from the superimposition of the scattered X-rays, are referred to as the *reflections* of the crystal. The structural information that can be deduced from a set of reflections is incomplete because much of what could have been derived from the original scattered X-rays is lost. More information can be recovered, however, by varying the orientation of the crystal with respect to the incident beam of X-rays to obtain multiple sets of reflections. A structure determined by X-ray crystallography cannot, therefore, be obtained from a single exposure of the crystal to X-rays. This has consequences in terms of both performing an X-ray diffraction experiment and interpreting the results. Exposure to an intense beam of X-rays will inevitably lead to structural degradation so the numbers of reflections that can be obtained from a single crystal are limited. Furthermore, sets of reflections are collected at different points in time so do not represent an instantaneous "snap shot" of the conformations of the molecules.

A structure derived by X-ray crystallography is, essentially a "double average" (Frauenfelder and Petsko 1980) across all the molecules in the crystal. The structural information derived from one particular set of reflections is a spatial average across all the molecules where the scattered X-rays combine constructively. In addition, through merging sets of reflections recorded at different time points, the molecular model constructed will be a temporal average over all the conformations adopted during the course of the experiment. The process of deriving a molecular structure from an X-ray diffraction pattern is far more technically challenging than implied by the brief summary above. Growing protein crystals of sufficient quality; deducing the phases of diffracted X-rays and deciding on the molecular model that best fits the electron density map are not trivial tasks. A more detailed account of macromolecular X-ray crystallography can be found in the following references: Durbin and Feher (1996), Blow (2002) and Drenth (2007).

1.3.2 Measuring conformational variability in crystal structures

The protein structure derived by X-ray crystallography is a consensus structure averaged over all crystallographically equivalent atoms during the course of the experiment. X-ray crystallography must, therefore, capture certain aspects of the dynamics and conformational variability of the protein molecules in the crystal. It is impossible to fit a molecular model of a protein to an electron density map precisely. There will always be a degree of uncertainty when assigning a position to every atom within the structure. Although undesirable from the perspective of structure determination, this uncertainty can be exploited when investigating conformational dynamics of proteins. Atoms whose positions cannot be established precisely might be expected to reside within the most dynamic regions of the molecule. Conversely, more rigid regions of the protein would be consistent with a greater level of accuracy. There-

fore, the uncertainty in locating atoms within a crystal structure can be used to quantify protein conformational dynamics.

There are two approaches to account for conformational variability in a crystal structure. The first, and least frequently employed, is to derive multiple models for the structure of the protein. If no single protein conformation can be fitted satisfactorily to the electron density map, then the protein is presented as an ensemble of multiple conformations each of which is consistent with the experimental data. Thus, the protein's flexibility is represented by the conformational diversity of the models. In X-ray crystallography, it is unusual for a structure to be published with multiple models for the whole protein. Instead a single model is published that includes alternate locations for the most dynamic regions of the molecule. Typically, these are alternate locations representing different rotomers for amino acid side chains. The second, and most common, approach to quantifying structural uncertainty is to assign an "uncertainty value" to the coordinates of every atom located by X-ray crystallography. These uncertainty measures are derived by interpreting the X-ray diffraction data in probabilistic terms.

The simplest probabilistic model to account for the variability in a crystal structure is the *isotropic model* where the deviations in the positions of crystallographically equivalent atoms are modelled as spherically symmetrical Gaussian probability density functions. Atoms are assumed to fluctuate with equal probability in all directions with the extent of the displacements following a normal distribution. Whilst the locations of these atoms can never be determined exactly, they are most likely to be found within some spherical region of space centred on the average coordinates. Conceptually, the greater the variability in the coordinates of the crystallographically equivalent atoms, the larger the radius of this sphere and, subsequently, the lower the degree of precision associated with locating these atoms. The isotropic model is parametrised by a single variable: $< |u|^2 >$, the mean square magnitude of the displacements of crystallographically equivalent atoms. This leads to the definition of an isotropic B-factor which, by convention, is calculated as:

$$B_{iso} = 8\pi^2 < |u|^2 > \tag{1.1}$$

Thus, an isotropic B-factor describes the "degree of indeterminacy" associated with locating the position of an atom from a X-ray diffraction pattern. At higher resolution, a more sophisticated anisotropic model of atomic displacements is frequently employed. Anisotropic refinement models the locations of atoms with trivariate Gaussian probability density functions allowing for independent fluctuations in three orthogonal spatial directions. Hence, the regions of space most likely to be occupied by the atoms are modelled as ellipsoids rather than spheres. Unlike the isotropic model, where the uncertainty in the locations the atom is parametrised by a single variable for each atom, the atom's mean-square displacement $< |u|^2 >$, the anisotropic model requires six variables per atom. Anisotropic atomic

displacement parameters (Trueblood *et al.* 1996) are typically expressed as a matrix of covariances between the displacements in the directions of the three Cartesian axes. For a displacement vector $\mathbf{u}$ with Cartesian components $(\Delta x, \Delta y, \Delta z)^T$, the Cartesian covariance matrix U^C (Trueblood *et al.* 1996) is given by:

$$U^C = \begin{pmatrix} <(\Delta x)^2> & <\Delta x \Delta y> & <\Delta x \Delta z> \\ <\Delta x \Delta y> & <(\Delta y)^2> & <\Delta y \Delta z> \\ <\Delta x \Delta z> & <\Delta y \Delta z> & <(\Delta z)^2> \end{pmatrix} \tag{1.2}$$

Matrix U^C is real and symmetric and therefore, by the Spectral Theorem, can be orthogonally diagonalised in the form:

$$U^C = RDR^T$$

where matrix R is the orthogonal matrix whose columns are the eigenvectors ($\mathbf{e_i}$) of U^C and D is the diagonal matrix of the corresponding eigenvalues (λ_i). Thus,

$$D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} <(\Delta e_1)^2> & 0 & 0 \\ 0 & <(\Delta e_2)^2> & 0 \\ 0 & 0 & <(\Delta e_3)^2> \end{pmatrix}$$

In diagonal form, the anisotropic displacements are easier to interpret geometrically. The eigenvalues are the mean-square displacements ($<(\Delta e_i)^2>$) of the atom about its mean position in the directions of the the eigenvectors. In terms of the trivariate Gaussian probability density function, the eigenvalues are the second moments along the principal axes of all the ellipsoidal surfaces of equal probability density. Hence, the eigenvectors define the *directions* of anisotropic fluctuations while the *extent* of these movements are given by the eigenvalues. Furthermore, the maximum and minimum anisotropic mean-square displacements along the orthogonal axes correspond to the largest and smallest eigenvalues respectively. The anisotropic model reduces to the isotropic model in the case where the mean-square displacements are equal in all three directions.

The degree to which an atom's movements deviate from a spherically symmetric distribution can be calculated as the ratio of the minimum and maximum eigenvalues:

$$Anisotropy = \frac{\lambda_{min}}{\lambda_{max}} = \frac{<(\Delta e_{min})^2>}{<(\Delta e_{max})^2>} \tag{1.3}$$

The anisotropy ratio, by definition, must lie in the range zero to one with the upper limit corresponding to the perfectly spherically symmetric (isotropic) distribution of atomic displacements. Atomic displacement distributions become more ellipsoidal (anisotropic) as the ratio approaches zero.

An "equivalent" to the isotropic B-factor, B_{iso}^{equiv}, can be derived from the entries in the cov-

ariance matrix U^C. The calculation approximates an isotropic mean square displacement by taking the mean of the mean-square displacements in the directions of the three eigenvectors. This quantity is equivalent to the mean of the mean-square displacements in the directions of the three Cartesian axes and, consequently, will result in the same expression as that of the isotropic B-factor if the mean-square displacements are equal in every direction.

$$\begin{aligned} B_{iso}^{eq} &= 8\pi^2 \frac{\left(<(\Delta e_1)^2> + <(\Delta e_2)^2> + <(\Delta e_3)^2>\right)}{3} \\ &= \frac{8\pi^2}{3}\left(\lambda_1 + \lambda_2 + \lambda_3\right) \\ &= \frac{8\pi^2}{3}\mathrm{Trace}(U^C) \\ &= \frac{8\pi^2}{3}\left(<(\Delta x)^2> + <(\Delta y)^2> + <(\Delta z)^2>\right) \\ &= 8\pi^2 <|u|^2> = B_{iso} \quad \text{if} \quad <(\Delta x)^2> = <(\Delta y)^2> = <(\Delta z)^2> \end{aligned} \tag{1.4}$$

Collectively, the parameters that define the isotropic and anisotropic models of atomic fluctuation are known as Atomic Displacement Parameters (ADPs). The term is rarely used in the case of isotropic ADPs which are almost always referred to as B-factors. Unfortunately, there is no consensus on the nomenclature used to describe anisotropic models. In the literature, ADP is sometimes only used in the case of anisotropic models as an abbreviation for the term Anisotropic Displacement Parameter (Trueblood *et al.* 1996). To avoid confusion, this thesis will refer to these parameters as Anisotropic Atomic Displacement Parameters (AADPs).

ADPs have traditionally been attributed to the temperature dependent oscillations of atoms under the constraints of bond geometry. Conceptually, these are the high-frequency perturbations (stretching, bending and rotation) of chemical bonds in a molecular structure. Hence, ADPs are often, incorrectly, referred to as *temperature factors*. The term temperature factor is a misnomer because thermal fluctuations cannot account for all the structural variation in a crystal. The International Union of Crystallography (IUCr) define atomic displacement vectors as deviations from the ideal lattice structure that incorporate the effects of both atomic motion and "static displacive disorder" (Trueblood *et al.* 1996; Merritt 2012). The causes of static disorder are not always apparent, but contributing factors could include irregularities in the arrangement of molecules across the crystal lattice (Petsko and Ringe 1984) or the existence of proteins "locked" in many alternate conformations (Meinhold and Smith 2005). ADPs are, in a sense, "complex error terms" encapsulating the extent to which the proposed model deviates from what can be deduced from the observed X-ray diffraction data. Therefore, there is always the risk that, in the case of a poorly refined structure, the ADPs are not a complete reflection of the protein's dynamics.

The lack of high resolution structures refined with anisotropic atomic displacement parameters has meant that structural bioinformaticians have, until recently, focused their attention almost exclusively on isotropic B-factors. It is generally assumed that atoms with high B-factors are indicative of highly dynamic regions of a protein. Even though B-factor values are not exclusively determined by the effects of atomic motion, some degree of correlation between B-factors and protein flexibility in the crystal can be expected. Nonetheless, such relationships may not, necessarily, be translatable to proteins moving freely in solution under physiological conditions. The symmetrical arrangement of proteins within a crystal is far removed from a dilute aqueous solution and the "crowded" heterogeneous environment of the cell. In addition, the extensive and highly regular protein-protein contacts of the crystal lattice can alter a protein's conformation and dynamics (Eastman *et al.* 1999; Eyal *et al.* 2005; Hinsen 2008). Proteins within a crystal lattice may also undergo certain types of rigid body motion that may have little or no relevance to conformational variability *in vivo*. For example, atomic displacements may be dominated by the effects of collective whole protein or domain oscillations governed by the arrangement of the proteins within the lattice. Unfortunately, there is no consensus on the extent to which these large-scale rigid body movements might contribute to B-factors (Kuriyan and Weis 1991; Meinhold and Smith 2005; Soheilifard *et al.* 2008).

Despite a crystal being far removed from the typical environment of a protein, there is evidence to suggest that the dynamics of proteins within crystal may resemble motion *in vivo*. Work by Norvell *et al.* (1975) demonstrated that, similar to myoglobin *in vitro*, amide groups of crystallised myoglobin can undergo hydrogen-deuterium exchange when soaked in heavy water. Furthermore, since deuterium was able to permeate into the core of the myoglobin molecules, myoglobin must have been undergoing similar "breathing" movements to those hypothesised to account for oxygen exchange *in vivo*. The native-like dynamics of myoglobin crystals are not atypical since many proteins retain some degree of functionality in the crystalline state (Mozzarelli and Rossi 1996). Proof of concept comes from studies showing that many enzymes remain catalytically active once crystallised (Makinen and Fink 1977) implying that the enzymes remain conformationally flexible within the crystal lattice. The structure of the lattice can, however, affect catalytic activity as seen in carboxypeptidase A where activity is dependent on the crystal's space group (Lipscomb 1973). Therefore, since the inhibitory effects of different lattice packing arrangements varies, the dynamics of crystalline proteins cannot always be assumed to be equivalent to the dynamics of proteins *in vivo*. In addition, the conditions under which the biological activity of a crystal is assayed may be very different to the conditions used to record the diffraction pattern. For example, to avoid damage by exposure to high intensity X-rays, a crystallographic experiment is typically performed at cryogenic temperatures. Hence, the B-factors derived from a crystal structure may be very different to the atomic fluctuations of a protein in a crystal at room temperature.

1.3.3 Alternatives to X-ray crystallography

X-ray crystallography is not always the most appropriate technique for studying how a protein's flexibility may relate to its biological function. In addition to the concerns regarding the crystalline state discussed previously, studying protein dynamics by X-ray crystallography has several major drawbacks. Foremost is the difficulty in obtaining samples of protein in sufficient quantities to yield diffracting crystals. The nature of X-ray diffraction also imposes restrictions on what can be deduced from crystallographic data. X-ray diffraction data can only report on the average conformational variability of the crystal and is unable to probe the movements of individual molecules. Quantifying conformational flexibility is limited to the fast, low amplitude equilibrium dynamics of the molecules as opposed to slower conformational rearrangements or the triggered changes of non-equilibrium dynamics. Although this thesis focuses predominately on crystallographic data, many other experimental methods can be employed to study protein conformational dynamics. Whilst by no means being exhaustive, the following section outlines some of the alternatives to classical X-ray diffraction studies.

Cryo-electron microscopy

Cryo-electron microscopy allows for the direct observation of individual protein molecules both in isolation and as components of sub-cellular structures. However, due to the lower resolution of the images, typically of the order of tens of angstrom, it is not possible to generate molecular models at the atomic scale. Consequently, cryo-electron microscopy is frequently employed to study large macromolecular complexes such as a ribosome or components of the cytoskeleton (Purdy *et al.* 2014; Bai *et al.* 2015). At this scale, cryo-electron microscopy can be used to study global conformational changes involving the rearrangement of subunits or domains. For example, conformational variability has been observed in complexes of the chaperonin proteins GroEL-GroES (Ludtke *et al.* 2004; Chen *et al.* 2006) and the enzyme pyruvate dehydrogenase (Zhou *et al.* 2001).

NMR spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy, like X-ray crystallography, is primarily perceived as a method for structural determination rather than quantifying conformational dynamics. On the contrary, since NMR signals are affected by a protein's conformational dynamics, analyses of NMR spectra can reveal much about protein motion at the atomic scale (Mittermaier and Kay 2009; Marion 2013). One of the key advantages of NMR is that macromolecules are studied in aqueous solution, which is arguably closer to the cellular environment than the crystalline state.

The application of NMR spectroscopy to probe protein dynamics is a wide-ranging field that encompasses many different techniques. The earliest NMR studies focused on spin relaxation measurements of the nuclei of $^{15}N-^{1}H$ bonds to track the fast ps − ns dynamics of the polypeptide backbone (Kay *et al.* 1989). A useful flexibility metric that can be derived from spin relaxation experiments is the order parameter, s^2, which quantifies the degree of consistency in the relative orientations of NMR detectable bonds in the structure. Order parameters are somewhat analogous to B-factors in that they can quantify the degree of conformational variability at a particular position along the protein's backbone. An amino acid that has a peptide bond with order parameters close to one would be expected to be highly constrained, while values close to zero would indicate a high degree of flexibility. For example, Feng *et al.* (1998) reported that the surface loops of the *E. coli* cold shock protein CspA have low order parameters in comparison to the rest of the protein. In addition, order parameter analysis can reveal features of functional significance. In the case of the Syrian hamster prion protein PrP, the low order parameters of the extended N-terminal region suggested that it was highly disordered and possibly had a role in the aggregation process that leads to neurodegenerative disease (Donne *et al.* 1997).

The main limitation with spin relaxation studies is that it is only possible to identify those regions of proteins that are undergoing very fast conformational fluctuations. Furthermore, unless different types of NMR probe are followed simultaneously, such as both the $^{15}N-^{1}H$ and $^{13}C-^{13}C$ bonds of the backbone, it is very difficult to establish how the protein is moving (Fischer *et al.* 1998). There are, however, NMR techniques that operate in the µs − ms time window that can provide details about biologically significant dynamics. These methods report on changes to the local chemical environment of the NMR probes and, therefore, track major conformational changes in the protein's structure. Exchange Spectroscopy (EXSY) and Carr-Purcell Meiboom-Gill (CPMG) relaxation dispersion are two widely used techniques (Hansen *et al.* 2008; Mittermaier and Kay 2009; Kleckner and Foster 2011). The application of EXSY to study the DNase domain of the bacterial toxin colicin E9 revealed the slow interconversion between two distinct conformational states and suggested that the structural rearrangements were driven by the *cis-trans* isomerisation of peptide bonds (Whittaker *et al.* 1998). Using CPMG, Eisenmesser *et al.* (2005) characterised the extended concerted dynamics of the enzyme prolyl *cis-trans* isomerase cyclophilin A and hypothesised a role for these movements in catalysis.

In addition to probing the dynamics of proteins in solution, NMR spectroscopy can also be applied to proteins in the solid-state, and NMR data of protein crystals has been used to determine whether B-factors are a true reflection of dynamics. A study by Reichert *et al.* (2012) found no significant correlation between B-factors values and a protein's internal dynamics as measured by solid state NMR. This is in direct contrast to studies comparing B-factors derived from ambient temperature crystallography to protein dynamics measured by solution NMR where the two methods are found to be in general agreement (Clore and

Schwieters 2006; Fenwick *et al.* 2014). These contradictory findings may be a consequence of the limited number of proteins studied so far by NMR in comparison to X-ray crystallography. Hopefully, as more data becomes available, it should be possible to establish whether NMR supports B-factor data or is, in fact, a far superior method for measuring dynamics at the atomic scale.

A study by Yang *et al.* (2007) took a different approach to the study of conformation dynamics using NMR. Rather than taking direct measurements, Yang *et al.* analysed the NMR structures deposited in the PDB. By measuring the variation between the multiple NMR models proposed for each PDB structure, a mean square fluctuation metric analogous to a B-factor was derived. However, over the 64 proteins analysed, the NMR metric only weakly correlated with the B-factor data. Interestingly, a stronger correlation was observed between the NMR metric and the predictions of a simple computer model (a Gaussian network model).

Although seemingly an ideal technique for studying protein dynamics, there are technical limitations preventing NMR being universally applicable. Not all atoms are detectable by NMR, and, when probing biological macromolecules, investigations are usually limited to hydrogen atoms and their interactions with the isotopes ^{13}C or ^{15}N. In addition, resolving NMR spectra to identify individual atoms imposes an upper limit on the size of the proteins that can be analysed. NMR spectra, similar to X-ray diffraction patterns, are signals that are generated by an ensemble of molecules. Thus, like X-ray crystallography, NMR cannot follow the conformational changes of individual proteins, but reports, instead, on the average dynamics of the population as a whole.

Classical spectroscopy

Classical spectroscopy is usually associated with biochemical analysis: the identification and assaying of chemical compounds through the absorption or emission spectra of certain chemical groups. As such, spectroscopy has traditionally been often overlooked in structural biology, but is now enjoying something of a renaissance through the application of spectroscopic methods to study protein dynamics. Classical spectroscopy is unique in being able to probe dynamics over a wide range of time scales under near-physiological conditions, and can, therefore, complement X-ray diffraction, NMR and cryo-electron microscopy studies. Ranan spectroscopy can measure the vibrational modes of amide bonds, aromatic residues and haem prosthetic groups within protein molecules to track conformational change (Spiro *et al.* 1990; Balakrishnan *et al.* 2008). Two dimensional infra-red (IR) spectroscopy, can probe the movements of a protein's backbone (Ganim *et al.* 2008) and has been successfully applied to characterise the structure and dynamics of the transmembrane region of a protein (Mukherjee *et al.* 2006), a structural motif that is notoriously difficult to study by crystallography or NMR. With Fluorescence Resonance Energy Transfer (FRET), often referred to as a "molecular ruler" (Stryer 1978), it is possible to study the dynamics involved with

protein-protein interactions; for example, the dimerisation of receptor protein-tyrosine phosphatase alpha in cell signalling (Tertoolen *et al.* 2001). In addition, FRET has also been applied to measure how the distances between regions within the same protein vary as the molecule undergoes conformational change. These studies are proving to be invaluable in the study of the dynamics of protein folding (Ha *et al.* 1996; Jia *et al.* 1999; Deniz *et al.* 2000; Schuler and Eaton 2008) where the high degree of structural disorder makes it impossible to study by crystallographic methods.

Mass spectrometry

Primarily an analytical technique, mass spectrometry is frequently used in conjunction with isotope exchange experiments to study conformational dynamics. In the case of hydrogen-deuterium exchange experiments, mass spectrometry can locate regions of a protein exchanging protons with the solvent with greater far greater sensitivity than is possible with the size exclusion chromatographic techniques used previously (Wales and Engen 2006; Konermann *et al.* 2011). Recently, mass spectrometry has been used to quantify conformational variability directly. Rather than undergoing fragmentation, the ionised proteins remain intact as they pass through the mass spectrometer. The trajectory through the spectrometer's electric field becomes dependent on the protein's shape as opposed to its molecular weight. Thus, in simple terms, the distribution of locations where the macromolecules hit the detector can be interpreted as a reflection of conformational variability in the protein sample (Koeniger *et al.* 2006).

Other X-ray methods

Although synonymous with crystallographic diffraction, the interaction between X-rays and matter can be exploited in other ways to study molecular structure and dynamics. Small angle X-ray scattering can provide information about the size and shapes of macromolecules, and, unlike X-ray diffraction, can be applied to both solid and liquid samples. The nanometre resolution of small angle X-ray scattering is poor compared to crystallography, but makes it possible to investigate the structure of large macromolecules and their complexes at the level of domains and subunits (Mertens and Svergun 2010; Kikhney and Svergun 2015). Working at the atomic scale, Laue X-ray crystallography is an alternative to classical X-ray diffraction where the X-ray source, rather than being "monochromatic", spans a range of wavelengths. The use of "polychromatic" X-rays means that Laue crystallography only requires a single exposure to generate a sufficient number of reflections for structural determination. Therefore, by drastically reducing the duration of an experiment, the Laue methodology allows for *time resolved crystallography*: following the conformational changes of molecules within the crystal in real time (Bourgeois *et al.* 2003; Schotte *et al.* 2003).

Laue crystallography is not the only technique that has reduced the exposure time in X-ray experiments. Femtosecond X-ray Crystallography (FX) has been made possible by the development of X-ray Free-Electron Lasers (XFELs) as sources of X-rays. An XFEL generates focused pulses of high energy X-rays to the extent that a 50 fs pulse exposes a crystal to the same number of X-ray photons as delivered by a conventional synchrotron in one second (Cohen *et al.* 2014). Exposure to a XFEL source completely annihilates the sample but, before the sample is destroyed, a high resolution diffraction pattern is produced. This is not as wasteful as it first appears since a diffraction pattern can be obtained from tiny nanocrystals consisting of only a few hundred unit cells (Hunter and Fromme 2011). However, many thousands of diffraction patterns are required to derive a model of a protein's structure. The solution is Serial Femtosecond X-ray Crystallography (SFX) where a continuous stream of nanocrystals are fed into the path of the X-rays (Chapman *et al.* 2011).

SFX with a XFEL is based on the principle of "diffraction before destruction" (Neutze *et al.* 2000; Schlichting and Miao 2012) and this has several advantages over conventional X-ray crystallography. The short duration of X-ray exposure means that there is no radiation damage (Lomb *et al.* 2011) which can be a major source of error when interpreting the diffraction pattern. In addition, since the precaution of cryogenic cooling is unnecessary with SFX, structures can be determined at room temperature and, therefore, proteins may adopt conformations that are more representative of the protein's dynamics *in vivo* (Keedy *et al.* 2015). The use of nanocrystals means that SFX can study proteins, such as membrane proteins, that are difficult to grow as large high quality crystals. This is reflected in the success of SFX in elucidating the structures of G-protein coupled receptors; for example, the human serotonin receptor (Liu *et al.* 2013); human rhodopsin (Kang *et al.* 2015); and the angiotensin II type 1 receptor (Zhang *et al.* 2015). Interestingly, Liu *et al.* (2013) made use of B-factors to illustrate the difference between the restrained transmembrane helices and the more flexible extracellular loops in the serotonin receptor. It could be argued that the combination of ambient temperatures, small crystals and minimal radiation damage makes B-factors obtained by SFX better measures of conformational dynamics than B-factors derived by traditional crystallographic methods.

The short exposure times of SFX lends itself to time resolved crystallography. The femtosecond X-ray pulses generated by a XFEL are quick enough to capture the transient intermediate conformational states of an enzymatic reaction. By carefully synchronising SFX with the activation of light sensitive proteins, two recent studies have been able to take molecular "snapshots" of the conformational changes that occur during these light activated reactions. Kupitz *et al.* (2014) detected conformational changes in the protein environment surrounding the oxygen evolving complex of photosystem II at 5 angstrom resolution. Tenboer *et al.* (2014) were able to resolve the conformational changes at the chromophore of the photoactive yellow protein at atomic resolution (1.6 angstrom). Reassuringly, the conformational changes observed in the photoactive yellow protein were consistent with the results obtained

from time resolved Laue crystallography.

1.4 Computational models of protein dynamics

Computational modelling of protein dynamics can overcome the two main limitations of experimental approaches. Firstly, a computational model gives a complete picture of the dynamics of a protein molecules; that is, at any given point in time, the locations of all atoms are known precisely. Secondly, in contrast to experimental techniques that can only report the average dynamics of a large ensemble of proteins, a computer simulation tracks the movements of individual molecules. Furthermore, in order to model dynamics, a computer simulation must also model the interactions between atoms that drive the dynamics. Therefore, the value of computer simulation is not only the ability to observe the dynamics of single molecules, but to explore how these dynamics might arise. In their review of the techniques available to study protein dynamics, Henzler-Wildman and Kern (2007) interpret the situation perceptively when they contrast experimental methods reporting "what is moving" to computational simulations that explain "why things move".

Computer simulation should not, nevertheless, be considered a panacea in the study of protein conformational dynamics. In the absence of algorithms that can accurately predict *de novo* protein folding, a computer model must be based upon an experimentally determined protein structure. Moreover, in modelling an experimental structure, it is necessary to make many simplifying assumptions. Proteins are very large molecules with complex chemistry. Even a modestly sized protein, between one and two hundred amino acids in length, will be composed of several thousand atoms. The protein's dynamics are not solely determined by the interactions between its constituent atoms, but will also involve interactions between the water, salts and other biomolecules present within the cellular environment. Incorporating every conceivable factor into a model would be computationally prohibitive, so a compromise must be made between a model's accuracy and its level of detail. Thus, a model must always be interpreted in the knowledge that, at best, the simulated dynamics are only an approximation, and, at worst, the motion may be entirely unrealistic.

Of all the approaches taken to model protein dynamics, this thesis will only consider Molecular Dynamics (MD), which describes the movements of atoms and molecules under the formalism of classical physics. On the surface, this approach may appear somewhat anachronistic, since it might be expected that quantum physics would be the appropriate framework to describe matter at the atomic scale. However, a full quantum description of all the atoms in a protein molecule is neither computationally feasible nor strictly necessary in order to model protein conformational dynamics. Under the Born-Oppenheimer approximation, electronic structure can be simplified to the point where molecules are assumed to have a fixed chemical structure. Conformational dynamics can then be described in terms of the

movements of atoms under the constraints of molecular topology; that is, the freedom that is permitted by bond geometries and steric effects. Atoms move under the laws of Newtonian mechanics under the influence of classical electrostatic forces and VdW interactions. While MD can adequately describe certain characteristics of proteins, the over-simplification of an atom's electronic structure means that many chemical phenomena, which could potentially influence dynamics, are either excluded from the model or greatly simplified. For example, no chemical reactions can take place since bonds can neither be broken nor formed. Hydrogen bonding and electron delocalisation over pi-bonds are not explicitly represented and their effects are accounted for through adjustments to bond geometries and the forces between atoms. Protein acid-base chemistry is non-existent with no dissociation of water molecules and no changes to the ionisation states of amino acid side chains.

MD is a somewhat vague umbrella term that encompasses a number of different approaches to modelling protein conformational dynamics using classical physics. The following sections outline the main areas of research that are categorised as MD.

1.4.1 All-atom simulations

An all-atom MD simulation models molecules in their entirety. Models of molecular structure, *the molecular topology*, accounts for all the bonded and non-bonded interactions of every single atom of every molecule. As a result, all-atom simulations are the most computationally demanding of the MD simulation methods. In simple terms, an all-atom simulation models the movements of molecules by solving the Newtonian equations of motion for every atom. These calculations are far from trivial, requiring the solution of large systems of coupled differential equations. In general, it is not possible to derive closed solutions for the equation of motion so numerical methods are employed instead.

The procedure of running a simulation can be divided in three distinct stages. Firstly, through consideration of molecular structure and chemistry, the equations of motions of the atoms are formulated. Secondly, the equations of motion are solved incrementally to obtain the simulation's trajectory: the time series recording the evolution of the potential energies, coordinates and velocities of every atom in the simulation. Finally, the trajectory is analysed to visualise the molecular motion and to reveal the interactions that drive these movements.

The process of translating a model of a protein into a system of equations of motion has been greatly simplified by the development of standard MD force fields for macromolecules. A force field defines how molecules are parametrised and how the forces between atoms are calculated. Although individual force fields differ in the details of how they parametrise molecules, most follow the same underlying principles. A molecular topology applies geometric constraints to bond lengths and angles depending on functional group chemistry. In proteins, for example, these constraints prevent free rotation about the peptide bond and may

also define permissible ranges for the backbone torsion angles of certain amino acid types. The forces between atoms are calculated by first classifying every atom in the molecule as belonging to a particular "atom type" that defines how it will interact with other atoms. For example, the hydrogen atoms of water molecules are parametrised differently to the hydrogen atoms of the methyl group of an alanine side chain to account for their differing polarities. This is necessary because the potential functions modelling the interactions between atoms are dependent solely upon the atoms' "types" and their positions in space.

All force field potential functions have contributions from bonded and non-bonded interactions as illustrated in equation 1.5, adapted from Karplus and McCammon (1983) and Ponder and Case (2003), where $\mathbf{r}(t)$ is a vector representing the coordinates of every atom in the simulation at a given point in time t.

$$V(\mathbf{r}(t)) = V^{bonded}(\mathbf{r}(t)) + V^{non-bonded}(\mathbf{r}(t)) \tag{1.5}$$

$$V^{bonded}(\mathbf{r}(t)) = V^{length}(\mathbf{r}(t)) + V^{angle}(\mathbf{r}(t)) + V^{dihedral}(\mathbf{r}(t)) + V^{improper}(\mathbf{r}(t)) \tag{1.6}$$

$$V^{non-bonded}(\mathbf{r}(t)) = V^{electrostatic}(\mathbf{r}(t)) + V^{VdW}(\mathbf{r}(t)) \tag{1.7}$$

As discussed previously, it is not possible to model the chemistry of a system completely, so the bonded and non-bonded potentials are greatly simplified and only consider a subset of all possible contributions. The bonded potential (equation 1.6) considers bond lengths (stretching and compression); deviations in bond angles; dihedral angles (bond rotation) and improper dihedrals (bond rotation constraints in ring systems). The non-bonded potential (equation 1.7) only considers the electrostatic and VdW interactions between pairs of atoms. The forms of both the non-bonded and bonded potentials will depend on the types of atom involved to account for factors such as differences in bond geometry or an atoms electronegativity and polarisability in a particular functional group. Furthermore, for computational efficiency, the non-bonded potential does not incorporate all possible non-bonded interactions. Beyond a certain cutoff distance, the forces between atoms are assumed to be negligible.

Since the potential is a function of atom position alone, the forces acting on the atoms are conservative, and can be derived from the potential. Thus, by Newton's third law of motion, a system of coupled differential equations expressed in terms of the accelerations of the atoms is obtained (equation 1.8). Step-wise numerical integration of the differential equations gives the velocities and coordinates of every atom modelled by the simulation.

$$-\nabla V(\mathbf{r}(t)) = \mathbf{F}(\mathbf{r}(t)) = \sum_{i=1}^{N} m_i \ddot{\mathbf{r}}_\mathbf{i}(t) \tag{1.8}$$

where i indexes the atoms from 1 to N.
m_i is the mass of the ith atom.
$\mathbf{F}$ is a vector field describing the forces acting on the atoms.
$\mathbf{r}_i(t)$ is a vector representing the coordinates of the ith atom.
$\ddot{\mathbf{r}}_\mathbf{i}(t)$ is the acceleration of the ith atom.

The summary of the simulation methodology has overlooked many technical details involved with modelling large macromolecules; for example, the handling boundary conditions and the treatment of temperature and pressure at the microscopic scale. More complete accounts of all-atom MD simulations can be found in Allen and Tildesley (1987) and Haile (1992). The basic principle, however, remains the same. A simulation repeatedly recalculates the forces acting on the atoms to incrementally adjust their positions and velocities. The thousands of atoms comprising a typical protein simulation and the small femtosecond time-steps that are necessary for accurate integration make all-atom simulation extremely computationally intensive. To date, it is only feasible to simulate the protein dynamics over nanosecond and microsecond time-scales, but, with continual advances in computer technology, millisecond simulations are now within reach (Shaw *et al.* 2009). Nonetheless, due to the short durations of all-atom simulations, the conformational dynamics studied are predominantly processes involving fast small amplitude fluctuations. Moreover, it is not yet feasible to fully explore the conformational diversity of a large ensemble of proteins that is routinely observed by experimental methods such as NMR or X-ray diffraction.

1.4.2 Coarse-grained simulations

Coarse-grained MD permit longer simulations of proteins at the expense of the level of structural detail incorporated into the model. In contrast to all-atom MD, coarse-grained MD simulations do not attempt to model every atom and bond in a molecular structure. Instead, molecules are distilled into their basic functional components; the level of abstraction dependent on which features are considered to be important. For example, united atom force fields, such as the GROMOS family (Scott *et al.* 1999), simplify proteins by incorporating hydrogen atoms into the heavy atoms of functional groups. The MARTINI force field (Marrink *et al.* 2007) goes one step further, coalescing the atoms in each amino acid into one or more spherical particles. For very large proteins, such as antibodies (Chaudhri *et al.* 2012), it is possible to apply coarse-graining at the domain and subunit level. Thus, by reducing the resolution of the molecular models, the MD calculations are simplified and longer simulations can be run. Nevertheless, there is always the risk that, with greater levels of abstraction, less is understood about the molecular interactions that drive the dynamics.

1.4.3 Harmonic approximations

An alternate approach to lessening the computational demands of running a MD simulation is to simplify the equations of motion. Under the harmonic approximation, particles are assumed to behave as simple harmonic oscillators with potential functions that vary as quadratic functions of atom displacements (Brooks and Karplus 1983; Go *et al.* 1983; Levitt *et al.* 1985). The protein is assumed to have adopted its native minimal energy conformation and the extent of its conformational dynamics are small fluctuations that never deviate too far from the minimal energy structure. Hence, in terms of the protein's funnel shaped free energy landscape, harmonic approximations model the equilibrium dynamics of the protein within a small region at the base of the funnel. The advantage of this approach is that, by the technique of Normal Mode Analysis (NMA), the simplified equations of motion can be solved to decompose the equilibrium dynamics of the protein in terms of a spectrum of vibrational modes. Thus, NMA can discriminate between those regions of a protein's structure that undergo high-frequency, low-amplitude fluctuations and those that participate in the lower frequency global collective movements. Furthermore, since the equations of motion for NMA are time independent, NMA can reveal aspects of a protein's conformational dynamics that cannot be easily explored by classical all-atom MD. In particular, the large amplitude slow global movements that occur over micro and millisecond time scales (Ma 2005; Skjaerven *et al.* 2009). Nevertheless, although NMA can fully resolve the temporal aspect of protein flexibility, the harmonic approximation severely limits the extent to which spatial conformational variability can be modelled. Thus, harmonic models may be inappropriate if the protein's dynamics involve large scale structural rearrangements.

1.4.4 Elastic network models

In theory, NMA can fully characterise the temporal dynamics of a protein by identifying regions of the protein that undergo the most significant high and low frequency oscillations. However, solving equations of motion by NMA relies on linear algebraic methods that do not scale well for all-atom models of proteins in an aqueous environment. In contrast to MD, the limiting factor in NMA calculations is not computational time but the amount of memory required by a computer in order to process a system of equations where every atom is a harmonic oscillator whose motion is coupled to the movements of every other atom. Therefore, as in the case of MD, the logical approach to make the NMA calculations tractable is to reduce the complexity of the problem by coarse-graining the structure and applying a greatly simplified potential function. Taking exactly the same approach to coarse-graining as used in MD, NMA can probe conformational dynamics at varying levels of detail, ranging between the vibrational modes of individual amino acids to the collective motions of domains and subunits. A coarse grained model whose dynamics are described by NMA is known as an Elastic Network Model (ENM) (Tirion 1996; Bahar *et al.* 1997; Haliloglu *et al.* 1997). The

name is apt, since an ENM is often described as a "bead and spring" model. The "beads" represent the protein substructures modelled (amino acids, domains or subunits) and the "springs" represent the interactions between them. The system of springs maintains a semi-rigid structure that only permits the substructures to oscillate about their average positions through small elastic deformations. An investigation by Rueda *et al.* (2007) validated the ENM approach by showing similarities between the large amplitude fluctuations predicted by ENMs and the large scale deformations observed in all-atom MD simulations.

1.4.5 Validating computer models

Although computational methods can be extremely useful in the study of protein conformational dynamics, the field in still in its infancy. There is, to date, no compelling evidence to suggest that the predictions made by a computer simulations are sufficiently reliable to replace traditional experimental techniques. Studies that attempt to validate MD force fields typically focus their attention on the dynamics of short peptides (Aliev and Courtier-Murias 2010; Beauchamp *et al.* 2012; Cino *et al.* 2012). Furthermore, these studies are primarily structural; i.e., evaluating force fields on the basis of how similar the simulated peptides' conformations are to those observed by NMR and X-ray experiments. Thus, it could be argued that these studies do not necessarily address how well MD models conformational dynamics. A MD simulation may result in a peptide adopting the "correct" ensemble of conformations, but there are no checks on whether the conformational flexibility exhibited by the simulated peptides are realistic, nor are there any assurances that the folding pathways are similar to those followed by the peptides *in vivo*. Criticism of these studies is, perhaps, unduly harsh because the computational costs of MD simulation mean that the only feasible way to systematically compare MD force fields is to simulate short peptides. In addition, since quantifying protein flexibility experimentally is inherently difficult, it is understandable that MD simulations are typically validated against static structural measurements.

In contrast to MD, the conformational dynamics predicted by NMA, and ENM in particular, have been scrutinised very closely. These simulations cannot undergo major conformational change and can, therefore, only be evaluated in terms of their accuracy in modelling small conformational fluctuations. The limited conformational freedom of an ENM lends itself to comparisons with the tightly packed and structurally uniform protein lattices of X-ray crystallography. Furthermore, the harmonic oscillations derived from an ENM are analogous to the isotropic and anisotropic B-factors derived from an X-ray diffraction experiment. Accordingly, the close correspondence between X-ray data and the fluctuations predicted by an ENM makes validating these types of simulation far simpler than for MD. Studies that have systematically evaluated different ENMs by comparison with crystallographic data include work by Kundu *et al.* (2002), Eyal *et al.* (2007), Kondrashov *et al.* (2007), Xia and Wei (2013) and Opron *et al.* (2014). Interestingly, in all of this work, the level of agreement

between the computational models and experimental data is weak. The correlation coefficients between experimental atomic fluctuations measured by X-ray crystallography and the values predicted by the ENM are typically within the range range 0.5 to 0.6. Nonetheless, as discussed previously, measures of atomic movements derived from X-ray crystallography may not necessarily be an accurate reflection of the true dynamics of the protein. Therefore, it is difficult to say with any confidence whether the discrepancies between experimental X-ray data and the predictions of a computational model are due to deficiencies with the model or are simply the result of experimental imprecision.

Chapter 2

Methods

2.1 Software

The software developed for this project was implemented in the Java (OpenJDK version 1.7) and python (version 2.7) programming languages. In addition to the core Java language, the software made use of the Apache Commons Math libraries (version 3.2) (Commons Math Developers 2013), JAMA linear algebra library (version 1.0.3) (Hicklin *et al.* 2012), EclipseLink persistence libraries (version 2.4.2) (EclipseLink Project 2013) and the HyperSQL database (version 2.2.9) (HSQL Development Group 2012). The initial stages of software development were inspired by the BioJava bioinformatics libraries (version 3.0.7) (Prlić *et al.* 2012) and these libraries were used as a reference during testing. GNU R (version 3.1.1) (R Development Core Team 2008) was used for all statistical analysis and data visualisation with the following core packages: ggplot2 (version 1.0.0) (Wickham 2009), plyr (version 1.8.1) (Wickham 2011), moments (version 0.13) (Komsta and Novomestky 2012) and mixtools (version 1.0.1) (Benaglia *et al.* 2009).

Zero was defined as a number of magnitude less than 10^{-10} in all floating point calculations that involved a comparison with zero or a rounding down to zero. Van der Waals (VdW) radii and atomic masses were derived from the chemoinformatics data of the Blue Obelisk Data Repository (BODR) (version 10) (Guha *et al.* 2006; The Blue Obelisk Group 2013).

Molecules and MD trajectories were visualised using VMD (version 1.9.1) (Humphrey *et al.* 1996).

2.2 The PDB and the formatting of PDB data files

All the crystal structures analysed in this study were obtained from the PDB (Bernstein *et al.* 1977). The scope of the PDB is very broad, accepting macromolecular structures determined by X-ray, neutron and electron diffraction crystallography or solution NMR spectroscopy. The data files deposited in the PDB record not only the molecular structures, but also detailed accounts of the experimental procedures followed in order to derive these structures. For example, a structure determined by X-ray crystallography will include information characterising the crystals and outline how the diffraction pattern was generated, recorded and analysed.

At the most basic level, PDB data files are simple plain text files intended to be read sequentially line by line. Structural and experimental data are organised into sets of *records* and *remarks* which comply with strict formatting rules. In order to accommodate the diversity of data deposited in the PDB, the PDB file specification lists over sixty different record and remark types that can be included in a file. However, only a small subset of records and remarks are found in every file; the majority only being relevant to a specific type of macromolecular structure or experimental technique.

The sequence of records and remarks in a PDB file follows a logical ordering that organises the contents of the file into three distinct sections. A PDB file begins with meta-data providing information about the authors, macromolecules and details of the experiment. Following this title section is a description of the primary structure of all proteins and/or nucleic acids and, importantly, instructions on how the structure should be interpreted. For example, details of any covalent modifications or cross linking between the residues. In X-ray structures, this section will also account for any residues that are "missing" due to being unresolved by the experiment. Finally, the PDB file ends with the structure: the three dimensional coordinates of all the atoms detected by the experiment. In the case of NMR structures, the coordinate data will usually be presented in the form of multiple models rather than a single definitive structure.

Extracts from the PDB file for the scorpion toxin protein 1AHO (Smith *et al.* 1997) are quoted in figure 2.1 as an example of how a PDB file is formatted. Each line begins with a keyword that identifies the line as a record or a remark. In figure 2.1, the first three line begin with the keyword `REMARK` followed by the number `290` indicating that the lines are remarks of type "290" that are used to list the symmetries of a crystal. In this example, the three lines combine to give a matrix of homogeneous Cartesian coordinates corresponding to the identity symmetry. The next set of lines begin with the keyword `SEQRES` identifying these as "residue sequence" records that report the macromolecule's primary structure. Reading the `SEQRES` records reveals that the scorpion toxin is a sixty-four residue, single chain protein beginning with valine and terminating at histidine.

Figure 2.1: Examples of records and remarks in the PDB file for protein 1AHO (Smith *et al.* 1997)

```
REMARK 290   SMTRY1   1  1.000000  0.000000  0.000000        0.00000
REMARK 290   SMTRY2   1  0.000000  1.000000  0.000000        0.00000
REMARK 290   SMTRY3   1  0.000000  0.000000  1.000000        0.00000
...
SEQRES   1 A   64  VAL LYS ASP GLY TYR ILE VAL ASP ASP VAL ASN CYS THR
SEQRES   2 A   64  TYR PHE CYS GLY ARG ASN ALA TYR CYS ASN GLU GLU CYS
SEQRES   3 A   64  THR LYS LEU LYS GLY GLU SER GLY TYR CYS GLN TRP ALA
SEQRES   4 A   64  SER PRO TYR GLY ASN ALA CYS TYR CYS TYR LYS LEU PRO
SEQRES   5 A   64  ASP HIS VAL ARG THR LYS GLY PRO GLY ARG CYS HIS
```

2.2.1 Developing software to process PDB data files

This project was primarily concerned with the analysis of protein structures determined by X-ray crystallography. Tables 2.1 and 2.2 give a brief descriptions of the records and remarks that must be parsed when interpreting a protein crystal structure. Many programming libraries exist that can parse PDB data files; for example, the open bioinformatics tools BioJava (Prlić *et al.* 2012) and BioPython (Cock *et al.* 2009). Unfortunately, these software tools are designed for general use and lack the functionality to extract the more specialist crystallographic information from PDB data files. Furthermore, after experimenting with BioJava, it was decided that it was easier to write new software rather than re-engineer an existing software project. Developing a new PDB file parser, however, posed a dilemma. Creating a monolithic parser to process everything found within a PDB data file would be both infeasible and inefficient. Conversely, building a parser with a limited scope, processing only a small subset of records and remarks, could lead to software that is difficult to update when additional functionality is required. The solution was to design a flexible PDB file parser to which functionality could be added or removed as required.

Taking inspiration from the "chain of responsibility" software design pattern (Gamma *et al.* 1995), a light-weight PDB file parser was implemented that simply inspected every record and remark within a PDB file. The responsibility for analysing the data was delegated to pluggable modules. Each module was designed for a specific function; for example, listing the symmetries of the protein crystal or identifying the presence of modified amino acids in a structure. In this way, the functionality of the PDB file parsing software could be tailored for a specific purpose by including the relevant data processing modules.

Table 2.1: Overview of the important PDB file remarks necessary for processing protein crystal structures

Remark	Description
`REMARK 200`	Information relating to the crystallographic experiment. For example, the temperature under which the X-ray diffraction data was recorded.
`REMARK 290`	The symmetries of the unit cell. Applying these matrices to the coordinates of the atoms in the asymmetric unit reconstructs the arrangement of the molecules within the crystallographic unit cell.
`REMARK 465`	Residues that are missing from the structure due to being unresolved by crystallography.

Table 2.2: Overview of the important PDB file records necessary for processing protein crystal structures

Record	Description
HEADER	Name, deposition date and unique four character identifier for the structure.
KEYWDS	Keywords relevant to the structure.
COMPND	All the proteins present in the structure and their subunits if composed of more than one polypeptide chain.
SEQRES	The primary structure of all protein chains in the structure.
MODRES	Locates and describes all modified amino acid residues within the structure.
SSBOND	Lists all pairs of cysteine residues linked by disulphide bonds.
CRYST1	The geometry of the crystallographic unit cell. Specifies the lengths and angles of the parallelepiped representation of the unit cell and the space group classification of the crystal.
SCALE[1-3]	The "scale matrix" that defines the transform between Cartesian and crystallographic coordinates. The scale matrix transform, when combined with the symmetries listed in the REMARK 290 section, can be used to reconstruct the arrangement of molecules across the entire crystal lattice.
ORIGX[1-3]	An orthogonal transform that will convert the coordinates to the form that was originally submitted to the PDB. The transform may realign the unit cell if none of the lattice vectors coincide with the Cartesian axes. Useful when aligning unit cells to the periodic boundaries of a molecular dynamics simulation box.
NUMMDL	Lists the number of models presented for the structure. Multiple models are rare when a structure has been determined by crystallography, but it is typical for an NMR structure to contain ten or more models.
MODEL/ENDMDL	Records defining the beginning and end of the coordinate data for each model in PDB file containing multiple models.
ATOM	Atom coordinates for the protein molecules of the asymmetric unit and the associated isotropic B-factors and occupancy values.
HETATM	Identical to ATOM records for the atoms of non-protein molecules in the structure.
ANISOU	The Cartesian coordinate covariance matrix modelling the anisotropic displacements of an atom. Only present if the crystal structure has been refined using the anisotropic model of atomic displacements.
TER	Identifies the end of the structural data for a polypeptide chain.

2.2.2 Standardising structural data

It is unusual for a PDB file to define a single conformation for the three dimensional structure of a protein. A PDB file may present a protein's structure in the form of multiple models, each representing a valid conformation for the protein, consistent with the experimental data. However, in PDB files of protein crystals, the structures are very rarely presented in the form of multiple models. Instead, it is more common to describe a protein's conformational variability in terms of small deviations at the residue level. The reason being that, in a crystal structure, the adoption of multiple conformations is usually only confined to a few residues. An example of how multiple conformations are presented in a PDB file is illustrated for a glutamate residue in the scorpion toxin protein 1AHO (figure 2.2). In this example, there are four atoms in the glutamate side chain that are assigned two alternate sets of coordinates. The different positions for the atoms are differentiated by a label that prefixes the residue's name and, in this case, the two locations are labelled "A" and "B". The likelihood that a given atom will be located at either one of these positions is given by an occupancy value. For the four glutamate atoms, occupancy values of 0.5 are assigned to both locations and, therefore, all four atoms are equally likely to be positioned at either location. There is, unfortunately, the potential for ambiguity when specifying multiple conformations using alternate atom locations. Should atoms be assumed to occupy their alternate locations independently of one another? Or, will bonding constraints result in the atoms occupying each alternate location collectively? In the case of the glutamate side chain, this is the difference between the alternate atom locations representing two and $2^4 = 16$ different conformational states.

Figure 2.2: Specification of the two alternate conformations of the glutamate residue at position 24 in the PDB file for protein 1AHO (Smith *et al.* 1997)

```
--------------------------------------------------------------------------------
 1       2   3  4 5  6  7          8      9       10     11    12            13
--------------------------------------------------------------------------------
ATOM    334  N   GLU A  24      -4.881   3.516  -5.341  1.00  7.08           N
ATOM    335  CA  GLU A  24      -5.136   2.127  -5.742  1.00  7.97           C
ATOM    336  C   GLU A  24      -4.210   1.172  -5.012  1.00  7.47           C
ATOM    337  O   GLU A  24      -4.678   0.155  -4.494  1.00  8.46           O
ATOM    338  CB  GLU A  24      -4.965   1.952  -7.249  1.00 11.82           C
ATOM    339  CG AGLU A  24      -5.938   2.589  -8.197  0.50 14.37           C
ATOM    340  CG BGLU A  24      -5.083   0.492  -7.667  0.50 15.77           C
ATOM    341  CD AGLU A  24      -5.450   2.702  -9.627  0.50 17.13           C
ATOM    342  CD BGLU A  24      -5.516   0.208  -9.078  0.50 17.79           C
ATOM    343  OE1AGLU A  24      -4.535   1.980 -10.096  0.50 18.86           O
ATOM    344  OE1BGLU A  24      -5.560   1.170  -9.874  0.50 19.33           O
ATOM    345  OE2AGLU A  24      -5.987   3.569 -10.363  0.50 19.96           O
ATOM    346  OE2BGLU A  24      -5.825  -0.966  -9.404  0.50 19.68           O
ATOM    347  H   GLU A  24      -4.552   4.086  -5.967  1.00  6.89           H
ATOM    348  HA  GLU A  24      -6.084   1.908  -5.518  1.00  8.29           H
```

Key to the columns:

1 Record type (`ATOM`).

2 Unique numerical identifier for the atom.

3 Name of the atom (the atom's "type").

4 Alternate location label (`A`, `B` or blank for this residue).

5 Residue name (`GLU` - glutamate).

6 Chain identifier (chain `A`).

7 Residue number in the chain (`24`).

8-10 X, Y and Z coordinates for the atoms.

11 Occupancy value (value from 0.0 to 1.0).

12 The isotropic B-factor for the atom.

13 The atom's element.

The presence of atoms with alternate locations in a crystal structure posed a problem for this project. Specifically, because this thesis focused on the analysis of high resolution crystallographic data, many of the crystal structures under consideration did not assign atomic coordinates uniquely. In a structure refined with alternate atom locations, the permutations of all possible combinations of coordinates would make a thorough analysis of each distinct protein conformation prohibitive. For example, the structure 2FG1 (Cuff *et al.* 2005), a protein of unknown function produced by *Bacteroides thetaiotaomicron*, the atoms in 25 of its 158 residues are resolved with alternate locations. Of these 25 residues, 24 contain atoms with two possible sets of coordinates while one reside has atoms with three sets of coordinates. Assuming that the alternate locations should be grouped at the residue level, the total number of different protein conformers is:

$$3 \times 2^{24} > 10^7 \text{ distinct conformations}$$

If atoms are assumed to occupy alternate locations independently then, since there are 183 atoms with two sets of coordinates and 7 atoms with three sets of coordinates, the number of permutations explodes exponentially to:

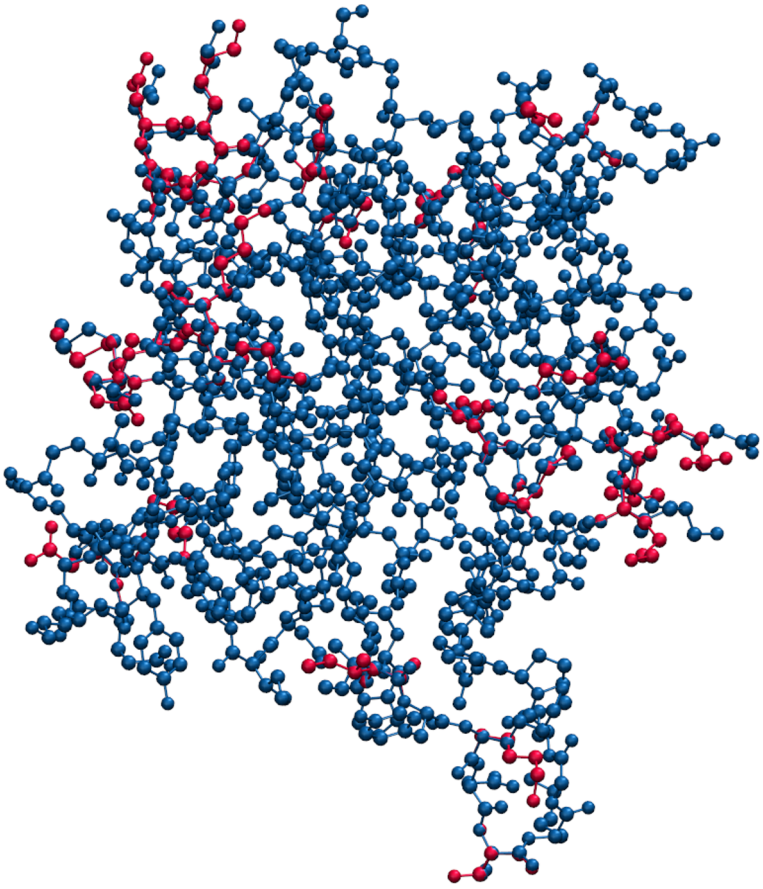
$$3^7 \times 2^{183} > 10^{58} \text{ distinct conformations}$$

As a compromise, only three protein conformers were considered:

- A maximum occupancy conformer where all atoms take the highest occupancy coordinates.
- A minimum occupancy conformer where all atoms take the lowest occupancy coordinates.
- A conformer where all coordinates are calculated as an occupancy value weighted mean.

The calculation of the weighted mean coordinates occasionally resulted in amino acid side chains with chemically impossible locations for atoms. For this reason, only the maximum and minimum occupancy conformations were considered when analysing protein structures. In the situation where two or more alternate locations had equal occupancy values, atoms were assigned on the alphabetic ordering of the alternate location identifiers. Atoms labelled with the the identifier "A" were assigned to the maximum occupancy structure while those atoms having an identifier coming last in the alphabetic sequence were assigned to the minimum occupancy structure. Although a somewhat arbitrary assignment, this allocation ensured that the maximum and minimum occupancy structures would be different at all locations where alternate sets of atomic coordinates were defined. For example, for the glutamate atoms in figure 2.2, the atoms labelled as "A" would be assigned to the max-

Figure 2.3: Superimposition of the minimum and maximum atom occupancy conformers of protein 2FG1 (Cuff *et al.* 2005). The conformer where all atoms are in their maximum occupancy positions is coloured red and the minimum occupancy conformer is coloured blue.



imum occupancy conformer while those labelled as "B" would be assigned to the minimum occupancy conformer.

Figure 2.3 compares the minimum and maximum occupancy conformers for the protein 2FG1 (Cuff *et al.* 2005). Surprisingly, given the high proportion of residues assigned multiple sets of coordinates, the superimposition of the two conformers reveals only minor differences between them. The root mean square deviation between all atoms in these conformers is 0.72 Å and, when only the backbone atoms are included in the calculation, the deviation is close to zero at 0.084 Å.

It could be argued that alternate location data could itself be used to derive a measure of local flexibility in a protein. However, in comparison with atomic displacement data, there is an insufficient number of atoms with alternate locations for a rigorous analysis. Alternate location data has limited value at present, but a comprehensive investigation may become feasible in the future as more high resolution crystal structures are deposited in the PDB.

2.2.3 Preparing PDB data files for analysis

PDB files were reformatted to remove all solvent and non-protein molecules. For crystal structures that had been resolved with atoms in multiple locations, the maximum and minimum occupancy conformers where derived as described previously. The coordinates of all missing atoms were assigned using Modeller (version 9.14) (Sali and Blundell 1993) by applying the *all-hydrogen* topology. Any structures with residues unresolved by crystallography were made complete through the generation of a single model with Modeller's *automodel* functionality. The PDB file's `SEQRES` records and missing residue remarks (`REMARK 465`) were used to locate any breaks in the structure and to identify which amino acids needed to be inserted to make the structure complete. The positions of all other atoms in the protein were held fixed when missing atoms were added so as not to perturb those atoms whose positions were known. Disulphide bonds were added if present in the `SSBOND` records of the original PDB file.

The complete protein structures created by Modeller were merged with the original PDB files to reincorporate important structural and crystallographic information such as the specification of the unit cell; crystal symmetries; and the isotropic and anisotropic displacement parameters for atoms. Any structural modifications made by Modeller, such as the conversion of selenomethionine (MSE) to methionine and changes to the protonation state of histidine were reversed to ensure consistency with the original crystallographic data.

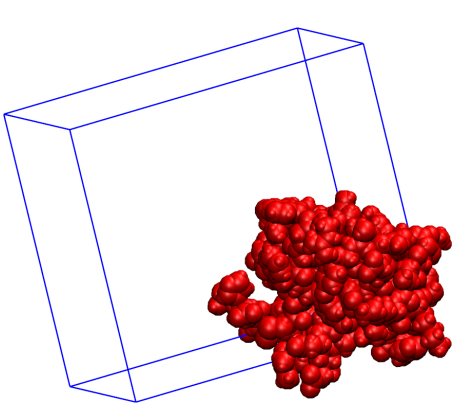
2.2.4 Reassembling the crystallographic unit cell

A PDB file of a protein crystal structure only publishes the atomic coordinates of the molecules that comprise the asymmetric unit of the crystal lattice. For most molecular biologists, the structures of the proteins in the asymmetric unit are sufficient. The asymmetric unit, however, does not provide any information about how the proteins are arranged across the crystal lattice. Therefore, it is impossible to accurately measure properties such as a protein's Solvent Accessible Surface Area (SASA) or atom packing density from the asymmetric unit alone. In order to account for the protein-protein interactions of the crystal in structural bioinformatics calculations, the crystallographic unit cell was reconstructed from the asymmetric unit. The unit cell was reconstructed using the crystal geometry and symmetry data recorded in the PDB file. The symmetries of the unit cell were parsed from the `REMARK 290` records of the PDB file and the scale matrix was obtained from the `SCALE[1-3]` records. The repeating structure of the crystal lattice was incorporated into the calculations by making the unit cell periodic in the direction of each lattice axis.

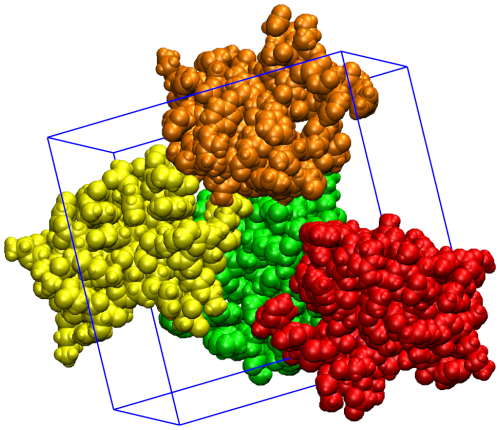
Figure 2.4 illustrates the process of reconstructing the crystallographic unit cell for the scorpion toxin protein 1AHO. The first image (figure 2.4a) shows the asymmetric unit positioned within the unit cell. The 1AHO crystals are orthorhombic and have unit cells that are rect-

angular prisms defined by three lattice basis vectors of unequal length. The unit cell's space group is P $2_12_12_1$ which has four symmetries: the identity symmetry and three 2_1 screw-axis symmetries in the direction of each lattice vector (two-fold rotations combined with a translation by half the length of the vector). Figure 2.4b shows the result of applying the four space group symmetries to the asymmetric unit to generates a unit cell of four proteins. The representation of the unit cell in figure 2.4b is somewhat misleading, suggesting a considerable amount of space between the proteins. The crystal lattice is a periodic structure generated by repeated translations of the unit cell in three dimensions and can be visualised using the analogy of the unit cell as a "building block". The crystal lattice is built up by stacking copies of the unit cell one on top of another. Adjacent copies of the unit cell will pack tightly together resulting in a compact structure with very little empty space between the proteins. Figure 2.4c illustrates the packing of the proteins across the lattice using a periodic representation for the protein molecules. The atoms of the periodic proteins reside entirely within the unit cell and, unlike figure 2.4b, the proteins of figure 2.4c wrap around to emerge from the opposite face when crossing the unit cell's boundaries. Thus, figure 2.4c is a visualisation of the unit cell showing how proteins from neighbouring cells encroach into the space.

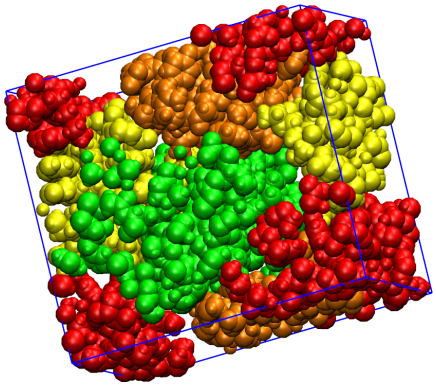
Figure 2.4: Sequence of steps illustrating the reconstruction of the unit cell for protein 1AHO (Smith *et al.* 1997). The edges of the unit cell are plotted as blue lines. The protein defined as the asymmetric unit by the PDB file is coloured red and the three other proteins of the unit cell are coloured orange, yellow and green.



(a) The asymmetric unit.

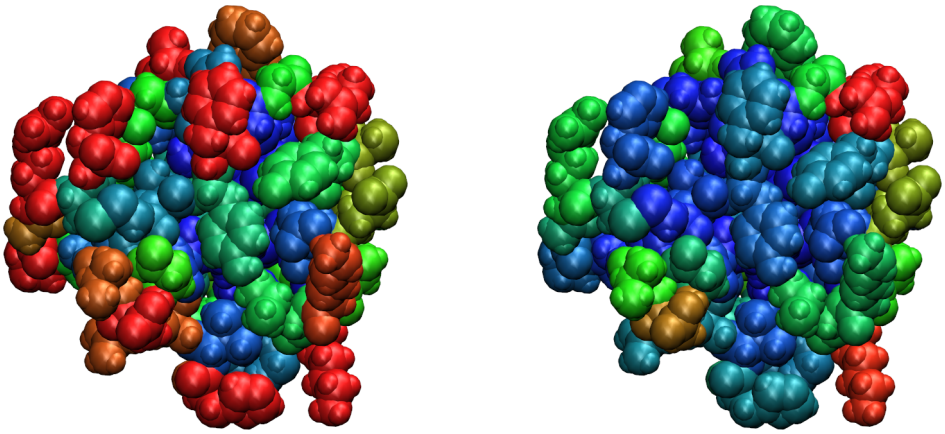


(b) The unit cell.



(c) The unit cell with periodic molecules.

Figure 2.5: Illustrating the effect of crystal contacts on the calculation of SASA for the protein 1AHO (Smith *et al.* 1997). The normalised SASA values are visualised by shading the amino acids using a range of colours from blue (low SASA) to red (high SASA).



(a) The asymmetric unit in isolation.

(b) The proteins of the unit cell

It is apparent from figure 2.4c that protein crystals are tightly packed structures held together by an extensive network of intermolecular contacts. The effect of these protein-protein contacts on structural bioinformatics calculations is illustrated in figure 2.5 where SASA is calculated for the scorpion toxin protein 1AHO. Figure 2.5a is a visualisation of normalised amino acid SASA values calculated for the asymmetric unit in isolation. As would be expected, the amino acids at the surface of the protein are coloured red indicating that a high proportion of their surface area is exposed to the surrounding solvent. In contrast, when the calculations are repeated for the proteins in the unit cell, there is a dramatic reduction in surface exposure (figure 2.5b). Many of the amino acids that appeared to be surface residues in the asymmetric unit have become occluded in the crystal lattice.

2.3 Outline of structural bioinformatics calculations

2.3.1 Solvent accessible surface area

The SASA (Lee and Richards 1971) of an atom was calculated using an implementation of the "rolling sphere" Shrake-Rupley algorithm (Shrake and Rupley 1973). It was necessary to implement a custom version of the algorithm for this thesis because standard software for calculating SASA, such as DSSP (Kabsch and Sander 1983), does not account for the structure of the crystal lattice. Following convention, the solvent probe was a sphere of radius of 1.4Å modelling a water molecule. The surfaces of atoms were described by spherical surface meshes of 128 points at the VdW radii. An approximately uniform distribution of points was achieved by projecting a golden section helix onto the surface of the sphere (Saff and Kuijlaars 1997; Hannay and Nye 2004; Swinbank and Purser 2006). The SASA of each atom was measured twice: once in the whole protein ($SASA^{prot}$) and once in the absence of all other atoms except for those within the same amino acid and the backbone atoms of the two sequentially adjacent residues ($SASA^{amino}$). Normalised SASA values were calculated by dividing $SASA^{prot}$ by $SASA^{amino}$. SASA values for amino acids were calculated by summing SASA values for all the constituent atoms. Normalised amino acid SASA values were calculated by dividing the $SASA^{prot}$ sum by the $SASA^{amino}$ sum.

2.3.2 Secondary structure assignment

Secondary structure was assigned using the Define Secondary Structure of Proteins (DSSP) program (version 2.2.1) (Kabsch and Sander 1983).

2.3.3 Distance from the protein surface

The depth of an atom was defined as the shortest Euclidean distance from the atom to an atom at the surface of the protein. A surface atom was defined as an atom with a SASA or normalised SASA greater than zero ($> 10^{-10}$).

2.3.4 Distance from the centre of mass

The Centre of Mass (COM) was calculated as a mass weighted average of the atomic coordinates. The distance between an atom and the COM was calculated as the Euclidean distance between the coordinates of the COM and the centre of the atom. For this calculation, it was not necessary to reconstruct the unit cell and account for lattice periodicity.

2.3.5 Alpha-carbon coordination number

The alpha-carbon coordination number (Nishikawa and Ooi 1980; Pollastri *et al.* 2002) was chosen as a simple measure of the number of contacts that an amino acid makes with its immediate neighbours. Coordination number was defined as the number of alpha-carbon atoms within a given radius of the alpha-carbon of the amino acid under consideration (excluding itself).

2.4 Algorithms for structural bioinformatics calculations in crystals

SASA, coordination number and the depths of atoms from the protein surface were all calculated using the periodic unit cells reconstructed from the asymmetric unit. All these calculations were dependent on measuring the shortest Euclidean distance between two atoms under the periodicity of the crystal lattice. From this calculation it was then possible to determine how close atoms were to each other in the crystal. Establishing an atom's immediate neighbours was essential for the calculation of SASA, coordination number and the depth to the protein's surface. The algorithms implemented to reconstruct the unit cell, measure distances between atoms and calculate SASA, coordination number and depth from the surface are described below in pseudocode.

Reconstruction of the unit cell

Applies the space group symmetries to all the chains of the asymmetric unit to generate a complete unit cell structure. All the atoms lie within the primary unit cell i.e., the parallelepiped, in the positive octant of the Cartesian coordinate system, with vertices defined by linear combinations of single lattice vectors.

Algorithm 1 Algorithm to reconstruct the unit cell in Cartesian coordinates

Require:

1: $chain$: the asymmetric unit chain
2: $\{\mathbf{M_{xyz}}\}$: the set of Cartesian symmetry matrices
3: $\mathbf{S}$: the scale matrix
4:
5: $unitCell \leftarrow []$
6: **for all** $\mathbf{M_i} \in \{\mathbf{M_{xyz}}\}$ **do**
7: $\quad chain_i^{new} \leftarrow \text{COPY}(chain)$
8: $\quad$ **for all** $atom \in chain_i^{new}$ **do**
9: $\quad\quad \mathbf{p^{xyz}} \leftarrow \text{GETCOORDINATES}(atom)$
10: $\quad\quad \mathbf{p^{abc}} \leftarrow \mathbf{SM_{xyz}p^{xyz}}$
11: $\quad\quad \mathbf{p^{abc}} \leftarrow \mathbf{p^{abc}} - \lfloor \mathbf{p^{abc}} \rfloor$
12: $\quad\quad \mathbf{p^{xyz}} \leftarrow \mathbf{S^{-1}p^{abc}}$
13: $\quad\quad \text{SETCOORDINATES}(atom, \mathbf{p^{xyz}})$
14: $\quad$ **end for**
15: $\quad \text{APPENDCHAIN}(unitCell, chain_i^{new})$
16: **end for**

Notes

5 Initialise a new unit cell structure.

6 Iterate over all the symmetries in the unit cell.

9 Create a new chain based on the original.

10 Apply the symmetry transformation and convert to crystal coordinates.

11 Truncate atom coordinates to lie within the primary unit cell.

12 Convert back to Cartesian coordinates.

15 Add the new chain to the structure.

Calculating the shortest distance in a periodic lattice

Calculation of the Euclidean distance between two points within the unit cell. The algorithm accounts for the periodicity of the crystal lattice to return the shortest possible distance. Essentially, the algorithm calculates the distance between the first point and the closest crystallographically equivalent second point.

Algorithm 2 Function to calculate the shortest distance between two points in the crystal lattice

Require:

1: $\mathbf{p_1^{xyz}}$ and $\mathbf{p_2^{xyz}}$: the two points under consideration
2: $\mathbf{S}$: the scale matrix of the unit cell
3:
4: **function** SHORTESTDISTANCE($\mathbf{p_1}$,$\mathbf{p_2}$,$\mathbf{S}$)
5: $\quad \mathbf{p_1^{abc}} \leftarrow \mathbf{S}\mathbf{p_1^{xyz}}$
6: $\quad \mathbf{p_2^{abc}} \leftarrow \mathbf{S}\mathbf{p_2^{xyz}}$
7: $\quad \mathbf{v_{1,2}^{abc}} \leftarrow \mathbf{p_1^{abc}} - \mathbf{p_2^{abc}}$
8: $\quad \mathbf{v_{1,2}^{abc}} \leftarrow \mathbf{v_{1,2}^{abc}} - \lfloor \mathbf{v_{1,2}^{abc}} \rfloor$
9: $\quad \mathbf{v_{1,2}^{xyz}} \leftarrow \mathbf{S^{-1}}\mathbf{v_{1,2}^{abc}}$
10: $\quad$ **return** $|\mathbf{v_{1,2}^{xyz}}|$
11: **end function**

Notes

5 Convert to crystal coordinates.

7 Calculate the crystal coordinate vector between atoms.

8 Truncate vector components to account for periodicity.

9 Convert vector back to Cartesian coordinates.

10 Return the distance (Euclidean norm).

Find all the neighbouring atoms about a given atom

Locates all the neighbouring atoms no farther than some cutoff distance from a given atom within a unit cell. The algorithm accounts for the periodicity of the crystal lattice and will "wrap around" the boundaries of the unit cell if necessary.

Algorithm 3 Function to find neighbouring atoms within the crystal lattice

Require:

1: a_i : the atom under consideration
2: r_c : the neighbourhood cutoff distance
3: $unitCell$: the reconstructed unit cell
4: $\mathbf{S}$: the scale matrix of the unit cell
5:
6: **function** FINDNEIGHBOURS(a_i,r_c,$unitCell$,$\mathbf{S}$)
7: $neighbours \leftarrow []$
8: $\mathbf{p_i} \leftarrow$ GETCOORDINATES(a_i)
9: **for all** $a_j \in unitCell : a_j \neq a_i$ **do**
10: $\mathbf{p_j} \leftarrow$ GETCOORDINATES(a_j)
11: $d \leftarrow$ SHORTESTDISTANCE($\mathbf{p_i}$,$\mathbf{p_j}$,$\mathbf{S}$)
12: **if** $d < r_c$ **then**
13: APPENDATOM($neighbours$,a_j)
14: **end if**
15: **end for**
16: **return** $neighbours$
17: **end function**

Notes

7 Initialise an empty list.

12 Add the neighbouring atom if within the cutoff distance.

16 Return all the neighbours of atom a_i.

Estimate the solvent accessible surface area of an atom with a unit cell

Applies the Shrake-Rupley algorithm (Shrake and Rupley 1973) to calculate SASA in the context of a periodic lattice structure. The algorithm accounts for lattice periodicity when determining how much of an atom's solvent accessible surface area is occluded by its immediate neighbours. The surface of an atom is defined as a spherical mesh of points approximately equidistant from one another (see the method section for the details of how the mesh was constructed).

Algorithm 4 Function to calculate the SASA of an atom within some group of atoms

Require:

a_i : the atom under consideration
atomGroup : the group of atoms (the unit cell or the amino acid and backbone atoms)
S : the scale matrix of the unit cell

function CALCULATESASA(a_i,*atomGroup*,**S**)
    $r_{probe} \leftarrow 1.4$
    $N \leftarrow 128$
    $r_{v,i} \leftarrow$ VANDERWALLSRADIUS(a_i)
    $\mathbf{p_i} \leftarrow$ GETCOORDINATES(a_i)
    $surfacePoints_i \leftarrow$ CONSTRUCTSURFACEPOINTS(p_i, $r_{v,i} + r_{probe}$, N)
    $neighbours_i \leftarrow []$
    for all $a_j \in G : a_j \neq a_i$ **do**
        $r_{v,j} \leftarrow$ VANDERWALLSRADIUS(a_j)
        $\mathbf{p_j} \leftarrow$ GETCOORDINATES(a_j)
        $d \leftarrow$ SHORTESTCARTESIANDISTANCE($\mathbf{p_i}$,$\mathbf{p_j}$,**S**)
        if $d < r_{v,i} + r_{v,j} + (2 \times r_{solv})$ **then**
            APPENDATOM(*neighbours*,a_j)
        end if
    end for
    for all $\mathbf{s} \in surfacePoints_i$ **do**
        for all $a_j \in neighbours$ **do**
            $\mathbf{x_j} \leftarrow$ GETCOORDINATES(a_j)
            $r_{v,j} \leftarrow$ VANDERWALLSRADIUS(a_j)
            $d \leftarrow$ SHORTESTCARTESIANDISTANCE(**s**,$\mathbf{p_j}$,**S**)
            if $d < r_{v,j} + r_{probe}$ **then**
                REMOVEPOINT(**s**,*surfacePoints*)
            end if
        end for
    end for
    $n \leftarrow$ NUMBEROFPOINTS(*surfacePoint*)
    $sasa \leftarrow 4\pi(r_{v,i} + r_{probe})^2 \times \frac{n}{N}$
    return *sasa*
end function

Notes

10 Create the spherical mesh of N surface points centred at point p_i with a radius of $r_{v,i} + r_{probe}$ (solvation radius).

11 Initialise an empty list to store neighbouring atoms.

12-19 Find all neighbouring atoms that limit the atom's exposure to the solvent.

17 Add the neighbour to the list if the solvent cannot fit between atoms.

20-29 Measure the extent of solvent exposure.

25 Eliminate the surface point if occluded by a neighbouring atom.

31 Calculate the proportion of exposed surface area based on the number of surface points remaining.

Estimate the normalised solvent accessible surface area of an amino acid

Calculates the normalised SASA of an amino acid within the unit cell. The SASA is simply the sum of the SASA values calculated for each atom of the amino acid. The SASA calculation is undertaken twice. The first calculation is for the amino acid surrounded by all other structure within the unit cell and accounts for the periodicity of the crystal lattice. The second calculation is of the amino acid in isolation except for the backbone atoms of its flanking residues in the protein chain. Dividing the SASA of the amino acid in the unit cell by the SASA of the amino acid in isolation gives the normalised SASA value.

Algorithm 5 Function to calculate the normalised SASA of an amino acid

Require:

1: *aminoAcid* : the amino acid under consideration
2: *unitCell* : the reconstructed unit cell
3: *segment* : the isolated amino acid plus the backbone atoms of flanking residues
4: **S** : the scale matrix of the unit cell
5:
6: **function** AMINOACIDNORMSASA(*aminoAcid*,*unitCell*,*segment*,**S**)
7: $sasaProtein \leftarrow 0$
8: $sasaIsolated \leftarrow 0$
9: **for all** $atom \in aminoAcid$ **do**
10: $sasaProtein \leftarrow sasaProtein +$ CALCULATESASA(*atom*, *unitCell*, **S**)
11:
12: $sasaIsolated \leftarrow sasaIsolated +$ CALCULATESASA(*atom*, *segment*, **S**)
13:
14: **end for**
15: $aminoAcidNormSasa \leftarrow \dfrac{sasaProtein}{sasaIsolated}$
16:
17: **return** *normSasa*
18: **end function**

Notes

7-8 Initialise total SASA values as zero.

9 Calculate SASA in the environment of the protein and add to the running total.

10 Calculate SASA for isolated amino acid (including any adjacent backbone atoms) and add to the running total.

12 Normalise SASA as the ratio of the two surface areas calculated.

Finding the atoms at the surface of a protein within a unit cell

Simple method to find all the surface atoms of a protein. The SASA values of each atom within the protein are calculated both in the periodic unit cell and in isolated parent amino acids. The ratio of the atom's SASA in the unit cell to the value in the isolated amino acid is calculated. All the atoms with a ratio greater than zero are considered to be exposed at the surface. Atoms whose raw SASA values are effectively zero are discounted immediately to avoid problems with division by zero or division by very small numbers.

Algorithm 6 Function to locate the surface atoms of a protein

Require:

1: *protein* : the protein under consideration
2: *unitCell* : the reconstructed unit cell
3: **S** : the scale matrix of the unit cell
4:
5: **function** FINDSURFACEATOMS(*protein*,*unitCell*,**S**)
6: $surfaceAtoms \leftarrow []$
7: **for all** *aminoAcid* ∈ *protein* **do**
8: *segment* ← CREATEISOLATEDSTRUCTURE(*aminoAcid*, *protein*)
9: **for all** *atom* ∈ *aminoAcid* **do**
10: *sasaProtein* ← CALCULATESASA(*atom*, *unitCell*, **S**)
11: *sasaIsolated* ← CALCULATESASA(*atom*, *segment*, **S**)
12: *ratio* ← 0
13: **if** *sasaProtein* and *sasaIsolated* are not zero **then**
14:
15: $ratio \leftarrow \frac{sasaProtein}{sasaIsolated}$
16:
17: **end if**
18: **if** $ratio > 0$ **then**
19: APPEND(*atom*, $surfaceAtoms$)
20: **end if**
21: **end for**
22: **end for**
23: **return** $surfaceAtoms$
24: **end function**

Notes

6 Initialise an empty list of surface atoms.

8 Construct isolated amino acid segment.

10 Calculate the SASA of the atom in the protein

11 Calculate the SASA of the atom in the isolated amino acid segment.

12 Determine whether the atom meets the criteria of a surface atom.

13 Add atom to the list of surface atoms if there is *any* degree of solvent exposure.

Calculation of the depth of an atom from the protein's surface

Builds upon the algorithm used to identify surface atoms to calculate surface depth. All the surface atoms of a protein are calculated using the algorithm described previously. The distance of atom from the surface is simply to shortest distance from that atom to any of the surface atoms.

Algorithm 7 Function to calculate the surface depth of an atom

Require:

1: $atom$: the atom under consideration
2: $surfaceAtoms$: the surface atoms of the protein
3:
4: **function** SURFACEDEPTH($atom$, $surfaceAtoms$)
5: **if** $atom \in surfaceAtoms$ **then**
6: **return** 0
7: **else**
8: $\mathbf{p_{atom}} \leftarrow$ GETCOORDINATES(atom)
9: $\forall \mathbf{p_{surf}} \in \{$ GETCOORDINATES(a) : $a \in surfaceAtoms\}$
10: **return** MIN(SHORTESTDISTANCE($\mathbf{p_{atom}}$,$\mathbf{p_{surf}}$,$\mathbf{S}$))
11: **end if**
12: **end function**

Notes

5 Define atoms at the surface as having zero "depth".

9 Find the minimum distance for the atom to any surface atom.

Calculation of amino acid coordination number within a unit cell

Builds upon the algorithm used to locate the immediate neighbours of a given atom within the unit cell. The unit cell is reconstructed using only the alpha-carbon atoms of the protein chains. All the neighbours within a cutoff distance of a given alpha-carbon are calculated using the algorithm described previously. The number of neighbouring alpha-carbons is the coordination number.

Algorithm 8 Function to calculate the α-carbon coordination number of an amino acid

Require:

1: $aminoAcid_i$: the amino acid under consideration
2: $unitCell$: the reconstructed unit cell
3: $\mathbf{S}$: the scale matrix of the unit cell
4: r_c : the cutoff radius
5:
6: **function** COUNTCOORDNUMBER($aminoAcid_i$, $unitCell$, $\mathbf{S}$, r_c)
7: $C_{\alpha,i} \leftarrow$ GETALPHACARBON($aminoAcid_i$)
8: $\mathbf{p_i} \leftarrow$ GETCOORDINATES($C_{\alpha,i}$)
9: $coordNumber \leftarrow 0$
10: **for all** $aminoAcid_j \in unitCell : aminoAcid_i \neq aminoAcid_j$ **do**
11: $C_{\alpha,j} \leftarrow$ GETALPHACARBON($aminoAcid_j$)
12: $\mathbf{p_j} \leftarrow$ GETCOORDINATES($C_{\alpha,j}$)
13: **if** SHORTESTDISTANCE($\mathbf{p_i}$,$\mathbf{p_j}$,$\mathbf{S}$) $< r_c$ **then**
14:
15: $coordNumber \leftarrow coordNumber + 1$
16: **end if**
17: **end for**
18: **return** $coordNumber$
19: **end function**

Notes

9 Initialise the coordination number count to zero.

13 Check if the atom is within the required local neighbourhood.

14 Increment the coordination number count by one.

Chapter 3

Evaluating isotropic B-factors as indicators of a protein's conformational dynamics

3.1 Introduction

The introduction outlined the reasoning behind the assumption that crystallographic B-factors are a reflection of the conformational dynamics of a protein within a crystal. In addition, an equally valid counterargument was discussed, proposing that B-factors may bear little or no relation to the underlying dynamics of the protein. The precision to which a protein structure can be determined by X-ray crystallography is affected by many variables, and conformational dynamics may not necessarily be the dominant factor. Crystal defects, experimental error, static structural disorder and global rigid body movements of a protein all offer plausible alternative explanations for B-factors.

Despite all the known limitations of B-factors, researchers continue to mine the B-factor data of the PDB in order to establish relationships between a protein's structure and its conformational dynamics. Analysis of B-factor distributions have been used to derive flexibility indices for individual amino acids (Karplus and Schulz 1985; Smith, Radivojac *et al.* 2003) and to relate conformational stability to side chain motility (Carugo and Argos 1997). B-factors have also been applied to the problems of predicting enzyme active sites (Yuan *et al.* 2003) and potential protein-protein interaction sites (Liu *et al.* 2010). The value molecular bioinformaticians still place on B-factor data is apparent from the fact that methods to predict protein B-factor profiles are continually being developed (Yuan *et al.* 2005; Schlessinger *et al.* 2006; Sonavane *et al.* 2013). In light of this considerable body of work, there is a clear need to re-evaluate the usefulness of B-factors as indicators of a protein's

conformational dynamics.

Two areas have been identified where previous research could usefully be extended. Firstly, scarcity of structural data meant that older studies had to make use of low resolution or low quality structures. Secondly, technological limitations meant that analyses only considered proteins in isolation rather than as elements of a crystal lattice. Given the increased number of high quality structures in the PDB and the greater computational power now available, it is now time to re-evaluate rigorously the relationship between crystallographic B-factors and protein conformational dynamics.

3.2 Aim

The aim is of this study is to make an up-to-date assessment of the value of isotropic B-factors as indicators of conformational variability in protein crystals. This study will begin by focusing on isotropic B-factor data rather than anisotropic atomic displacement parameters. The rationale being that isotropic B-factors are more frequently used as a surrogate measure of flexibility in bioinformatics research.

The extent to which isotropic B-factors reflect protein dynamics will be assessed using high resolution crystal structures. B-factor data will be analysed in relation to a set of protein structural properties that are expected to correlate with conformational variability. Furthermore, since the analysis is limited to the data deposited in the PDB, measurements will be restricted to static structural features of a protein that can be derived from the data present in a PDB file.

The static structural properties chosen for analysis are:

- Amino acid type
- Secondary structure
- Surface exposure measured as a normalised SASA
- Surface depth
- Distance from the protein's COM
- Local atom packing density

In addition, all protein-protein contacts in the crystal will be considered where this may affect the calculations.

3.3 Hypothesis

Using only high resolution crystallographic data and fully accounting for the crystal structure should make it possible to observe clear relationships between isotropic B-factors and the protein's structure. Specifically, an atom's B-factor value should be consistent with its expected conformational freedom given the structure of the crystal. If, however, B-factors are dominated by contributions from other effects, then it will be more difficult to discern any relationships.

3.4 Results and discussion

3.4.1 Creating the protein data set

A key objective in this project was to derive a suitable data set of high resolution proteins for analysis. Strict selection criteria were applied to ensure the quality of the crystallographic data.

The criteria applied to select proteins for analysis were:

- X-ray structures determined at near-atomic resolution ($\leq$1.5 Å).
- Structures refined purely isotropically.
- Single chain proteins.
- Chains of at least 50 amino acids.
- Low sequence homology between proteins ($< 30\%$).
- Cytosolic or extracellular proteins.
- Structures to be complete or near-complete with only a small number of unresolved amino acids.
- High quality crystal structures with low R ("reliability") indices.
- Proteins should not be bound to large cofactors or co-crystallised with large molecules.

The rationale behind these criteria is that, by limiting the analysis to a diverse collection of high quality structures, any relationships between B-factors and the structural properties of proteins will be more apparent than if the data set encompassed every isotropically refined crystal structure deposited in the PDB. Setting a near-atomic upper limit for the resolution (1.5 Å) made it more likely that any uncertainty in locating atoms within a structure would be due to the protein's conformational dynamics. Hence, the assertion that B-factors are reflection of conformational variability is far easier to justify. Furthermore, B-factors are less

likely to be catch-all error terms in structures that have been refined with low R ("reliability") indices, since a low R index is indicative of a structure in close agreement with the crystallographic data.

The analysis was restricted to single chain proteins to avoid any complications that might arise from the dynamics of multi-subunit protein complexes. The protein-protein interactions between the chains in a multi-subunit complex are different to those that hold the crystal lattice together (Carugo and Argos 1997). Consequently, the B-factors of atoms in crystals of multi-subunit complexes may be unlike those of comparable atoms in crystals of single chain proteins. Peptides (defined as proteins shorter than 50 amino acids) were similarly excluded since the lack of extended secondary and tertiary structure might result in unusual dynamics.

Low sequence homology between proteins guaranteed structural diversity and minimised bias from the inclusion of proteins whose crystal structures have been determined multiple times. For example, approximately 50 structures at near-atomic resolution have been deposited in the PDB for hen egg-white lysozyme.

Membrane proteins were excluded from the data set because their inclusion might have had a distorting effect on the analysis. The structures of membrane proteins are distinctly different to those of cytosolic or extracellular proteins. Membrane proteins span lipid bilayers and, therefore, have surfaces enriched with hydrophobic amino acids. In contrast, the hydrophobic amino acids of a cytosolic or extracellular protein are usually buried within its interior. In addition, membrane proteins must be co-crystallised with detergent molecules to provide stability in the absence of the cell membrane. The interaction between a membrane protein and detergent may perturb the dynamics, and result in B-factors that are not directly comparable with other proteins. By the same reasoning, any cytosolic or extracellular proteins co-crystallised with large molecules, either present as bound cofactors or as components of the crystallisation medium, were excluded.

The measurements of structural properties such as surface area and atom packing density requires complete protein structures where the location of every atom is known. However, this presents a problem for structures determined by X-ray crystallography since atoms with low electron density, hydrogen in particular, are rarely detected. Furthermore, highly mobile amino acids within unstructured regions of a protein may be completely missing from the structure or only partially resolved. Modelling software can deduce the most likely positions of any missing atoms within a crystal structure, but the structures generated can only be considered to be reliable if the majority of the protein's conformation is already known. For this reason, near complete crystal structures were favoured over those with long stretches of unresolved residues.

Full details of how the protein data set was obtained is given in the methods section of this chapter. The properties of the final dataset of 114 structures are summarised in table 3.1.

Table 3.1: Summary of the protein structures resolved isotropically.

Feature	Number of proteins	% of data set
Chain length [1]		
$51 - 99$	24	21.1
$100 - 299$	70	61.4
$300 - 532$	20	17.5
Resolution		
< 1.0Å	3	2.6
$1.0 - 1.5$Å	72	63.2
1.5Å	39	34.2
Space Group		
P 2_1 2_1 2_1	38	33.3
P 1 2_1 1	15	13.2
C 1 2 1	13	11.4
22 other space groups [2]	48	42.1
Structural Classification [3]		
all-α ($> 60\%$ α and $< 5\%$ β)	10	8.77
mostly α-helix ($> 60\%$ α and $> 5\%$ β)	1	0.88
all-β ($> 50\%$ β and $< 5\%$ α)	4	3.51
mostly β-structure ($> 50\%$ β and $> 5\%$ α)	0	0.00
$\alpha\beta$ proteins ($15 - 55\%$ α and $10 - 45\%$ β)	66	57.89
others	33	28.95
Alternate conformations [4]		
0%	47	41.2
$0 - 10\%$	55	48.3
$10 - 20\%$	12	10.5
$\geq 20\%$	0	0.0

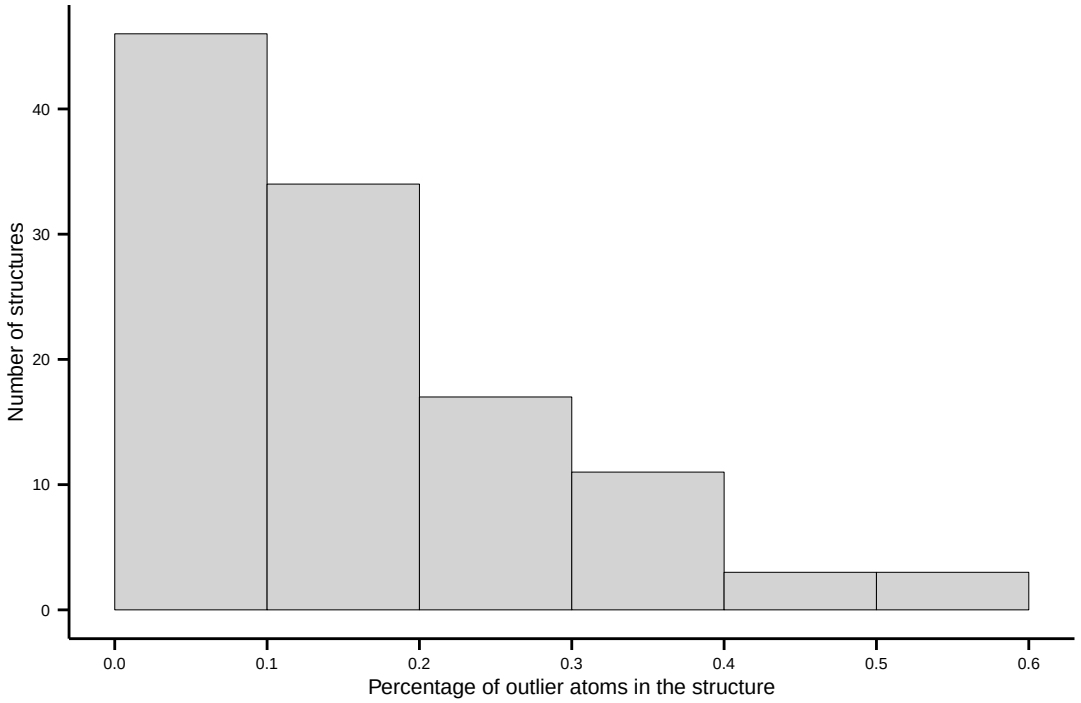
1 Median length 166.5. The minimum and maximum are 51 and 721 respectively.
2 Eleven space groups are represented by a single structure.
3 Using the domain structural classification of Michie *et al.* (1996).
4 Measured as the proportion of amino acids resolved with alternate conformations. The highest proportion is 19.2%.

3.4.2 Assessing the quality of the data set

The quality of the B-factor data was evaluated using a local averaging check based on the assumption that atoms in close proximity should have similar B-factors (Hendrickson 1985). The methodology is similar to the *ISOR* and *SIMU* restraints of the `SHELX` crystallographic refinement software (Sheldrick and Schneider 1997) where the variance in isotropic atomic displacements of spatially close atoms cannot exceed a certain threshold. An atom was marked as an "outlier" if its B-factor exceeded three standard deviations from the mean of its neighbours within a 5 Å radius. The number of outlier atoms was counted for each protein and expressed as a percentage of the total. A high proportion of outlier atoms was reasoned to be indicative of a structure that fitted poorly to the crystallographic data. The presence of such proteins in the data set would, therefore, obscure any relationships between B-factors and the protein structural properties under investigation.

The distribution of the proportion of outlier atoms over the whole data set is plotted in figure 3.1. The distribution clearly shows that the proportion of atoms with "atypical" B-factors is very low. Less than 1% of the atoms in each protein of the data set are outliers, and the proportion is even lower (less than 0.5%) for the majority of these proteins. Consequently, there is no evidence for any significant systematic anomalies in the B-factor data of the structures in the data set.

Figure 3.1: Distribution of B-factor outliers across the protein data set. The horizontal axis is the percentage of atoms in a crystal structure that have "atypical" B-factors and is a continuous scale. The data is binned in intervals of 0.1%. For example, the first bar shows that, for the majority of structures in the data set, the percentage of atoms in the structure with "atypical" B-factors is between 0 and 0.1%.

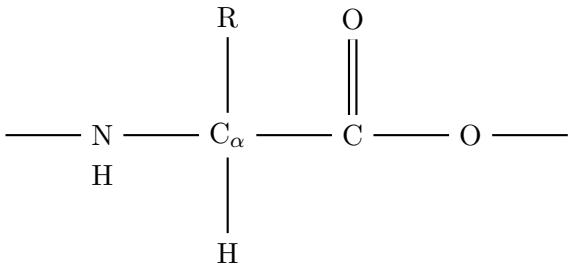


3.4.3 Distribution of alpha-carbon B-factors

The analysis of B-factors was focused on the alpha-carbons of proteins. The choice of alpha-carbons was guided by the assumption that, of all the atom types present in every amino acid, the movement of the alpha-carbon is most likely to be influenced by the local dynamics of the protein. The alpha-hydrogen was not considered suitable because it is not usually resolved by X-ray crystallography and the carbonyl and amine groups of the backbone were rejected because their movements are constrained by the chemistry of the peptide bond. Therefore, where defined, the B-factors of the alpha-carbons can be used as a standard, comparable measure of the conformational variability along the length of the protein backbone. Figure 3.2 is a schematic diagram of a generalised amino acid within a protein, illustrating the position of the alpha-carbon in relation to the other atoms of the backbone.

The isotropic B-factors of the atoms within each structure were normalised using the median-mad method. This technique was preferred over traditional mean-standard deviation ("z"-normalisation) because the use of robust statistics for central tenancy (median) and spread (Median Absolute Deviation (MAD)) make it less sensitive to distortion from atypical data (Wilcox 2010).

Figure 3.2: Generalised structure of an amino acid within a protein. The alpha-carbon (Cα) is positioned centrally with the amine and carbonyl groups of the peptide bonds at either side. The amino acid side chain is represented by the functional group "R".



The distribution of normalised alpha-carbon B-factors (figure 3.3) is highly positively skewed (skewness measure 3.021). It could be argued that the source of the skew in the distribution is a consequence of the B-factor data being "incomplete" due to there being alpha-carbons unresolved by crystallography. The data set is representative of the true distribution because there are 22,810 alpha-carbons in the data set of which only 123 (approximately 0.5%) were unresolved and had coordinates, but not B-factors, estimated by modelling software. Thus, the data set is almost complete with respect to alpha-carbon B-factors. It should be noted, however, that the unresolved atoms are likely to be located in the most mobile regions of a protein and would, therefore, be expected to have B-factors that lie to the far right of the distribution. The absence of these highly mobile atoms will contribute to the positive skewness of the distribution but, given the small numbers involved, it is unlikely to be a

Table 3.2: Parameters for the Gaussian mixture model. Values calculated after convergence of the EM algorithm ($\epsilon = 10^{-3}$).

$\boldsymbol{i}$ (component)	$\boldsymbol{\lambda_i}$ (proportion)	$\boldsymbol{\mu_i}$ (mean)	$\boldsymbol{\sigma_i}$ (standard deviation)
1	0.810	-0.132	1.115
2	0.190	3.598	3.460

significant factor.

The broad spread of atoms with high B-factors is an interesting feature of the data set and has been commented on previously (Parthasarathy and Murthy 1997; Smith, Radivojac *et al.* 2003). It is tempting to explain the shape of this distribution in terms of a two population model consisting of low B-factor interior atoms and a set of more flexible high B-factor atoms at the surface (Parthasarathy and Murthy 1997). This was investigated by fitting a two component Gaussian mixture model to the data. In this model, the probability density function for the alpha-carbon B-factors is assumed to be a weighted sum of two Gaussian distributions (equation 3.1).

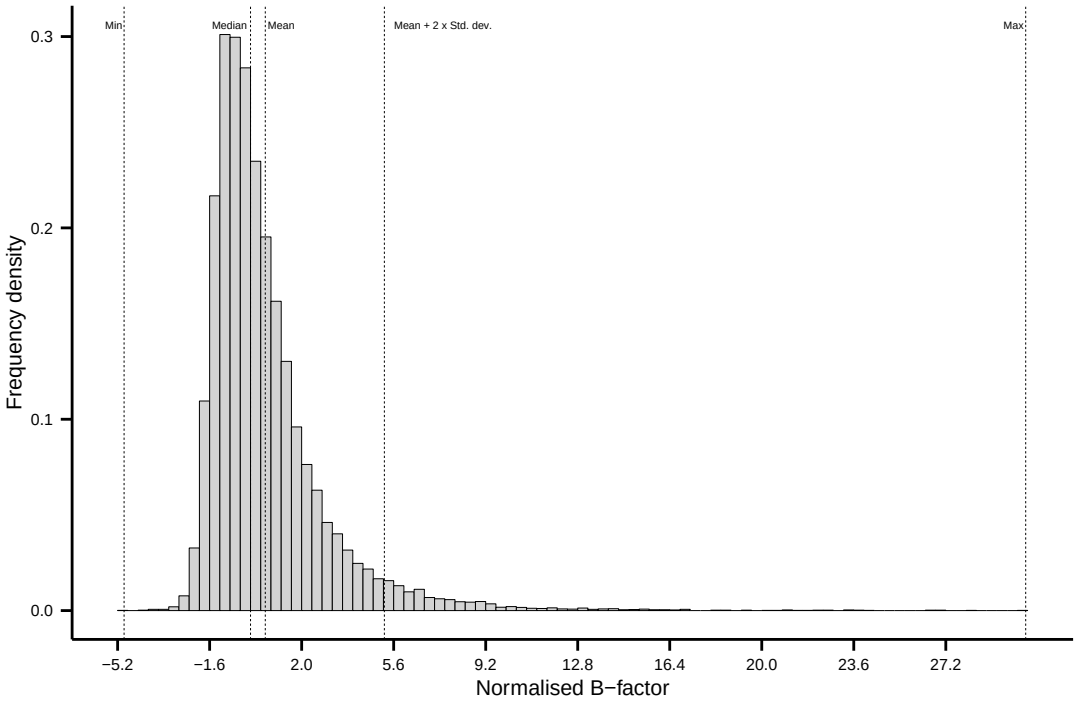
$$p(x) = \lambda_1 n(x; \mu_1, \sigma_1) + \lambda_2 n(x; \mu_2, \sigma_2) \tag{3.1}$$

In equation 3.1, $n(x; \mu, \sigma)$ is the density function for a Gaussian with mean and standard deviation of μ and σ respectively. The weightings of the two Gaussians, λ_1 and λ_2, correspond to the proportions of atoms in the two populations. The parameters of the mixture model were calculated using the EM algorithm as implemented by the `mixtools` R library (Benaglia *et al.* 2009). A plot comparing of the empirical distribution to the fitted mixture model is shown in figure 3.4 and the values of the model's parameters are recorded in table 3.2.

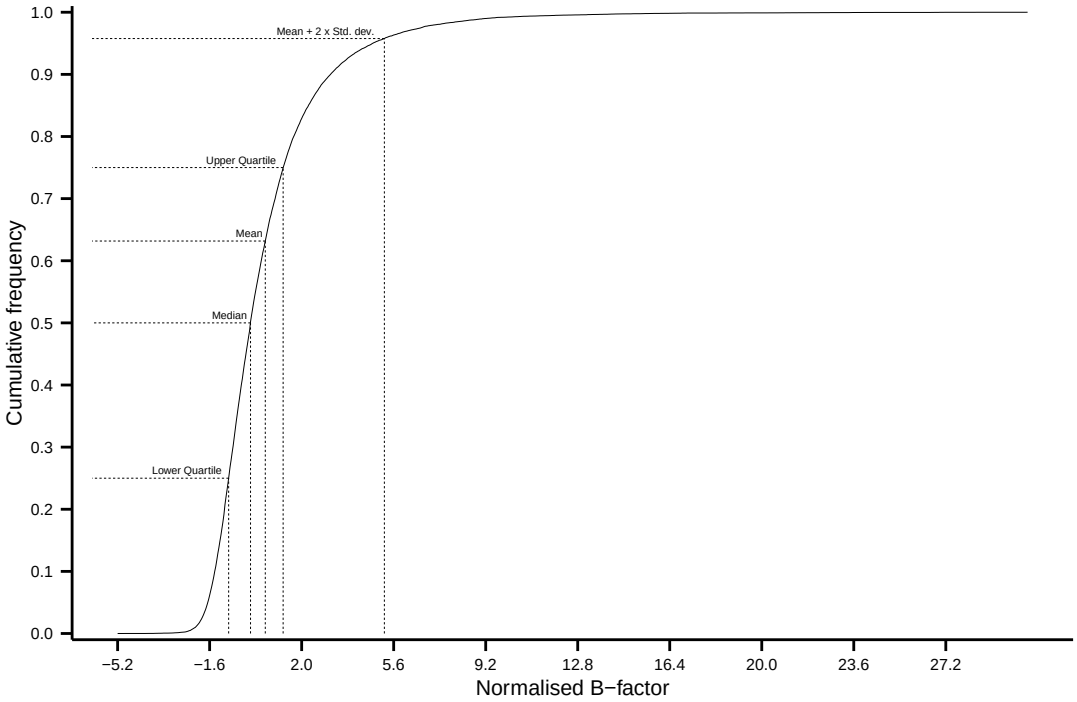
Visual inspection of the graphs in figure 3.4 suggest that a two component Gaussian mixture model is an approximate description of the B-factor data. This was confirmed quantitatively by a Kolmogorov-Smirnov test that gave a test statistic of 0.0374 (zero and one measuring maximal and minimal agreement respectively). The closeness of the fit is best visualised by a quantile-quantile plot of the empirical cumulative density function against that of the mixture model (figure 3.5) where perfect agreement would correspond to a straight line. Although figure 3.5 approximates a straight line, the plot deviates from this ideal over the whole data range. Increasing the number of components of the mixture model improved the fit, but a Gaussian mixture model could be made to fit the data to any level of precision given a sufficient number of components. More complex models, however, would be far more difficult to interpret in terms of protein structure.

Even though a Gaussian mixture was only an approximate model for the B-factor data, it was interesting to examine whether the two Gaussian components corresponded to interior and exterior atoms of the proteins. A coarse classification was applied where alpha-carbon

Figure 3.3: Distribution of alpha-carbon B-factors for the maximum occupancy structures.

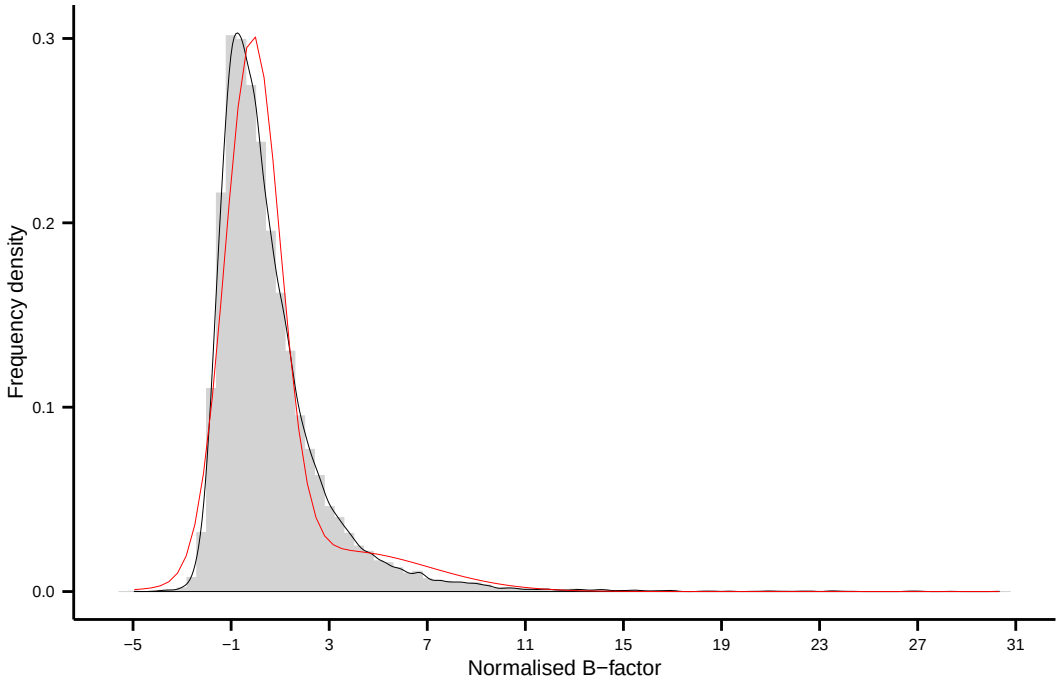


(a) Histogram of alpha-carbon B-factors



(b) Cumulative frequency distribution for alpha-carbon B-factors

Figure 3.4: Comparing the empirical distributions for the B-factors to a Gaussian mixture model. The histogram is shaded in grey and the empirical probability distribution is plotted as a black line. The mixture model is superimposed in red.



atoms were defined as interior atoms if their amino acids had a normalised SASA less than or equal to 0.05. An amino acid with a normalised SASA greater than 0.05 was defined as being at the surface. Density plots of the B-factor distributions of these two subsets of atoms are presented in figure 3.6 and their summary statistics in table 3.3. The results confirm that the interior atoms generally have lower B-factors than those at the surface of the protein. Nonetheless, there is a considerable degree of overlap between the ranges of B-factors exhibited by these two groups of atoms. Comparing the parameters of tables 3.2 and 3.3 reveals that the components of the mixture model do not correspond with the definitions of interior and surface atoms. Furthermore, alpha-carbons within the protein interior make up a smaller proportion of all atoms and exhibit a greater spread of B-factor values compared to the prediction of the mixture model. It could be argued that a better level of agreement might be possible by adjusting the definitions for interior and surface atoms. Nevertheless, this would be futile because, irrespective of the choice of the cutoff value used to classify the atoms, the B-factor distributions for both interior and surface atoms will remain highly skewed and deviate from normality. The factors that determine an atom's B-factor appear to be far more subtle than a simple binary classification based on the atom's location.

Figure 3.5: Quantile-quantile plot comparing the cumulative density function for the Gaussian mixture model against the empirical cumulative density function for the B-factors. Perfect agreement between the cumulative density functions is represented by the dashed line.

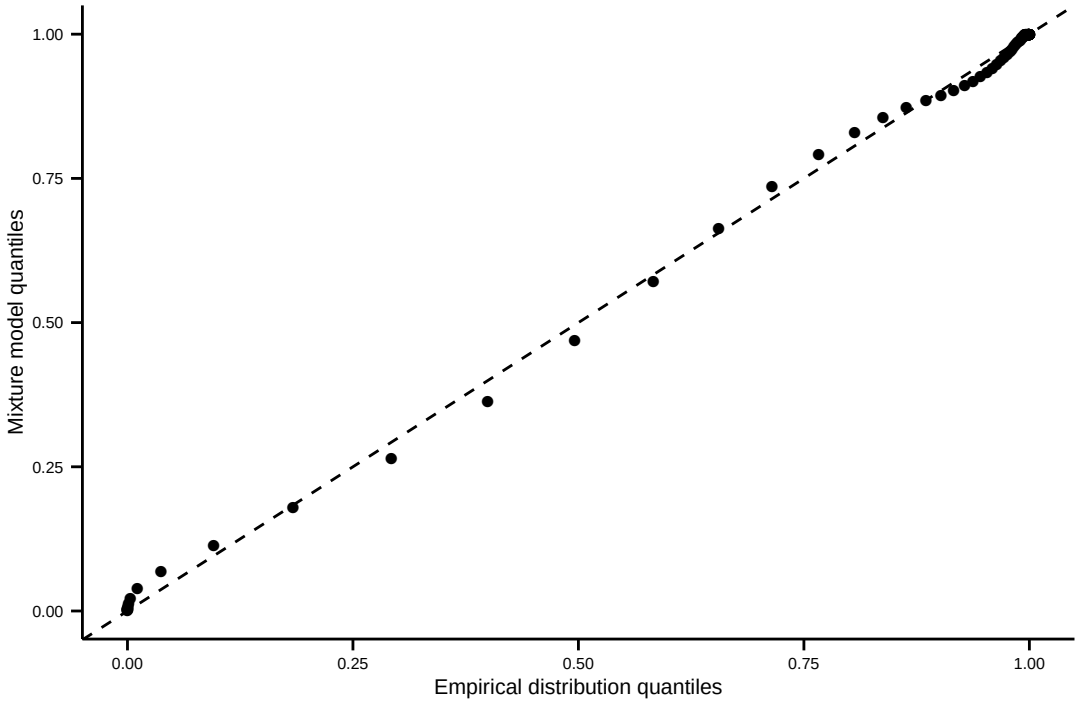


Figure 3.6: Comparing the distributions for the B-factors of interior and surface atoms.

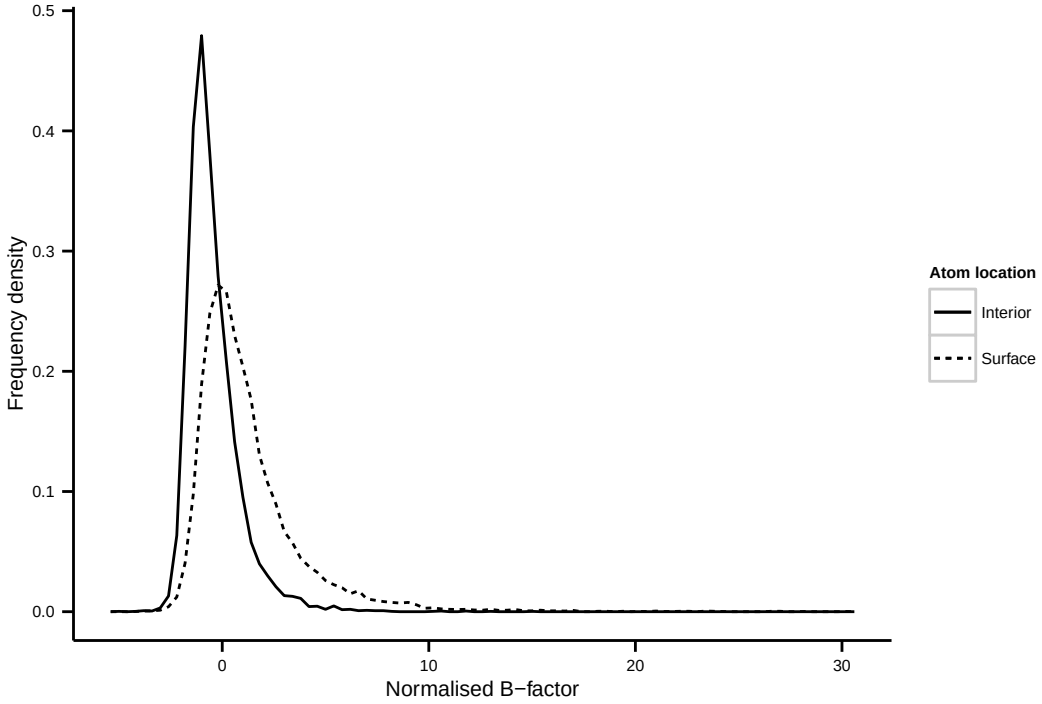


Table 3.3: Summary statistics for surface and interior alpha-carbon B-factors

Location	**Number (proportion)**		**Mean**	**Std. Dev.**	**Min.**	**Max.**
Interior	8790	0.39	-0.458	1.286	-4.947	14.927
Surface	13897	0.61	1.132	2.590	-4.011	30.326

3.4.4 Relating alpha-carbon B-factors to protein structure

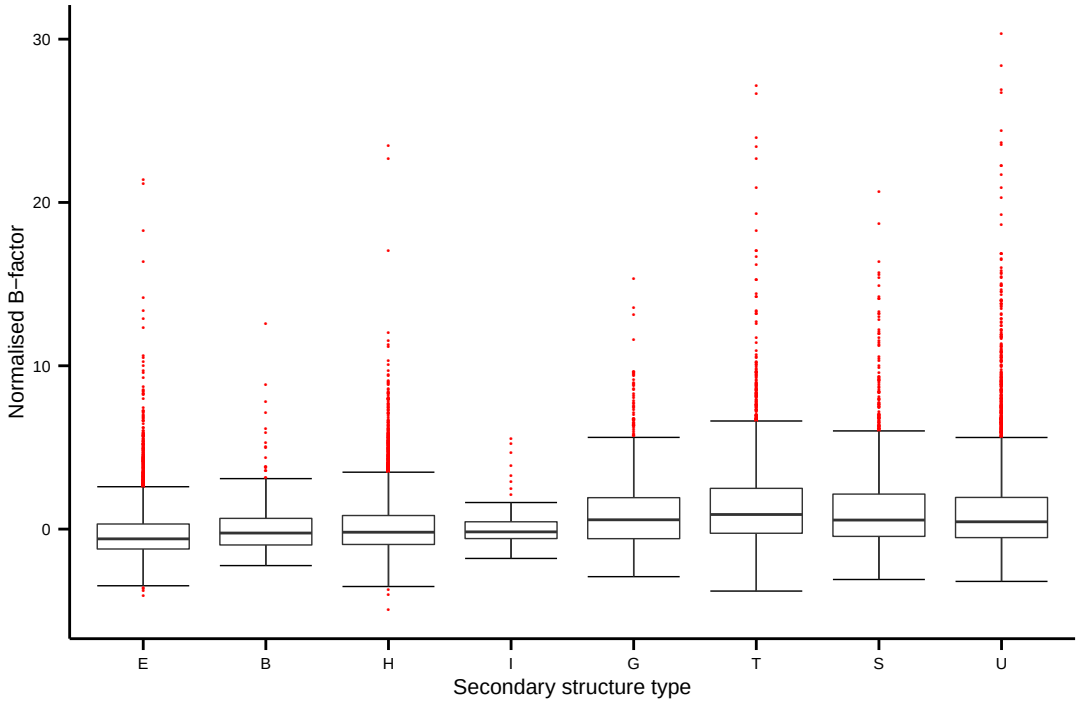
The initial exploration of the B-factor data set confirms that, at a very coarse level, B-factors are a reflection of the conformational variability of an atom. B-factors of alpha-carbons within the interior of a protein are typically lower than those at the surface which is consistent with amino acids having less conformation freedom at the protein's core. Nevertheless, there are evidently other factors that influence B-factor values as seen from the considerable overlap between the B-factors of surface and interior atoms.

In order to investigate whether B-factors are a reflection of conformational variability, the alpha-carbon data was subdivided according to various structural properties. The features chosen for analysis were: secondary structure classification, amino acid type, solvent exposure, distance from the surface, distance from the protein's COM and amino acid coordination number. All of these properties are widely accepted as correlates of conformational variability and would, therefore, be expected to correlate with B-factor values if the hypothesis is correct. Interestingly, irrespective of how the data set was subdivided in the subsequent analyses, similarly shaped positively skewed B-factor distributions were observed in all subsets. It was decided to visualise the data as boxplots in order to preserve as much information as possible about the shapes of the distributions. Boxplots were generated using the `ggplot2` package in GNU R (Wickham 2009). The horizontal bars that define the "boxes" of the boxplots are the 25%, 50% (median) and 75% quartiles of the data. The "whiskers" of the boxplots extend to the nearest data points within a distance of 1.5 times the interquartile range. All data points that lie beyond the range of the "whiskers" are classified as "outliers".

The majority of the boxplots presented in this thesis are plotted without the outlier data. Although comprising a small fraction of the total population, the outlying B-factor values were at the extreme ends of the data range. The inclusion of outlier data requires a very coarse scale when plotting the boxplots and can, therefore, obscure small but significant differences between boxplots. Figure 3.7 plots the boxplots with outliers for alpha-carbon B-factors grouped according to secondary structure classification. This example clearly shows that including the outliers makes it very difficult to distinguish between the B-factor distributions.

A striking feature of all the boxplots generated is the span of the "whiskers" and the wide range covered by the outlier data. This observation suggests that protein B-factor data is inherently highly variable. Quantitative analysis using summary statistics or fitting linear models was deemed inappropriate due to the broad scattering of the data. This thesis can, therefore, only report qualitative trends between B-factors and the structural properties investigated. Deriving reliable predictive quantitative relationships proved to be impossible.

Figure 3.7: Normalised alpha-carbon B-factors grouped according to secondary structure. Outliers are plotted as red points. Secondary structure labels are the DSSP classifications: E : β (extended); B : β (bridge); H : α-helix; I : π-helix; G : 3-10 helix; T : turn; S : bend; and U : unclassified ("coil").



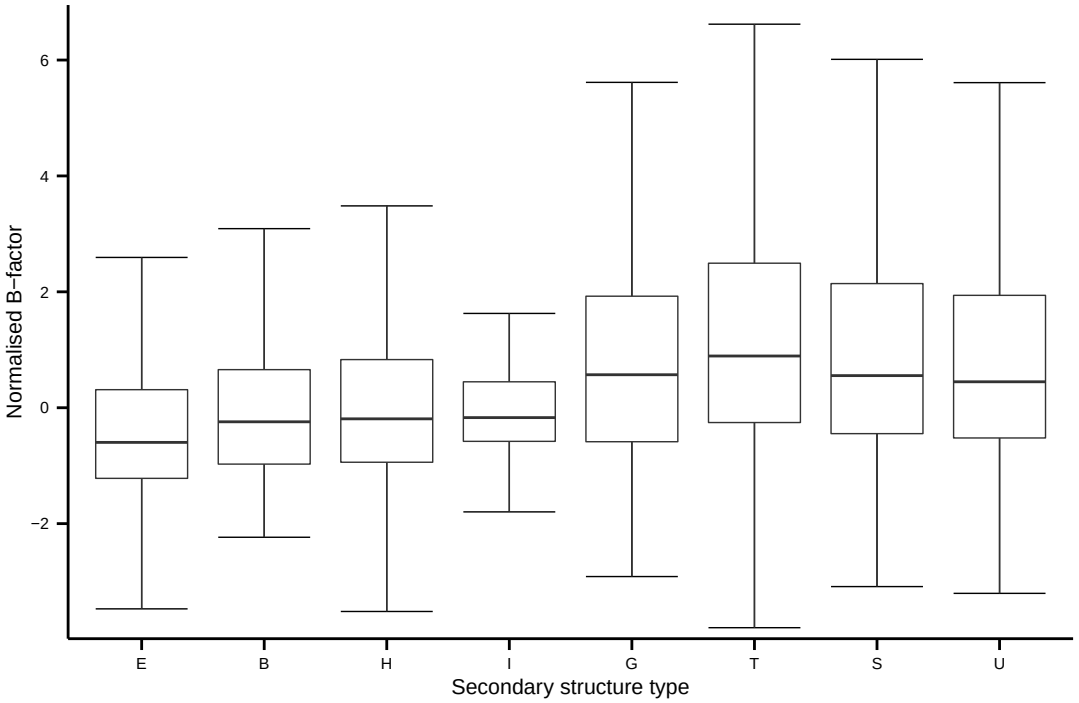
Secondary structure

Secondary structure is maintained by a regular network of hydrogen bonds between the amide groups of the protein's backbone. Adoption of secondary structure would, therefore, be expected to limit the conformational variability of the backbone atoms. Work by Yuan *et al.* (2003) and Sonavane *et al.* (2013) have shown that normalised B-factors of alpha-carbons in regions of α-helix or β-sheet are generally lower than those in other types of secondary structure. The results presented in figure 3.8a clearly show, as might be expected, that the alpha-carbon atoms of residues held within extended regions of regular secondary structure have the lowest B-factors. B-factors of delta-carbons were also considered in order to investigate whether the restraining effect of secondary structure might be just limited to the atoms of the protein backbone. The B-factors of the delta-carbon atoms (figure 3.8b) are higher than those of the alpha-carbons but, interestingly, still maintain differences between the secondary structure classifications. A possible explanation could be that the side chains of residues in regular secondary structure might be restrained through interactions that maintain higher level "super-secondary" structural motifs such as the "beta-sandwich" or "greek key".

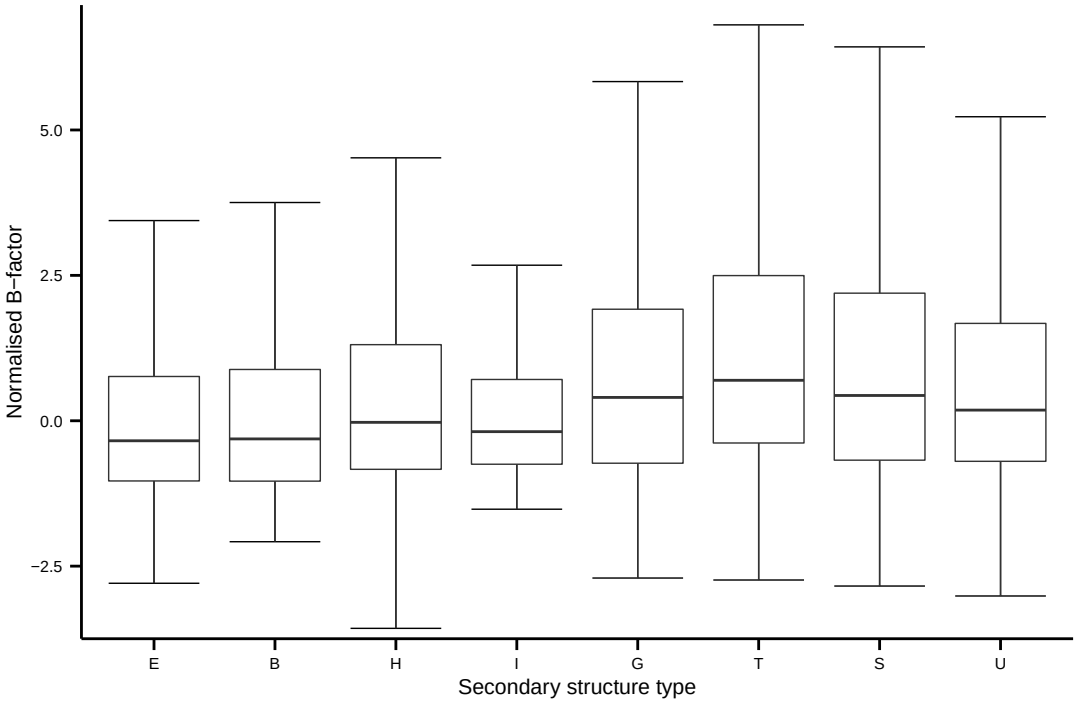
The effect of regular secondary structure on B-factors of alpha-carbon might be attributed to

a dampening of the thermal fluctuations of the backbone atoms; other contributors to the B-factor values, such as large scale rigid body motion and crystallographic defects, are unlikely to be lessened by secondary structure. Despite "temperature factor" being a misnomer, the analysis suggests that temperature dependent atomic motion must be a partial determiner of B-factors. Nonetheless, the extent to which B-factors are affected by thermal fluctuations cannot be easily quantified.

Figure 3.8: Normalised B-factors grouped according to secondary structure. Secondary structure labels are the DSSP classifications: E : β (extended); B : β (bridge); H : α-helix; I : π-helix; G : 3-10 helix; T : turn; S : bend; and U : unclassified ("coil").



(a) Normalised Alpha-carbon B-factors. The proportion of outliers was less or equal to 7% for all groupings



(b) Normalised Delta-carbon B-factors. The proportion of outliers was less than 7% for all groupings.

Amino acid type

The distributions of alpha-carbons B-factors among the twenty major amino acid types is shown in figure 3.9a. Along the horizontal axis of the figure (from left to right), the amino acids are broadly grouped into the following categories: "atypical", acidic, basic, polar and non-polar. Although there is considerable variation in the data, the overall trend observed is clear: amino acids with the most hydrophobic side chains have the lowest B-factors.

As might have been expected, the amino acid B-factor profiles do not deviate significantly from the isotropic B-factor profiles derived by Parthasarathy and Murthy (1997) and Smith, Radivojac *et al.* (2003) and discussed by Sonavane *et al.* (2013). Previous studies explain the differences between amino acids in terms of the relative frequencies of occurrence within the interiors of proteins. Hydrophobic amino acids are typically buried inside a protein, and the low B-factors are a reflection of the limited conformational freedom within the tightly packed interior. Conversely, hydrophilic amino acids at the surface would experience significantly less hindrance.

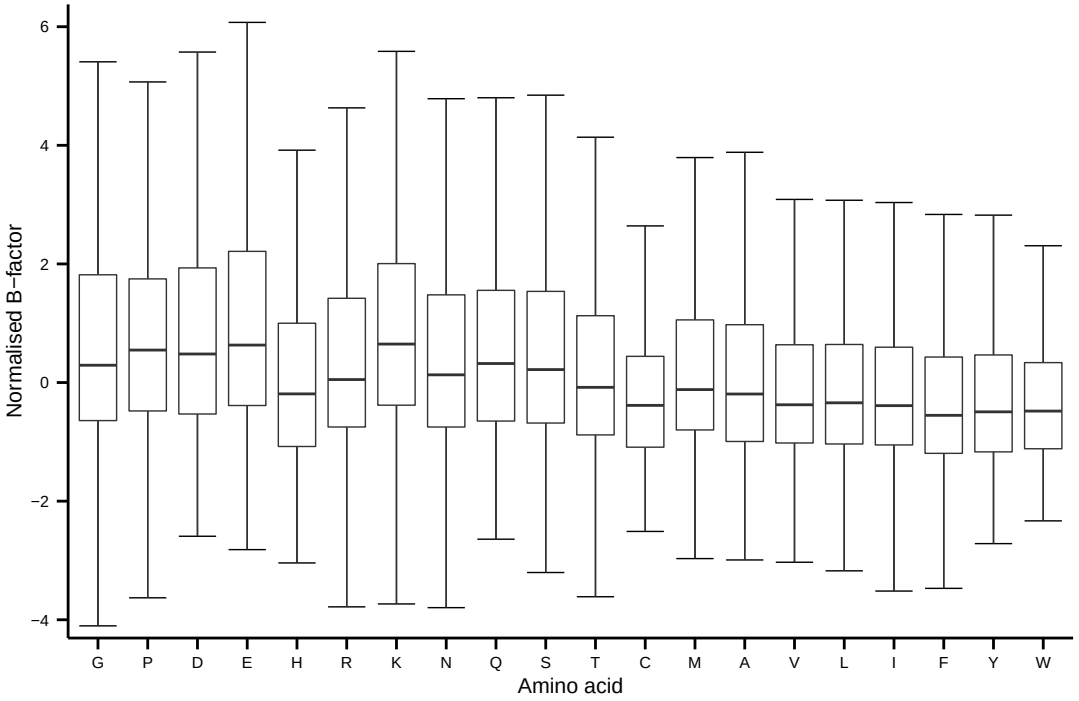
The analysis of Parthasarathy and Murthy (1997) acknowledges that B-factor distributions cannot be satisfactory explained by amino acid hydrophobicity alone. This is supported by the results of this thesis which suggest that all aspects of an amino acid side chain chemistry can have an effect on B-factor values. Glycine exhibits a very broad distribution of high B-factor values which can be attributed to the greater conformational freedom allowed by the absence of a side chain. Similarly, the low B-factors observed for cysteine and histidine may be a consequence of the restraining effect of the bonded and non-bonded interactions in which they participate. Cysteines would be severely restricted when linked via disulphide bridges, but the restrictions imposed on histidine are not so obvious. In enzymes, histidine arises more frequently at active sites in comparison to the rest of the structure (Bartlett *et al.* 2002; Holliday *et al.* 2007) strongly suggesting that histidine has a key functional role. The B-factors at an enzyme's active site are typically lower than those in equivalent environments at other locations (Yuan *et al.* 2003). Hence, it is feasible that the presence of catalytically active histidine residues may contribute to the low B-factor values of histidine observed in the data set.

The effect of side chain chemistry can be dramatically illustrated with proline. The alpha-carbon B-factor distribution of proline is comparable to that of glycine due to its constrained cyclic structure. Proline's delta-carbons (figure 3.9b) have very low B-factors compared to the highly mobile delta-carbons of other amino acids. The difference can be directly attributed to proline's delta-carbon being covalently bonded to the backbone nitrogen.

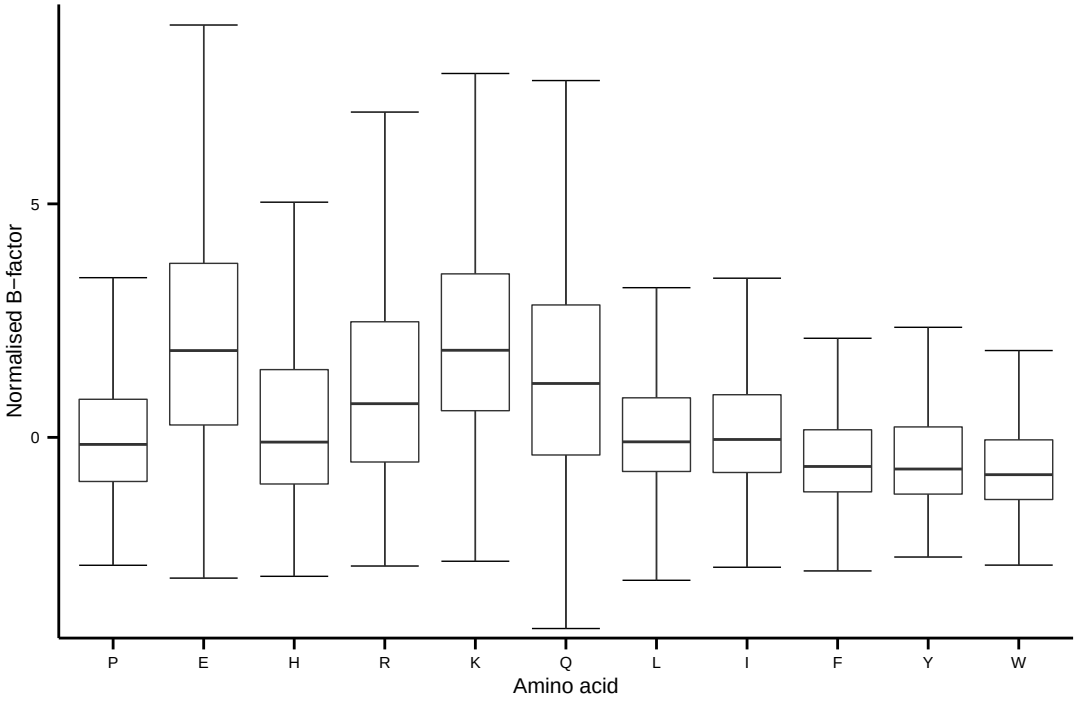
It is difficult to find a single, satisfactory explanation to account for the differences between the B-factors for the different amino acid types. Hydrophobicity is important because it determines whether a particular class of amino acid is likely to be located at the surface or

buried within the interior. In addition, amino acid chemistry and the role of the residue in the protein will have an effect. For example, the conformational dynamics of a cysteine residue will be very different depending on whether that residue is part of a disulphide bridge, located at an enzyme's active site or has no specific structural or functional role. Despite an uncertainty in the underlying causes, the data shows a general correlation between B-factor values and the amino acid types expected to exhibit the greatest conformational fluctuations. Nevertheless, the high degree of variability in the dataset makes it difficult to derive an amino acid "flexibility index" (Karplus and Schulz 1985) based on B-factor values alone.

Figure 3.9: Boxplots of normalised B-factors grouped according to amino acid type



(a) Normalised Alpha-carbon B-factors. The proportion of outliers was less than 7% in all groupings except for methionine (9.5%) and tryptophan (7.2%).



(b) Delta-carbon B-factors. The proportion of outliers was less than 6% in all groupings.

Solvent exposure and surface depth

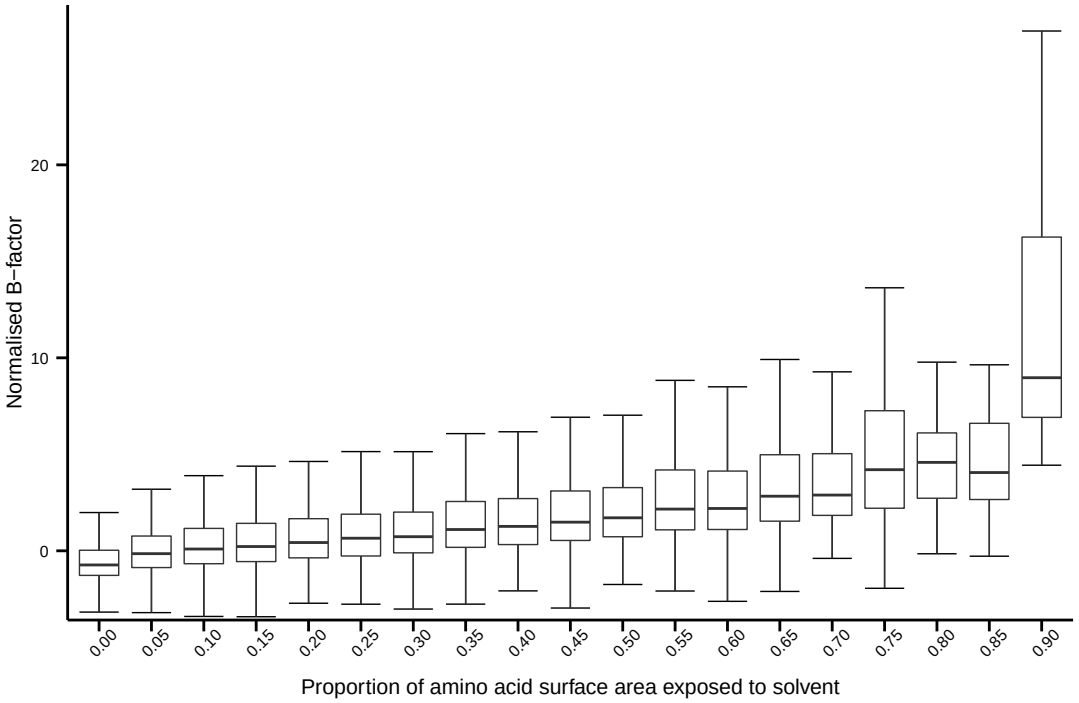
The amino acid analysis suggests that the location of an atom at the protein's surface or interior is an important factor in determining its B-factor. Consequently, it was decided to investigate surface effects in more detail. The first feature considered was the exposure of an amino acid's atoms to the solvent. Exposure was measured as a normalised ratio of an amino acid's SASA in the protein to the SASA of the same amino acid in the same conformation and surrounded by the backbone atoms of two sequentially adjacent amino acids. Initially, only the normalised SASA ratios for individual alpha-carbon atoms were measured, but this was found to be poor measure of surface exposure because not all surface amino acids have an alpha-carbon in direct contact with the solvent. It was, therefore, necessary to calculate the SASA of the whole amino acid even though only the B-factors of the alpha-carbons are considered.

The relationship between normalised amino acid SASA and alpha-carbon B-factors is visualised in figure 3.10. The general trend observed is an increase in the B-factor values as solvent exposure increases. The simplest interpretation is that greater solvent exposure will mean fewer contacts with neighbouring amino acids and, thus, greater conformational freedom. It is also conceivable that the dynamics of the protein, when in direct contact with the solvent, could be influenced by the motion of the solvent's atoms through a "buffeting" action.

The effect of solvent exposure on B-factor values have been studied extensively. Carugo and Argos (1997) interpreted the increase in mobility of amino acid side chains with increased solvent exposure as being a consequence of the removal of restrictions that limit conformation freedom in the tightly packed protein core. A strong correlation between solvent exposure and B-factor values was observed in the α-helices of hemerythrins. The B-factor profiles of these proteins exhibited a periodicity that coincided with that of the exposure of residues in the external α-helices (Sheriff *et al.* 1985). However, the effect of solvent exposure on B-factor values may be more difficult to pinpoint in the general case. An analysis by Zhang *et al.* (2009) deduced that the solvent can exert a long range influence of the conformational dynamics of a protein. Specifically, an alpha-carbon's B-factor value can be affected by the neighbouring amino acids' exposure to the solvent.

There are other, equally feasible, explanations for the correlation between SASA and B-factor values that are independent of a protein's inherent conformational flexibility. The protein surface may be more likely to exhibit static conformational variation as opposed to the buried residues of the core. Atoms at the surface may also undergo greater displacements as a consequence of whole protein rigid body librations as described by the Translation Libration Screw (TLS) model (Schomaker and Trueblood 1968). It is also feasible that, rather than being local effects, the protein-protein contacts of the crystal lattice may cause perturbations affecting dynamics over a wide area of the protein surface.

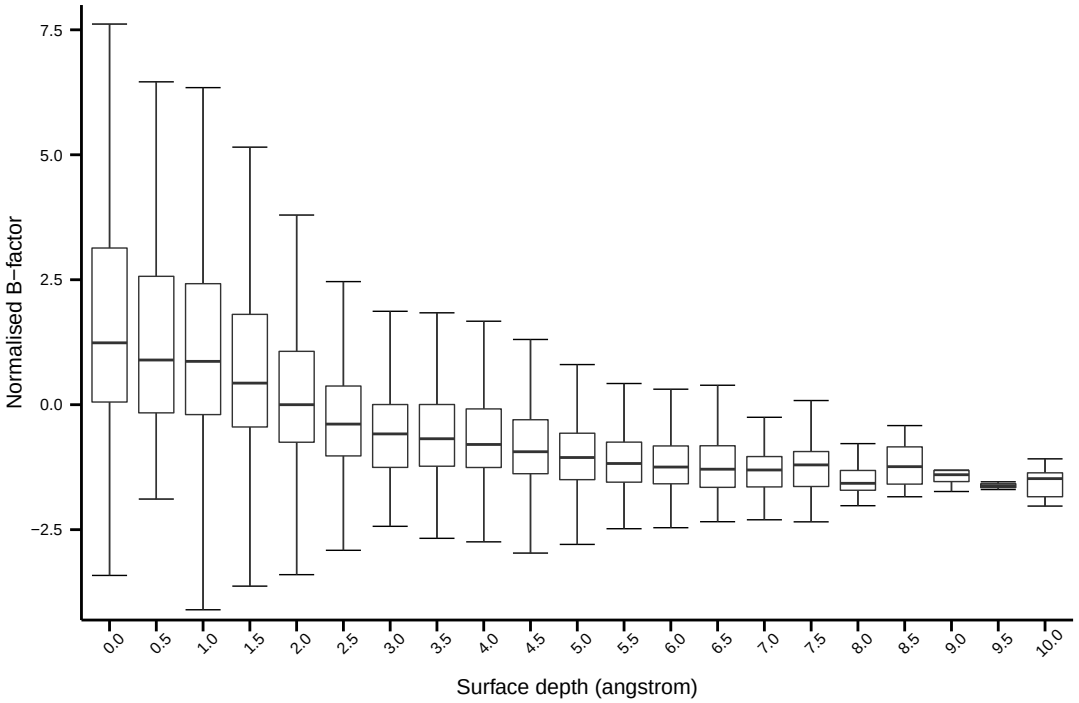
Figure 3.10: Boxplots of alpha-carbon B-factors grouped according to normalised SASA for the amino acid. The bin width is 0.05 units except for the final bin (0.9 to 1.0). The proportions of all outliers were less than 7% in each grouping except at 0.3 (7.2%).



Calculating an amino acid SASA is a coarse metric of surface exposure. Therefore, the alternative approach of measuring the depths of atoms from the protein surface was also pursued. The relationship between atom depth and B-factors is the reciprocal to that observed for amino acid SASA ratios (figure 3.11). The general trend can be explained with similar arguments to those used when considering amino acid SASA. Interestingly, the isotropic B-factor analysis by Sonavane *et al.* (2013) found that the mean normalised B-factor values of alpha-carbons decreased by 0.1055 for every 0.5 Å from the protein surface. With respect to median-mad normalised B-factors, the decrease in the median values of the boxplots in figure 3.11 is approximately 0.13 for every 0.5 Å (estimated by linear regression). However, due to the considerable degree of variation in the data, it is impossible to attach much significance to this value.

It would be remiss not to discuss the limitations of the methodology as a reason for the high variability observed when relating surface exposure to B-factor values. Of all the protein properties investigated, surface exposure is the one attribute that cannot be quantified with precision. The need to implement a customised version of the SASA algorithm meant that complex structural features such as the presence of internal cavities were not accounted for. Nonetheless, it can be argued that the algorithm used is adequate because none of the proteins in this study are multi-subunit complexes which are more likely to form structures

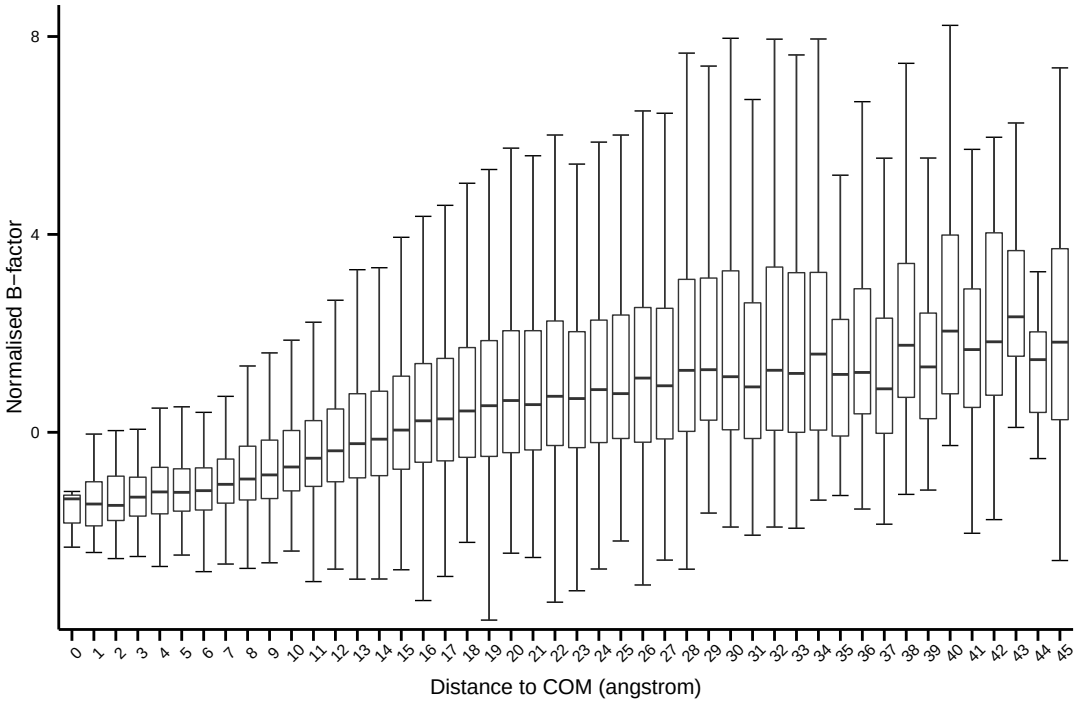
Figure 3.11: Boxplots of alpha-carbon B-factors grouped according to the atom's distance from the surface. The bin width is 0.5 Å. The proportions of all outliers were less than 6% in each grouping except ≥7.0 Å (5.7-16.7%).



with large internal channels or pockets. Although, measuring SASA to a high degree of precision is desirable, it is more important that the SASA measurements account for the protein-protein contacts across the crystal lattice.

Another source of error in the SASA calculations was the omission of any surface occlusion effects from the non-protein molecules in the crystal. Although large ligands and cofactors were excluded, small organic molecules and salts were present in some structures. The proteins may have also been tightly bound to ion cofactors such as zinc, cadmium, iron or copper. Despite being a source of error, it would have been impractical to account for the presence of every compound in the crystal and, in particular, in those cases where the molecules were only partially resolved or completely unresolved. The fact that only crystals containing small non-protein compounds were permitted and that over 114 diverse proteins were examined should mean that these oversights are insignificant in the overall analysis. Furthermore, different compounds may affect protein dynamics in ways that are completely independent of surface occlusion. For example, the influence of an ionic compound's electrostatic interactions with the protein are likely to perturb dynamics to a greater extent than any solvent shielding effects. Although an interesting area of research, a full consideration of the effects of intermolecular forces is outside the scope of this work.

Figure 3.12: Boxplots of alpha-carbon B-factors grouped according to the atoms's distance to the protein's COM. The bin width is 1 Å except for the final bin (≥45 Å). The proportions of all outliers were less than 7% in each grouping in the range 3–36 Å and up to 21% otherwise

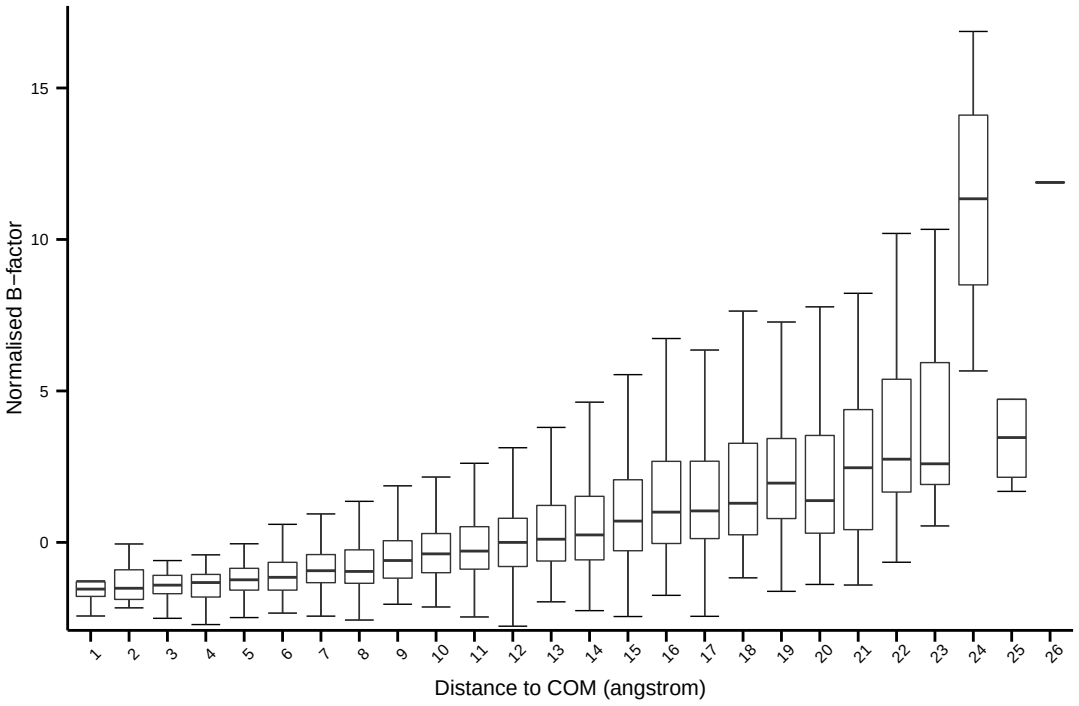


Distance of the atom from the protein's centre of mass

The relationship between B-factors and distance to the COM was investigated to determine if there was a significant contribution to B-factor values from the movements of the protein as a rigid body. In a simplified version of the TLS model, atomic displacements due to whole protein librations ("rocking" movements) are assumed to be proportional to the square of the distance of the atom to the protein's COM (Kundu *et al.* 2002). This simple model is not supported by the results (figure 3.12) since the B-factor values appear to tail-off rather than increase quadratically as the distance to the COM increases.

Figure 3.12 reveals something about the shapes of the proteins in the data set. If all the proteins are assumed to be approximately spherical with a uniform distribution of mass, then the distance to the COM would be negatively correlated with surface depth. On the basis of the previous results, B-factors would be expected to increase as the distance to the COM increases. Not only do the results deviate from this hypothetical relationship, but the range of distances to the COM are five-fold higher than those measured for surface depth. Hence, the data set must exhibit a diversity of different protein shapes. This hypothesis was tested by filtering the dataset to include only the most "spherical" proteins. A spherically shaped protein was defined to be a structure where the standard deviation in the distances from the

Figure 3.13: Boxplots of alpha-carbon B-factors grouped according to the atoms's distance to the protein's COM for the most spherical proteins in the data set. The bin width is 1Å. The proportions of all outliers were less than 8% in each grouping except for 1,18,23 and 25 Å at 16.7,8.2,10.0 and 20.0% respectively.



COM to the atoms at the surface are less than 3.5 Å. Applying this condition selected only 31 from the 114 proteins of the data set. Arguably, the proteins selected are not particularly spherical in shape, but this criteria filtered out the most irregularly shaped proteins whilst still retaining a sufficient number for study. Repeating the analysis with these spherical proteins gave a result that broadly supported the hypothesis. The maximum distance to the COM was reduced from 45 Å to approximately 26 Å and a positive correlation was observed between the distance to the COM and the B-factors (figure 3.13). Nevertheless, it would be inappropriate to draw any definitive conclusions from such a small sample.

It is not possible to explain the relationship between the distance to the COM and B-factors without a much more detailed analysis of the shapes of the proteins. Proteins consisting of multiple domains would be expected to undergo rigid-body motions more complex than simple librations. Furthermore, there would be no linear correlation between surface exposure and distance from the COM for an irregularly shaped protein.

Amino acid coordination number

The investigation into the effect of distance to the COM prompted a consideration of how the shape and mass distribution in a protein might affect B-factor values. Shape is very difficult to quantify, but density is easier to measure. Amino acid coordination number (Nishikawa and Ooi 1980; Pollastri *et al.* 2002), also known as the contact number, was used as a simple indicator of how closely amino acids were packed together.

Before calculating the coordination numbers, a suitable cutoff distance was sought. Ideally, the cutoff distance should extend from each alpha-carbon atom to encompass only the alpha-carbons of immediately adjacent amino acids. The cutoff distance was selected by analysing the distribution of the inter-alpha-carbon distances over all the proteins of the data set. All inter-alpha-carbon distances less than 20 Å were measured for each protein. Distances were converted to frequencies using a bin size of 0.5 Å and expressed as normalised ratios by dividing by the total number of distinct alpha-carbon pairings. The median ratio for each distance increment was calculated and plotted (figure 3.14). Two maxima are apparent from the graph. The first peak at 3.5–4 Å corresponds to the average distance between alpha-carbons of sequentially adjacent amino acids. The second broader peak lying within the range 5–8 Å was interpreted to be the range of distances between alpha-carbons of spatially adjacent amino acids. Therefore, 8 Å was chosen as a suitable distance to calculate alpha-carbon coordination numbers. Reassuringly, the cutoff was consistent with the results of a similar analysis by Halle (2002), indicating that the structures are representative of proteins in general.

The results of the coordination number analysis are shown in figure 3.15. There is a decrease in the normalised alpha-carbon B-factor as the coordination number of the amino acids increase. High coordination numbers indicate a tightly packed region of the protein and, it is hypothesised that, the corresponding low B-factors are a direct consequence of limited conformational freedom. Hindrance to movement could be a result of insufficient free space or strong non-bonded interactions between other atoms in close proximity. Irrespective of the mechanisms involved, the effect of coordination number will only impact protein dynamics in terms of thermal fluctuations and local rigid body motion.

In this analysis, no consideration was made to the effect of the different amino acid types and their relative sizes on coordination number. For example, a tightly packed group of aromatic amino acids might have a lower alpha-carbon coordination number compared to a sparse cluster of amino acids with less bulky side chains. This issue was addressed by recalculating the coordination numbers using all neighbouring atoms over the same 8Å radius rather than just alpha-carbons. The trend observed was the same (figure 3.16), suggesting that the relationship between local packing density and B-factors is the same irrespective of how the packing is measured. These findings are consistent with the theoretical Local Density Model (LDM) proposed by Halle (2002) where isotropic B-factor values are governed by atomic

Figure 3.14: Distribution of alpha-carbon to alpha-carbon distances for the maximum occupancy protein structures of the dataset. Lower and upper quartiles are represented with error bars to indicate the spread of proportions.

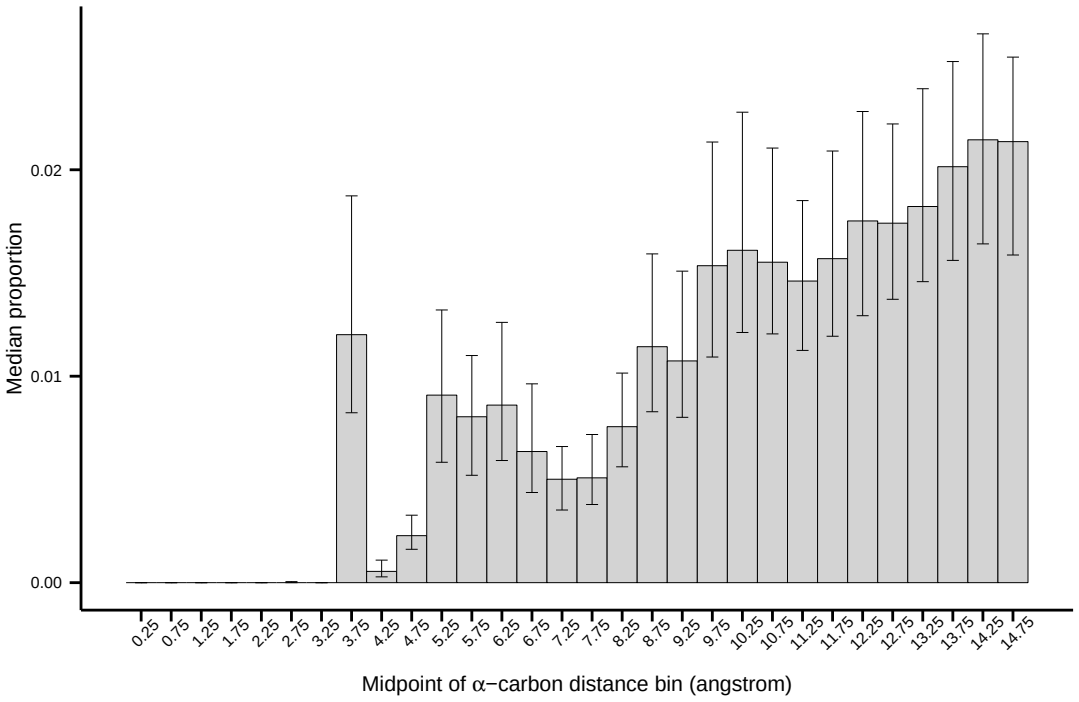


Figure 3.15: Boxplots of alpha-carbon B-factors grouped according to the coordination number of the amino acid.The proportions of all outliers were less than 6% in each grouping except at 3 (7.1%).

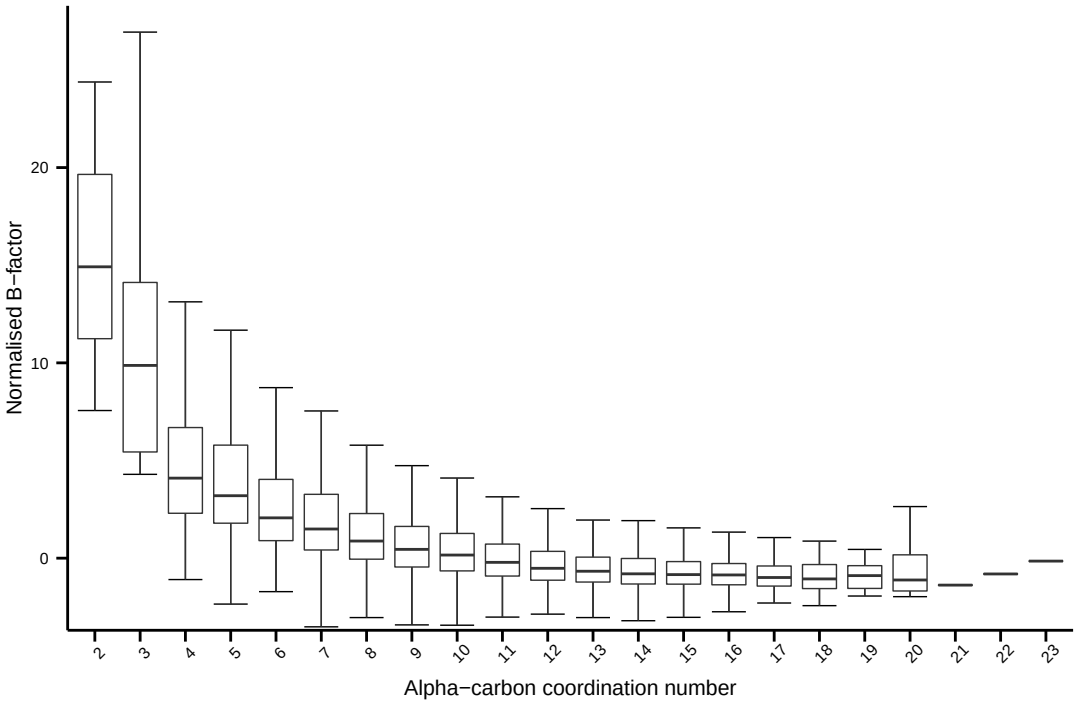
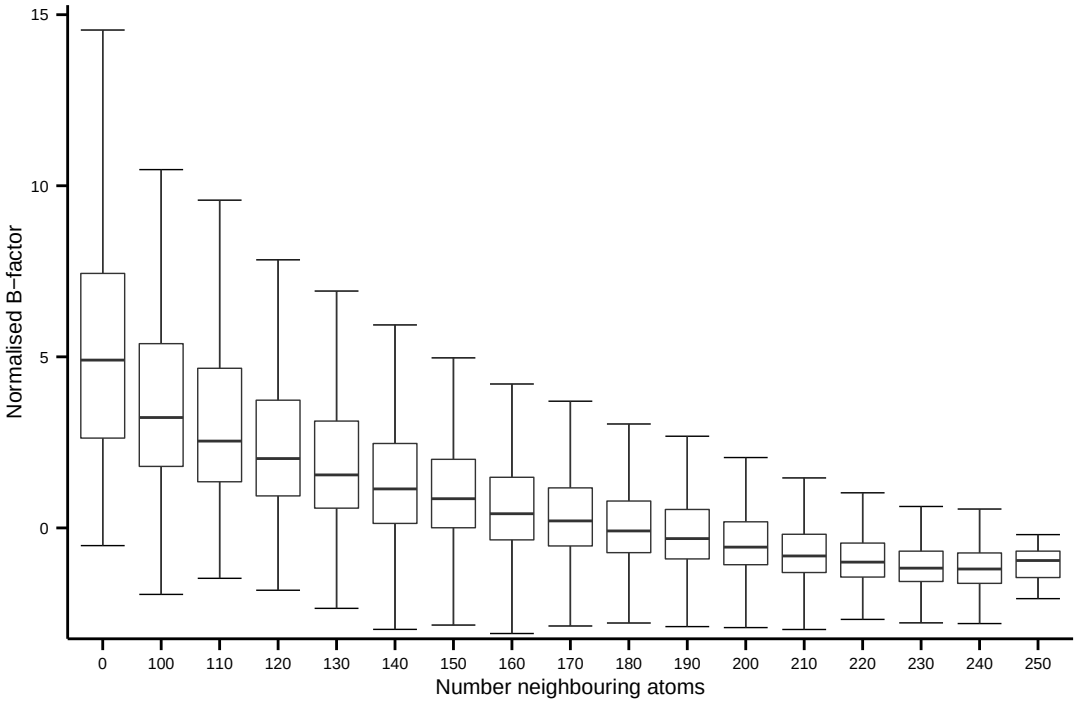


Figure 3.16: Boxplots of alpha-carbon B-factors grouped according to the number of neighbouring atoms. The proportions of all outliers were less than 6% in each grouping.



packing densities within the crystal.

Although coordination number appears to be an ideal parameter to explain B-factor values, there are potential shortcomings with measuring local density in this way. Surface amino acids will always have low coordination numbers even when their side chains are internalised and packed tightly together. A more rigorous treatment of surface amino acids would be necessary to account for this effect.

Effect of protein contacts

The protein-protein contacts of the crystal lattice must be taken into account if SASA and coordination numbers are to be accurately calculated. This was the main reason the structure of the crystal lattice was reconstructed in the analyses. Failure to do so would have resulted in underestimating packing density and overestimating SASA in regions where proteins interact. Despite the greater accuracy, calculations involving lattice symmetries are more complex and computationally intensive than the equivalent calculations on single isolated proteins. Consequently, it was questioned whether there is sufficient benefit in reconstructing periodic unit cells when analysing B-factor data. In order to answer this question, the calculations were repeated using only the crystals' asymmetric units in the absence of any lattice symmetries.

The effect of crystal contacts on the relationship between coordination number and B-factors is shown in figure 3.17. The graph superimposes the B-factor profiles against coordination number calculated for both crystal lattices and isolated proteins. The effect of protein-protein contacts in the crystal is apparent from figure 3.17 where the B-factors of amino acids with low coordination numbers (exposed residues at the surface) are much higher when measured in the crystal lattice. Presumably, this effect is due to the incorrect assignment of low coordination numbers to atoms close to protein-protein interaction sites. These atoms would have less freedom in movement and, therefore, lower B-factors are to be expected. However, the extent to which crystal contacts dampen B-factors values cannot be quantified since there is no data for equivalent non-crystalline structures. Furthermore, due to the high degree of variability in B-factor values across the dataset, any attempt to apply a "correction factor" to estimate B-factor values in the absence of crystal contacts would be highly unreliable.

Figure 3.17: Effect of crystal contacts on coordination number. Median normalised alpha-carbon B-factors plotted against coordination number for single proteins and proteins in the crystal. The lower and upper quartiles are plotted as upward and downward pointing triangles respectively

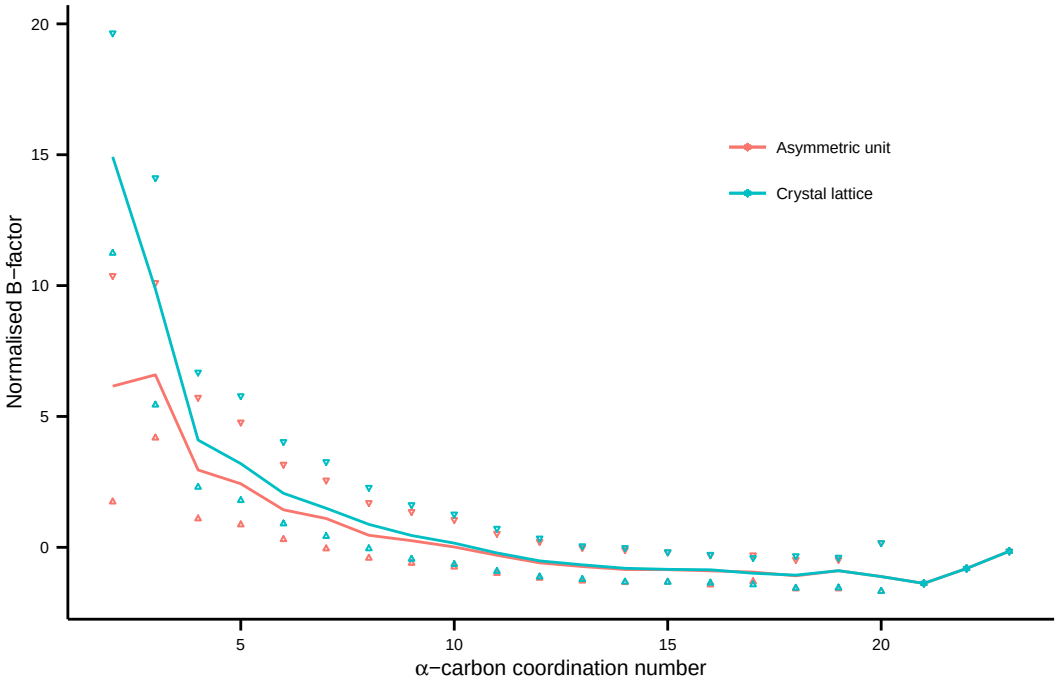


Table 3.4: Correlations between structural properties measured for proteins in the crystal lattice. Spearman correlation coefficients calculated to 3 d.p. Equivalent correlation coefficients measured for the isolated asymmetric units are in italics. *Note:* the distance to the COM is always measured in an isolated protein.

	SASA	**Depth**	**COM distance**	**Coord. Number**
SASA	1.000	-0.675 *-0.693*	0.386 *-0.384*	-0.777 *-0.697*
Depth		1.000	-0.324 *-0.327*	0.652 *0.593*
COM distance			1.000	-0.383 *-0.353*
Coord. Number				1.000

3.4.5 Correlations between structural properties

The protein attributes measured in this study are not completely independent of one another. Moving through the protein, from the surface to the interior, surface depth and coordination number increase while distance to the COM and SASA decrease. Hence, positive correlation is presumed between surface depth and coordination number and both quantities would be expected to correlate negatively against solvent exposure and the distance to the COM. Correlation coefficients were calculated between all the structural properties in order to expose any redundancy in the analysis. Spearman's method was used to calculate the correlations since neither linear relationships nor normally distributed data can be assumed. The calculations (table 3.4) are consistent with the expected relationships and, unsurprisingly, the weakest correlations are those involving the distance to the COM which can be attributed to the diversity of protein shapes discussed previously. A reasonable correlation between coordination number and solvent exposure supports the assertion that coordination number, and related measures, can be used as a computationally inexpensive estimate for solvent exposure (Hamelryck 2005). From the perspective of structural bioinformatics, this has obvious applications when approximating solvent exposure for proteins.

An unanticipated finding is the difference between the correlations for the properties measured in a crystal lattice compared to those measured for isolated proteins. These differences, particularly those involving coordination number, highlight why it is important to consider the lattice when analysing crystallographic data.

3.4.6 Strategies to reduce the variation in B-factor data

The analysis discussed previously broadly supports the hypothesis that B-factors are indicators of conformational variability within the crystal. However, contrary to what might have been expected, high quality crystallographic data and fully accounting for crystal contacts did not make it possible to derive quantitative relationships between B-factors and the structural properties that correlate with conformational dynamics. B-factor values are highly variable across the proteins in the data set and this makes it impossible to deduce anything other than broad qualitative trends from the data. Subsequently, this led to an investigation into strategies that could be employed to reduce the variability of the B-factor data. The two approaches taken were: transforming the data with a mathematical function and applying different normalisation techniques to standardise the B-factors of each protein.

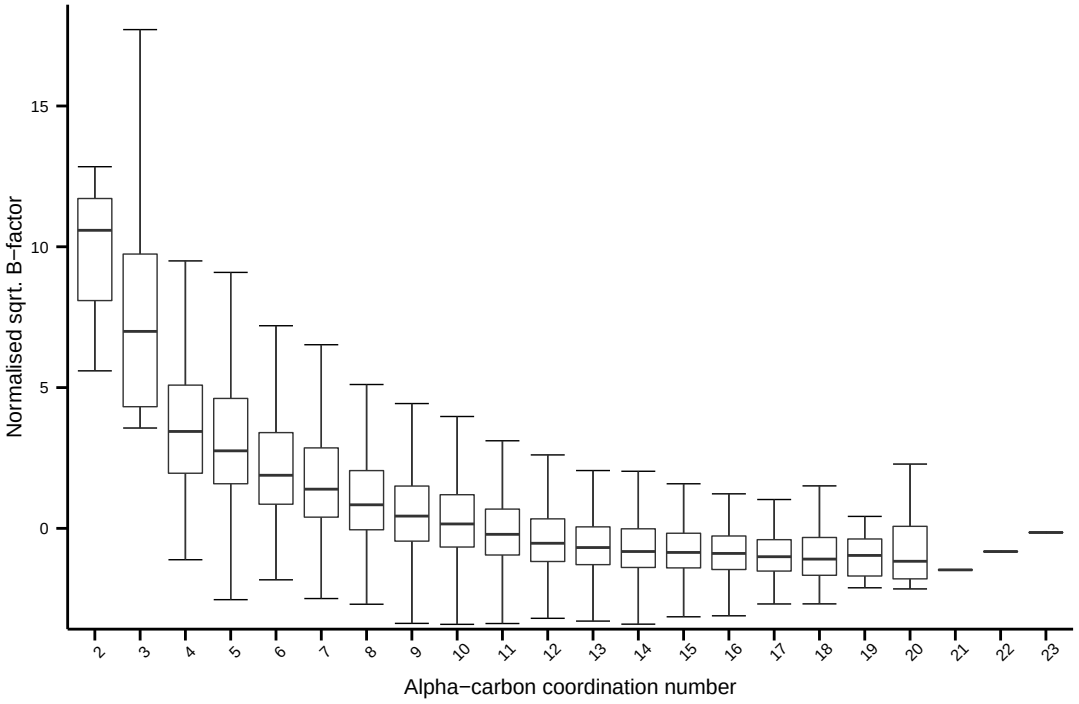
Transforming B-factors

As has been commented previously, the spread of B-factor values is highly positively skewed. The asymmetry is partly a consequence of the definition of isotropic B-factors as mean square distances i.e., quadratic functions of atomic displacements (equation 1.1). It was therefore logical to compensate for this effect by taking the square root of the B-factor. Applying the square root transform to the data reduced the degree of skew, but the distribution still remained positively skewed (skewness measure reduced from 3.021 to 1.922). Interestingly, the skew can be reduced further by transforming the B-factor data with the natural logarithm function (skewness measure at 1.185). However, unlike the square root, the natural logarithm of a B-factor cannot be easily interpreted in terms of atomic fluctuations. Despite reducing the skewness of the data set, neither the square root nor the natural logarithm appeared to reduce the spread of B-factor values. Furthermore, the trends observed remained the same irrespective of whether or not a transform was applied. As an example, figures 3.18a and 3.18b are boxplots between B-factors and alpha-carbon coordination number when the square root and natural logarithm are applied. A comparison between the transformed data in figures 3.18a and 3.18b with the original data in figure 3.15 suggests that there is no advantage in applying these transforms.

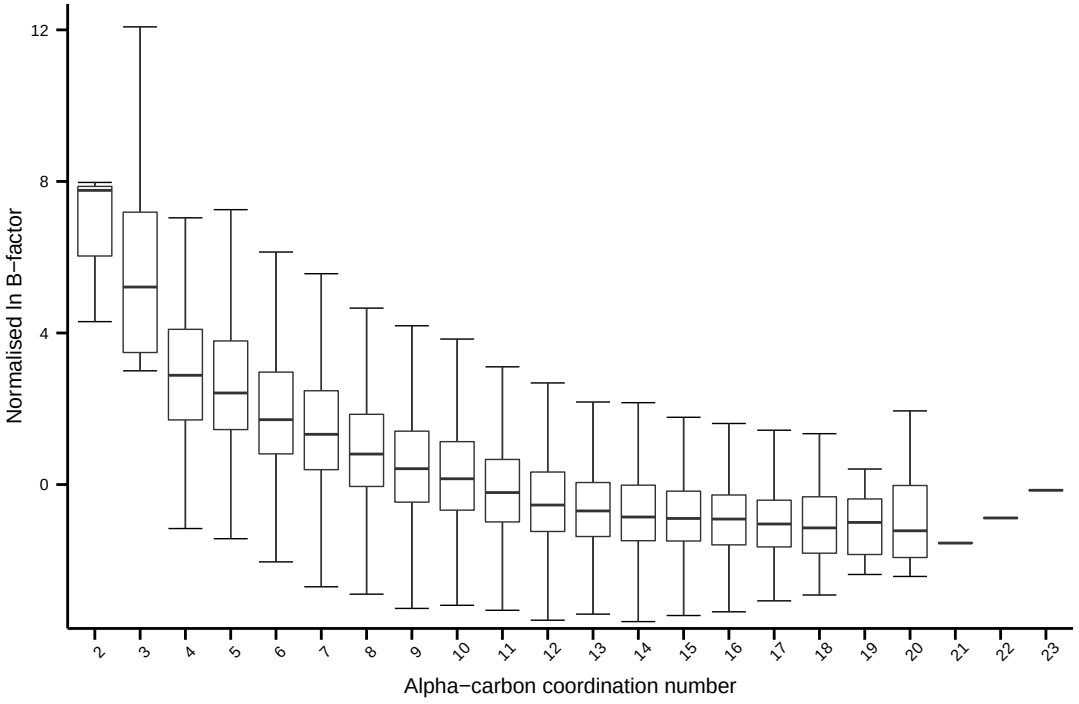
Normalising B-factors

It is feasible that the high level of variability in the B-factor data might be a consequence of the differing conditions under which crystal structures were determined. The structures have been derived from protein crystals across a range of different temperatures, pressures, pH values and solvents. In addition, different laboratories will have used different equipment and followed different methodologies to obtain, interpret and refine the crystallographic data. In response, most, if not all, previous research on B-factors has attempted to standardise

Figure 3.18: Boxplots of transformed alpha-carbon B-factors grouped according to the coordination number of the amino acid



(a) Square root transform. Proportion of outliers less than 6% in all groupings.



(b) Natural logarithm transform. Proportion of outliers less than 5% in all groupings.

the data by normalising the B-factors within each protein. Unfortunately, there is no clear consensus on which normalisation techniques are the most successful at eliminating the variability between structures.

In the absence of evidence from previous research, three widely used normalisation methods were compared to assess what effect, if any, normalisation would have on the B-factor data. The three normalisation techniques compared were: mean-standard deviation ("z"-normalisation); median-mad (z-normalization using robust statistics) and min-max scaling. Full definitions of these normalisation methods are given in the methods section of this chapter.

A simple metric was devised to quantify the extent to which normalisation reduced the spread of B-factor values. The definition of this metric is given in equation 3.2 below:

$$\text{Score} = \frac{\left|\text{median}\left\{B^{norm}_{dynamic}\right\} - \text{median}\left\{B^{norm}_{rigid}\right\}\right|}{\max\left[\text{mad}\left\{B^{norm}_{dynamic}\right\}, \text{mad}\left\{B^{norm}_{rigid}\right\}\right]} \tag{3.2}$$

where,

$B^{norm}_{dynamic}$ are the normalised B-factors of a subset of alpha-carbons in a region of the protein expected to be conformationally dynamic.

B^{norm}_{rigid} are the normalised B-factors of a sunset of alpha-carbons in a region of the protein expected to be conformationally rigid.

The scores calculated using equation 3.2 should always be positive and, over all criteria used to divide atoms into "dynamic" and "rigid" subsets, the normalisation method that consistently scores highly will be the method that is most effective at standardising B-factors. The rationale being that an optimal normalisation method should maximise the difference between the average normalised B-factors of the most dynamic and rigid atoms whilst minimising the spread of normalised B-factors for those two groups of atoms. Hence, in the numerator of equation 3.2, the difference between the median normalised B-factors should be large. Simultaneously, in the denominator, the maximum MAD across both sets of normalised B-factors should be small. Robust measure of location (median) and spread (MAD) are used to limit distortion from atypically high or low normalised B-factors.

The score defined by equation 3.2 is independent of the criteria used to define subsets of alpha-carbon expected to be conformationally dynamic or rigid. Using the results described previously, a number of different definitions of conformationally dynamic and rigid groups of atoms were formulated. These definitions are summarised in table 3.5.

The results of calculating the scores defined by equation 3.2 for each of the three normal-

Table 3.5: Criteria used to define dynamic and rigid groups of atoms

Atom subset criterion	Conformationally dynamic	Conformationally rigid
secondary structure	unclassifiable	extended β-structure
normalised amino acid SASA	> 0.5	< 0.01
coordination number	≤ 5	≥ 15
amino acid type	aspartate	tryptophan

isation methods are given in table 3.6. For comparison, the calculations included the scores for B-factors that had not been normalised. In addition, a criterion that did not distinguish between conformationally dynamic and rigid atoms was included as a control. This control simply selected 5000 alpha-carbon atoms at random for the dynamic and rigid subsets and would, therefore, always be expected to give a score close to zero. The table also includes the results for the normalisation of B-factors that were transformed with the square root or the natural logarithm functions. This was done for completeness as it is conceivable that a combination of normalisation and data transformation may be the optimal solution.

The calculations in table 3.6 demonstrate that normalisation has an effect on reducing the variability in the B-factor data. Irrespective of how the atoms were divided into dynamic and rigid subsets, both mean-standard deviation and median-mad normalisation methods gave scores that showed an improvement compared to when no normalisation was applied. There is some evidence to suggest that median-mad normalisation offers a fractional improvement over mean-standard deviation. However, since the underlying distributions for the metric scores are unknown, it is impossible to confirm whether these differences are statistically significant. Table 3.7 gives a rough indication of the level of "background noise" expected with the calculations. In the case of mean-standard deviation and median-mad normalisation, these values are comparable to the differences between the scores for these two methods in table 3.6. Therefore, the effects of median-mad and mean-standard deviation normalisation appear to be roughly equivalent. Min-max scaling, whilst improving the data slightly, does not appear to eliminate inconsistencies in B-factor values as effectively as the other two methods. A possible explanation is that min-max scaling is based on the assumption that the atoms with the highest and lowest B-factors are equivalent in all protein structures. While this may be a reasonable assumption when comparing proteins with a high degree of structural similarity, it may not be appropriate for the diverse set of structures considered here. It is also noteworthy that both median-mad and mean-standard deviation normalisations of transformed B-factors show a slight improvement over the untransformed data. However, because the normalisation scores for the raw and transformed B-factors are not strictly comparable, it is impossible to say whether this is significant.

To confirm that normalisation is beneficial, the analysis was repeated using the raw B-factor

Table 3.6: Comparing B-factor normalisation methods

Atom subset criterion	**transform**	**Normalisation**	**Score**
secondary structure	none	none	0.6575
		mean-sd	0.8897
extended β-structure: 5181 atoms		median-mad	0.9346
unclassified structure : 4365 atoms		min-max	0.8452
	square root	none	0.6826
		mean-sd	0.9311
		median-mad	0.9646
		min-max	0.8880
	natural log	none	0.7159
		mean-sd	0.9774
		median-mad	0.9896
		min-max	0.9350
solvent exposure	none	none	1.5117
		mean-sd	2.0923
SASA $>$ 0.5 : 1605 atoms		median-mad	2.1816
SASA $<$ 0.01 : 5643 atoms		min-max	1.7912
	square root	none	1.6953
		mean-sd	2.3715
		median-mad	2.4625
		min-max	2.0361
	natural log	none	1.9629
		mean-sd	2.7344
		median-mad	2.7939
		min-max	2.3513
coordination number	none	none	1.8561
		mean-sd	2.2594
$\geq$ 15 : 1533 atoms		median-mad	2.2663
$\leq$ 5 : 538 atoms		min-max	1.9514
	square root	none	2.1880
		mean-sd	2.5895
		median-mad	2.5970
		min-max	2.2584
	natural log	none	2.5778
		mean-sd	3.0203
		median-mad	2.9764
		min-max	2.7593
amino acid type	none	none	0.8227
		mean-sd	0.8415
Aspartate : 1428 atoms		median-mad	0.9133
Tryptophan : 360 atoms		min-max	0.8733
	square root	none	0.8611
		mean-sd	0.8969
		median-mad	0.9365
		min-max	0.9231
	natural log	none	0.9283
		mean-sd	0.9520
		median-mad	0.9697
		min-max	0.9598

Table 3.7: Comparing B-factor normalisation methods with random selections of atoms. The random selection of 5000 atoms for the two groups was repeated 5 times. The score reported is the mean of these 5 samples and the associated standard deviation.

transform	**Normalisation**	**Score (mean ± sd)**
none	none	0.0279 ± 0.0125
	mean-sd	0.0497 ± 0.0299
	median-mad	0.0122 ± 0.0141
	min-max	0.0374 ± 0.0233
square root	none	0.0445 ± 0.0234
	mean-sd	0.0399 ± 0.0424
	median-mad	0.0210 ± 0.0109
	min-max	0.0251 ± 0.0258
natural log	none	0.0243 ± 0.0234
	mean-sd	0.0291 ± 0.0114
	median-mad	0.0125 ± 0.0191
	min-max	0.0312 ± 0.0308

data. Figures 3.19 and 3.20 are the boxplots for raw B-factors against the distance to the protein's COM and surface. When compared to the normalised data in figures 3.12 and 3.11, normalisation has reduced some of the variability in the data. The relative sizes of the interquartile ranges and boxplot "whiskers" are reduced when compared to the plots where no normalisation has been applied. Furthermore, the trends observed appear to be clearer in the normalised plots. A comparison between figures 3.12 and 3.19 for the COM distance analysis provides a particularly striking example. Median normalised B-factors increase smoothly as the distance to the COM increases while median B-factors in the non-normalised plot wildly fluctuate once the distance to the COM exceeds 30 Å. In the case of surface depth, the normalised boxplots clearly show a reduction in both the median and the spread of B-factor values the deeper the atom is buried within the protein. This is consistent with the model that atoms within the protein interior have less conformational freedom compared to those near the surface. Although the non-normalised boxplot (figure 3.20) shows a reduction in the median B-factor values as depth increases, the reduction in the spread in B-factor values is less apparent.

Figure 3.19: Boxplots of alpha-carbon B-factors grouped according to the atoms's distance to the protein's COM. The bin width is 1 Å except for the final bin (≥ 45 Å). The proportions of all outliers were less than 5% in each grouping in the range 3–42 Å and up to 20% otherwise

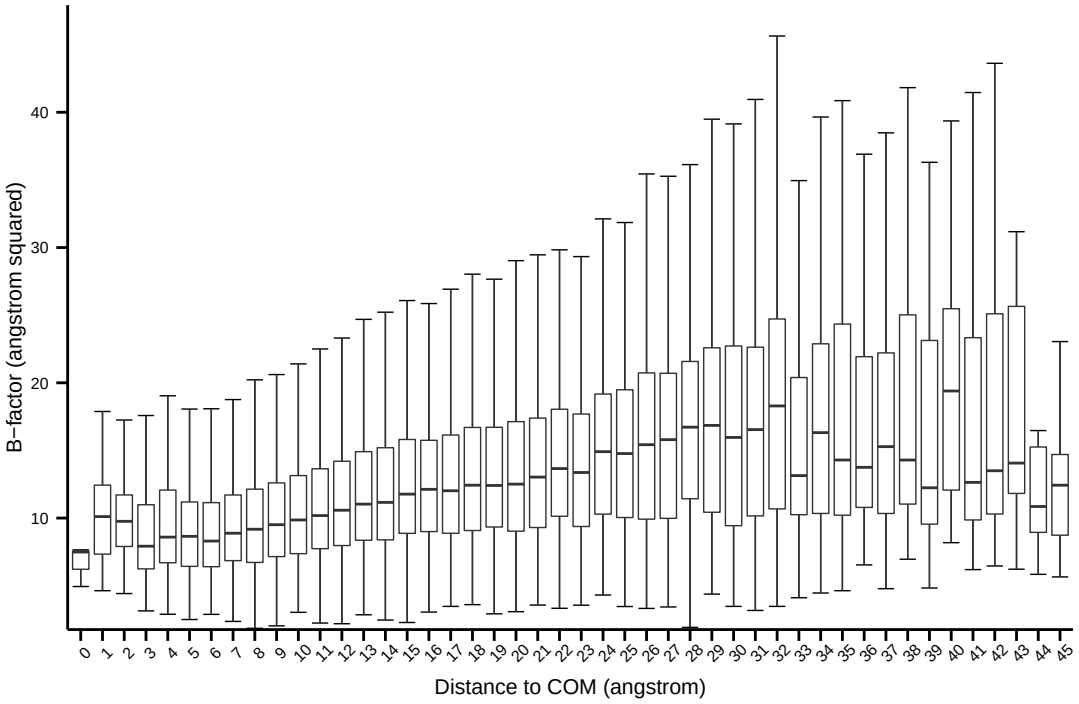
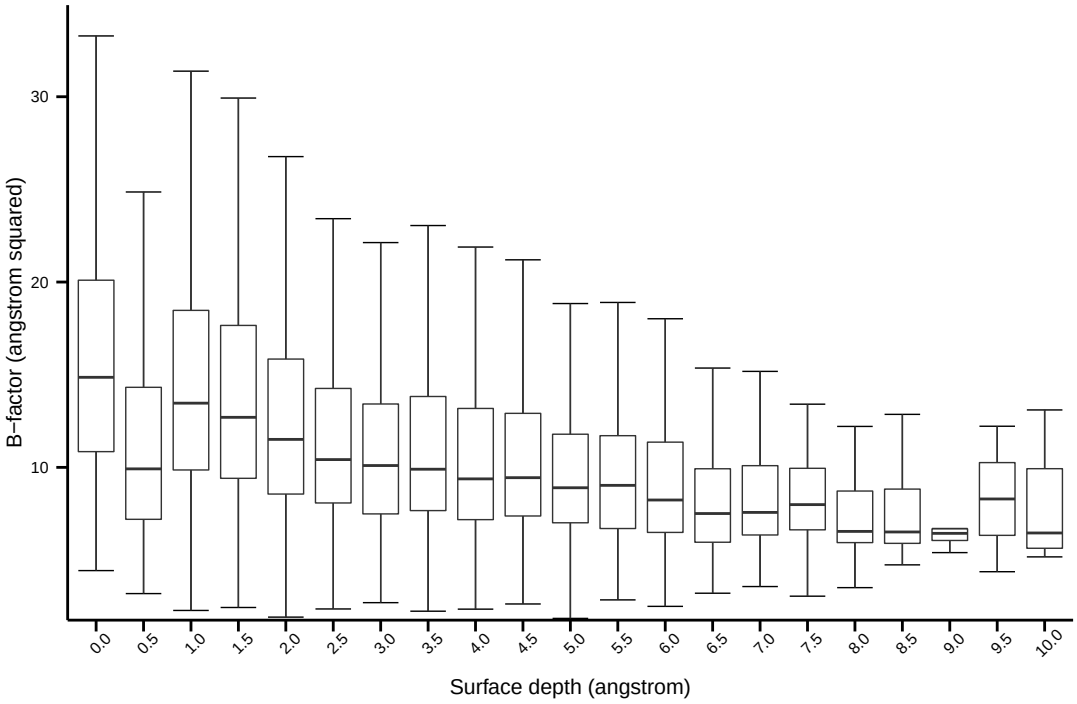


Figure 3.20: Boxplots of alpha-carbon B-factors grouped according to the atom's distance from the surface. The bin width is 0.5 Å. The proportions of all outliers were less than 5% in each grouping except ≥7.0 Å (5.6-16.7%).

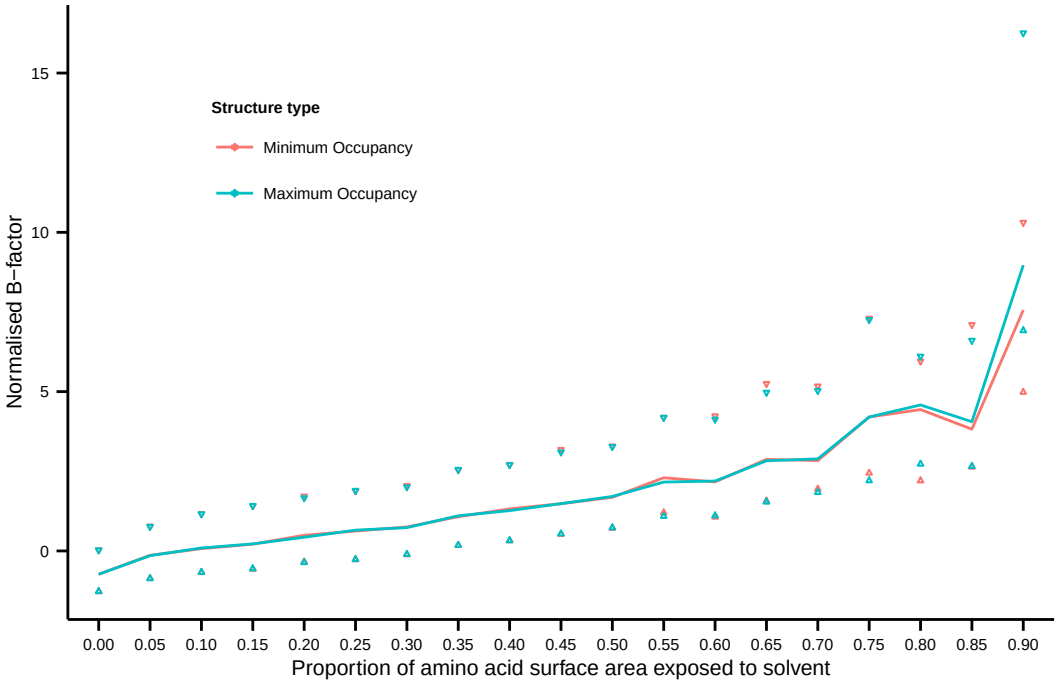


3.4.7 Effect of atom occupancy

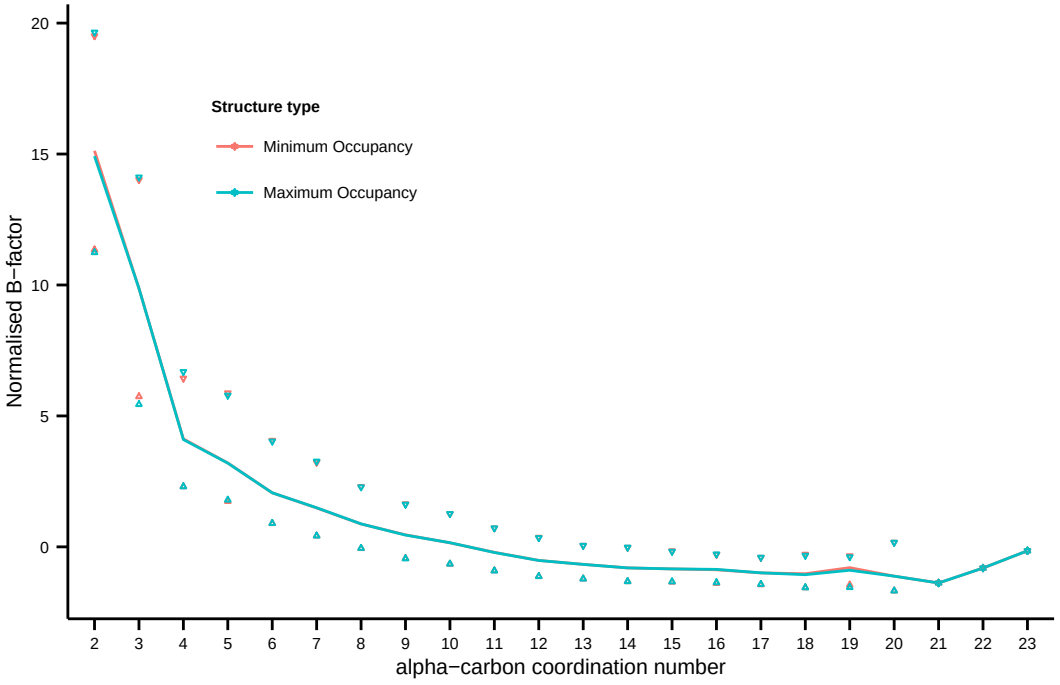
All the results described so far have been derived from structures where all the atoms are in their maximum occupancy positions. Interestingly, more than half of the structures in the data set have been resolved with atoms in alternative conformations. This led to the consideration of whether the results would be different if these other conformations were examined. The analysis was repeated using structures where all atoms were in their minimum occupancy positions. It was hoped that, by comparing these two "extremes" of conformation, any differences due to the alternate locations would be apparent. Surprisingly, the analysis of the minimum occupancy structures generated results that were almost identical to those of the maximum occupancy structures. Figures 3.21a and 3.21b compare the results obtained from the minimum and maximum occupancy structures with respect to SASA and alpha-carbon coordination number.

The almost identical results obtained could be explained by the observation that, for the majority of protein structures resolved with atoms in alternate locations, only a small proportion of the overall structure is affected. Of the 77 proteins in more than one conformation, the conformational variation in 55 (71%) of these proteins occurs in fewer than 10% of their amino acids (see table 3.1). Nevertheless, even small changes to the orientations of a few side chains at the protein's surface could change which residues are exposed to or shielded from the solvent. This explains the minor discrepancies between the minimum and maximum occupancy structures in the analysis of SASA (figure 3.21a) compared to the almost complete agreement in the analysis of coordination number (figure 3.21b). It is not surprising that the calculation of SASA is more sensitive to small variations in atom coordinates, particularly when the atoms affected are likely to be the highly mobile side chains of residues at or near to the protein's surface.

Figure 3.21: Comparison of the results of B-factor analysis when minimum and maximum occupancy structures were used. The median B-factor calculated over each interval is plotted as a line. The lower and upper quartiles are represented by upward and downward pointing triangles.



(a) B-factors grouped according to amino acid SASA using the same intervals as figure 3.10



(b) B-factors grouped according to alpha-carbon coordination number using the same intervals as figure 3.15

3.4.8 Combining structural properties

The results described above show that there are weak correlations between B-factor values and structural properties that are likely to influence the conformational variability. The relationships observed, however, are only very general and the trends cannot be easily quantified. The high degree of variability inherent in the data set is apparent from the broad interquartile ranges of the boxplots used to visualise the data (figures 3.8a to 3.15). It was speculated that it might be possible to establish a more convincing connection between B-factor values and conformational dynamics if, instead of considering each structural property in isolation, the effect of all the properties in combination could be investigated. The assumption that each structural property in isolation would give some indication of the underlying dynamics of the protein may have been too simplistic. It is, perhaps, more likely that the dynamics of a protein, and consequently atom B-factor values, are influenced by a complex interplay between different structural factors.

Support Vector Machines (SVM) were employed to establish whether B-factor values could be predicted from all the protein structural properties in combination. SVM are a machine learning technique that can be used for both regression and classification tasks. Initially, SVM were used for regression to predict alpha-carbon B-factor values given the amino acid type; secondary structure; normalised SASA of the amino acid; distance to the protein surface; distance to the protein's COM; and coordination number within an 8Å radius for the atom. The regression model was derived by training on a random selection of 75% of proteins from the data set. The model was then tested by comparing the model's predictions to the experimentally determined B-factors for the remaining 25% of the proteins. The accuracy of the SVM model was quantified by calculating the Pearson and Spearman correlation coefficients between the predicted B-factor values and the experimental data. The process of training and testing the SVM model was repeated five times using different randomly selected sets of proteins to ensure that the results were not dependent on the choice of proteins used for training or testing. The results are presented in table 3.8 for raw B-factor values and B-factors normalised by the mean-standard deviation and median-mad methods.

Table 3.8: Correlation coefficients for five independent randomised SVM regression analyses of the isotropic B-factor data set

Correlation coefficients						
No Normalisation		**Mean-SD**		**Median-MAD**		
Pearson	**Spearman**	**Pearson**	**Spearman**	**Pearson**	**Spearman**	
0.378	0.367	0.594	0.650	0.590	0.619	
0.401	0.411	0.609	0.659	0.617	0.618	
0.289	0.302	0.573	0.632	0.624	0.651	
0.371	0.392	0.574	0.645	0.599	0.616	
0.281	0.282	0.568	0.644	0.574	0.610	
0.34 ± 0.06	0.35 ± 0.06	0.58 ± 0.02	0.65 ± 0.01	0.60 ± 0.02	0.62 ± 0.02	(mean ± SD)

The most striking feature in table 3.8 is the difference between the results for raw B-factors and B-factors that have been normalised. The benefit of applying normalisation is apparent from the near doubling of the correlation coefficients when either mean-standard deviation or median-mad normalisation is used. The difference between the correlation coefficients for the normalised and raw B-factor data is statistically significant. Testing whether the distributions of correlation coefficients were equal using the Mann-Whitney test gave $p < 0.01$ for both Pearson and Spearman coefficients. Furthermore, regression performed equally well irrespective of the choice of normalisation method. There were no statistically significant differences between the distributions of correlation coefficients when comparing the results derived from median-mad and mean-standard deviation normalised B-factors (Mann-Whitney tests at $p > 0.05$).

The SVM analysis was repeated using only alpha-carbon coordination number as the independent variable. The results are given in table 3.9 and, interestingly, the accuracy of the predictions appear to be close to that obtained when all the structural properties were used in the analysis. These results suggest that, of all the structural properties considered, atom packing density might have the greatest influence on determining B-factor values. An attempt was made to find a minimal subset of structural properties that could be used for B-factor predictions. Whilst the search was not exhaustive, alpha-carbon coordination numbers always appeared to be the dominant variable.

Table 3.9: Correlation coefficients for five independent randomised SVM regression analyses of median-mad normalised B-factors where the alpha-carbon coordination number is the only independent variable.

Correlation coefficients	
Pearson	**Spearman**
0.522	0.519
0.449	0.492
0.507	0.511
0.479	0.513
0.498	0.521
0.49 ± 0.03	0.51 ± 0.01

(mean ± SD)

Although there appears to be some relationship between B-factor values and static structural properties of proteins, SVM regression cannot predict B-factor values with a high level of precision. Subsequently, it was considered whether SVM might be more successful when applied to the easier problem of classifying atoms as being either "flexible" or "rigid". The classification analysis was run identically to the regression analysis with the exception that the alpha-carbon B-factor values were replaced with a label of "flexible" or "rigid". Labels were assigned by inspecting the normalised B-factor values and assigning the label of "flexible" or "rigid" depending on whether the values were greater than or less than zero. A normalised B-factor greater than zero indicated that an atom had a higher B-factor than

Table 3.10: Percentage of atoms correctly classified as being "flexible" or "rigid" in five independent randomised SVM classification analyses of the median-mad normalised B-factor data set.

Choice of independent variables	
All structural properties	**Alpha-carbon coordination number**
73.8%	69.3%
72.2%	67.8%
72.5%	69.1%
72.8%	67.1%
73.5%	69.0%
73 ± 0.7%	68 ± 1.0%

(mean ± SD)

a typical atom of the structure and, thus, could be considered to be more "flexible". Conversely, a normalised B-factor less than or equal to zero was indicative of a more "rigid" atom. The results of five independent rounds of SVM classification are given in table 3.10 which presents the percentage of atoms in test sets whose labels were correctly predicted. The table also includes the results when the SVM classification task was repeated using alpha-carbon coordination number as the only independent variable. The results of the classification tasks in table 3.10 are not a great improvement over the results of the regression analysis. Irrespective of the choice of independent variables, the SVM classifiers can only predict "flexible" or "rigid" atoms with an accuracy of around 70% which means that there is a significant proportion of atoms whose B-factors are not consistent with what might be expected given the proteins' structures.

3.5 Methods

3.5.1 Deriving the protein data set

The set of protein X-rays structures used in the analysis was derived in a two stage process. The first step queried the PDB using the RESTful web services of the Research Collaboratory for Structural Bioinformatics (RCSB) website (September 2014). A query was submitted in the form of XML that requested the PDB identifiers, resolutions and R-factors of all experimental X-ray structures resolved to a maximum resolution of 1.5 Å. In addition, the query limited the search to single chain proteins at least 50 residues in length and sharing no more than 30% sequence homology. These criteria produced a list of 1591 candidate proteins.

The second processing step screened the candidate proteins using more rigorous criteria than were available through the web services. All structures were downloaded from the PDB FTP repository and each PDB file was parsed to analyse the protein structure and the details of the crystallographic experiment. Structures were excluded if any of the reported R-factors were greater than 0.25 or if there were more than three consecutive unresolved residues. Any inconsistencies between the sequence and structure of the protein, as specified by the `SEQRES` and `ATOM` records, resulted in exclusion and only MSE was accepted as a modified residue. Proteins complexed with large cofactors or ligands (defined as molecules with more than 10 resolved atoms) were also eliminated. Membrane proteins were discarded, as far as was possible, on the basis of whether the PDB file contained the text "membrane" or "channel" in the title or keyword meta-data. Structures that included any anisotropic displacement data (`ANISOU` records) were also eliminated from this study to ensure that the structures had been refined with isotropic B-factors.

3.5.2 Structural calculations

Minimum and maximum occupancy structures were derived for each protein in the data set following the procedure described in the methods chapter. Each structure underwent two rounds of processing. The first stage ran calculations on the asymmetric units, treating them as isolated proteins. The second stage of processing reconstructed each protein crystal's periodic unit cell to recalculate properties such as SASA, coordination number and surface depth that are affected by the structure of the crystal lattice. The results of all the calculations were stored in a HyperSQL database (HSQL Development Group 2012) allowing for efficient querying and reformatting of the data for subsequent statistical analysis and visualisation. The structural processing software was implemented as a multi-threaded Java application in order to make efficient use of computational resources and limit processing bottlenecks.

The results of the calculations were analysed using the scripts written in the programming

language R (R Development Core Team 2008) which is optimised for statistical computation and data visualisation. A Java program was developed to run queries against the database and to output the results in a format that could be read by the R scripts. The program also applied all post-processing operations to the data such as the normalisation and mathematical transformation of B-factor values.

3.5.3 B-factor normalisation methods

The reasoning behind normalisation is that it might correct for any variation in B-factor values arising from the different conditions under which structures are determined. Therefore, normalisation was applied to each structure individually and no attempt was made to normalise the B-factors across all the proteins collectively. All normalisation techniques rely on atoms being grouped into of sets the same "type". Atoms were classified according to their type as defined by the names of the atoms in the `ATOM` records of the PDB file. Branch digits were ignored to ensure that similar atoms were made equivalent; for example, the two gamma carbon atoms of valine (`CG1` and `CG2`) were both considered to be the same. Care was taken to account for white space in the atom names as not to confuse "`SE_`" (selenium) with "`_SE`" (an epsilon sulphur).

Atoms where the coordinates had been assigned by the modelling software were not included in the normalisation calculations. Atoms were also excluded if their B-factors could not be normalised i.e., if all the atoms of a particular type had equal B-factors or where there was only one atom of a particular type in the structure.

Mean-standard deviation ("z") normalisation

The mean and population standard deviation were calculated for each atom type in the structure and used to normalise the B-factors following the method of Carugo and Argos (1997). For example, given a protein structure containing atoms of type X, the normalised B-factors $B_{X,i}^{norm}$ of the ith atom of this type was calculated as:

$$B_{X,i}^{norm} = \frac{B_{X,i} - < B_X >}{\sigma_{pop}(B_X)} \tag{3.3}$$

where $B_{X,i}$ is the raw B-factor and $< B_X >$ and $\sigma_{pop}(B_X)$ are the B-factor mean and population standard deviation for all the atoms of type X in the structure.

Median-mad normalisation

Median-mad normalisation is similar to mean-standard deviation normalisation but uses the more robust statistics of the median and MAD as the measures of location and spread. The normalised B-factors $B_{X,i}^{norm}$ of the ith atom of type X in a structure was calculated as:

$$B_{X,i}^{norm} = \frac{B_{X,i} - \text{median}(B_X)}{\text{mad}(B_X)} \tag{3.4}$$

where $B_{X,i}$ is the raw B-factor and $\text{median}(B_X)$ and $\text{mad}(B_X)$ are the B-factor median and MAD for all the atoms of type X in the structure.

Minimum-maximum normalisation

Unlike both the previous methods, minimum-maximum normalisation ensures that the B-factors for all structures lie within the same interval. A linear scaling is applied so that the lowest B-factor in the structure maps to a value of zero and the highest to one. This approach is similar to the unitary normalisation employed by Schneider *et al.* (2014) who scaled the B-factors of DNA and protein complexes to values in the range 1-100. The normalised B-factors $B_{X,i}^{norm}$ of the ith atom of type X in a structure were therefore calculated as:

$$B_{X,i}^{norm} = \frac{B_{X,i} - \min(B_X)}{\max(B_X) - \min(B_X)} \tag{3.5}$$

where $B_{X,i}$ is the raw B-factor and $\min(B_X)$ and $\max(B_X)$ are the minimum and maximum B-factors values for the atoms of type X in the structure.

3.5.4 Machine learning using support vector machines

The implementation of the SVM regression and classification algorithms were provided by the `e1071` R package (Meyer *et al.* 2014) that incorporates the `LIBSVM` SVM library developed by Chang and Lin (2011). The alpha-carbon B-factor data set required preprocessing to convert the two non-numeric discrete variables (amino acid type and secondary structure classification) into a suitable input format. Both variables were converted to binary vectors; for example, the amino acid types of the alpha-carbon atoms were converted to twenty element binary vectors. Each element of the vector took a value of either one ("true") or zero ("false") to indicate which of the twenty standard amino acid types was assigned to the atom. Thus, for each vector, only one element could take a value of one while all the other elements were set to zero. Secondary structure classification was treated in exactly the same way using eight element binary vectors to represent the seven DSSP secondary structure types plus an unclassifiable/"random coil" category. The classification task applied an additional

preprocessing step which replaced the B-factor values of the atoms with labels of "rigid" or "flexible" . B-factor values were first normalised by the median-mad method described above and then each atom was assigned one of the two labels. Atoms with normalised B-factors greater than zero were labelled as being "flexible" while those with normalised B-factors equal to or less than zero were labelled as "rigid".

The `tune.svm` function of the `e1071` package was used to find suitable values to use for the SVM gamma and cost parameters. SVM were tuned using values for gamma and cost over the range 10^n for $n = -4, -3, \ldots, 3, 4$ for a small number of proteins. This located an optimal value for gamma within the range 0.001 to 0.1 and costs from 10 to 1000. Tuning was then repeated using ten randomly selected proteins. The values of the cost parameter were set to 0.1, 10, 100 or 1000 and the gamma values to 10^{-n} for $n = 2, 3$ and 4. This process was repeated five times to determine suitable values for the cost and gamma parameters. The values chosen for the cost and gamma varied depending on the data set used. For the raw B-factors, the optimal value for gamma was 0.01 and 10 for the cost. For median-mad normalised data the gamma value was 0.001 with a cost of 1000. In the case of mean-standard deviation normalisation, a gamma value of 0.01 and a cost of 100 was chosen. Nonetheless, the choice of these parameters was not clear-cut, gamma values from 0.001 to 0.01 and costs 10 to 1000 all scored similarly under tuning.

The SVM models for both B-factor regression and classification tasks was derived using the `svm.model` function of the `e1071` package. Default parameters were chosen with the exception of the cost and gamma parameters and the option to normalise/scale the input data when deriving the models. Scaling was not applicable to the binary vectors representing amino acid type and secondary structure classification. The SVM models were derived by training on a random selection of 75% of the proteins in the data set. Training with fewer proteins gave models that performed poorly while increasing the size of the training set significantly increased the computational time without any improvement in the accuracy of the predictions. The choice of whether the SVM model generated could be used for regression or classification was determined automatically by the `svm.model` function. Supplying numeric B-factors produced a SVM model for regression and replacing the B-factor values with labels gave a SVM model that could be used for classification.

Chapter 4

Evaluating anisotropic B-factors as indicators of a protein's conformational dynamics

4.1 Introduction

From the work described in chapter 3 of this thesis, it is not possible to establish a clear relationship between isotropic B-factors and structural properties expected to correlate with conformational flexibility. A plausible explanation could be that the isotropic model is an inadequate description of the fluctuations of crystallographically equivalent atoms. In contrast, the alternative anisotropic model may characterise the movements of atoms within a crystal structure more realistically. Furthermore, since the anisotropic refinement of a structure requires good quality crystallographic data at a very high resolution, AADPs may be less susceptible to distortion from model error than isotropic B-factors. Therefore, the analysis of chapter 3 was repeated using only high resolution protein structures that had been refined anisotropically. Unlike isotropic B-factors, there are no examples in the literature of recent analyses where AADP data has been related to static structural properties of proteins. Most work relating to AADPs is in the context of validating theoretical models of protein dynamics and, in particular, the harmonic atomic oscillations predicted by elastic network models (Eyal *et al.* 2007; Kondrashov *et al.* 2007; Hafner and Zheng 2011). The motivation for a classical analysis of AADPs was not only to fill a gap in the literature, but to investigate whether AADPs might be more insightful and informative indicators of conformational flexibility in protein crystals than isotropic B-factors.

Unlike the isotropic model, which is parametrised by a single variable, the anisotropic model uses six parameters to characterise the fluctuations of atoms. Consequently, it is not possible

to make a direct one-to-one comparison between isotropic B-factors and AADPs. Nonetheless, of the the six AADPs, three of these variables represent mean-square displacements analogous to an isotropic B-factor.

4.2 Aim

The overarching aim of this study is to establish whether the conformational variability in a protein crystal can be more accurately described by AADPs than isotropic B-factors. A set of high resolution protein crystal structures resolved with AADPs will be analysed to investigate whether there are any relationships between an atom's AADPs and the environment in which it is located in the crystal. The influence of amino acid type; secondary structure; depth from the surface; exposure to the solvent and atom packing density will be considered. The results from chapter 3 supported the assertion that these static structural properties were correlates of conformational variability even though no quantitative relationships with B-factor values could be formulated. It is therefore reasonable to assume that, if AADPs are a more accurately model for the atomic displacements in a crystal, it will be easier to observe the effect of protein crystal structure on conformational dynamics by analysing AADPs rather than isotropic B-factors.

The analysis of AADPs will focus on measures of atomic mean-square displacements rather than the variables that define the orientations of these movements. Although the asymmetry in the direction of atomic movements is a significant feature of the anisotropic model, the extent to which an atom moves is more important when considering conformational flexibility. Moreover, when considering a structurally diverse collection of protein crystals, it would be very difficult to determine how every structure should be orientated so that the directions of atomic fluctuations could be meaningfully compared. The AADPs considered in this study are:

- The three eigenvalues of the covariance matrices modelling the anisotropic movements of atoms.
- The anisotropy ratio.
- The equivalent isotropic B-factor.
- The anisotropic volume.

The values of the three eigenvalues are equal to mean-square displacements of the atoms. Specifically, the largest and smallest eigenvalues correspond to the mean-square deviations in the directions of maximal and minimal displacement respectively. Therefore, a secondary aim of this study will be to establish whether all three eigenvalues should be considered when quantifying conformational dynamics, or whether just one value is sufficient; for example,

the eigenvalue corresponding to the maximal deviations of an atom. In addition, three measures derived from the eigenvalues will also be considered: the anisotropy ratio; the "equivalent" isotropic B-factor and the anisotropic volume. Anisotropy is the ratio of the smallest eigenvalue to the largest (equation 1.2) and measures the degree of asymmetry in the movements of the atoms. The equivalent isotropic B-factor is derived from the mean of the three eigenvalues (equation 1.4) while the anisotropic volume is the product of the square roots of the three eigenvalues. Geometrically, the anisotropic volume is proportional to the volume of the region of space where an atom is expected to found at any given level of probability.

4.3 Hypothesis

The ADPs of protein structures refined anisotropically are more representative of the protein's dynamics than the ADPs of structures refined isotropically. Analysing the relationships between AADPs and the structural properties that influence protein conformational variability should yield better quality results compared to isotropic B-factors.

4.4 Results and discussion

4.4.1 Deriving the protein data set

A set of high resolution protein crystal structures was derived following exactly the same procedure that was used to obtain the protein data set described in chapter 3. The only difference in methodology was the selection of structures that had been refined using an anisotropic model for atomic fluctuations. Anisotropically refined structures were selected by inspecting the atomic coordinate records of the PDB data files and excluding structures that did not publish any AADP data. The higher resolution, and greater quality of the structures refined anisotropically, allowed for stricter criteria when selecting structures to include in the data set. The minimum resolution for the structures was reduced from 1.5 Å to 1.2 Å and the maximum R indices from 0.25 to 0.2.

The initial query submitted to the PDB generated a list of 491 candidate proteins that had been resolved to a resolution of 1.2 Å or higher. Subsequent filtering by the parsing of PDB data files resulted in a final data set of 120 proteins which are summarised in 4.1. Comparing table 4.1 to the summary data for the isotropic data set (table 3.1) reveals some notable similarities and differences. Both sets of proteins have similar compositions in terms of crystal structure with the three dominant space groups, P 2_1 2_1 2_1, P 1 2_1 1 and C 1 2 1, occurring with approximately the same frequency. There are some significant differences in terms of structure between the two sets of protein. The isotropic data set contains proteins

that are, on average, longer by 20 amino acids and there is a higher proportion of all-alpha proteins and mixed alpha-beta proteins. Moreover, the anisotropic data has a higher proportion of proteins that cannot be easily categorised with respect to secondary structure composition. As would be expected, the structures of the anisotropically resolved proteins are at the upper limits of X-ray resolution with approximately a third of all structures resolved at the sub-angstrom scale. As a direct consequence of the improved resolution, many more residues in the anisotropically refined structures are modelled with alternate conformations. Approximately 90% of all the anisotropically refined proteins are modelled with alternate conformations in contrast to just under 60% for the isotropically refined structures. Most strikingly, 17.5% of the proteins in the anisotropic data set have 20% or more of their amino acids in more than one conformation while there are no isotropically refined structures that have a similarly high proportion of residues in alternate conformations.

Table 4.1: Summary of the protein structures resolved anisotropically.

Feature	Number of proteins	% of data set
Chain length [1]		
$51 - 99$	30	25.0
$100 - 299$	80	66.7
$300 - 532$	10	8.3
Resolution		
< 1.0Å	41	34.2
$1.0 - 1.2$Å	79	65.8
Space Group		
P 2_1 2_1 2_1	42	35.0
P 1 2_1 1	21	17.5
C 1 2 1	11	9.2
23 other space groups [2]	46	38.3
Structural Classification [3]		
all-α ($> 60\%$ α and $< 5\%$ β)	8	6.7
mostly α-helix ($> 60\%$ α and $> 5\%$ β)	1	0.8
all-β ($> 50\%$ β and $< 5\%$ α)	8	6.7
mostly β-structure ($> 50\%$ β and $> 5\%$ α)	1	0.8
$\alpha\beta$ proteins ($15 - 55\%$ α and $10 - 45\%$ β)	54	45.0
others	48	40.0
Alternate conformations [4]		
0%	11	9.2
$0 - 10\%$	55	45.8
$10 - 20\%$	33	27.5
$\geq 20\%$	21	17.5

1 Median length 146.5. The minimum and maximum are 51 and 532 respectively.
2 Twelve space groups are represented by a single structure.
3 Using the domain structural classification of Michie *et al.* (1996).
4 Measured as the proportion of amino acids resolved with alternate conformations. The highest proportion is 27%.

4.4.2 Initial choice of anisotropic atomic displacement parameter for analysis

The analysis of AADPs began by concentrating on the largest eigenvalue of the AADP covariance matrix. As the largest eigenvalue, it was reasonable to assume that this parameter might be the most suitable to highlight differences between atoms. Furthermore, focusing the analysis on a single mean-square displacement parameter allowed for a direct comparison with the previous work on isotropic B-factors. To simplify the nomenclature used to discuss the AADPs analysed in this thesis, the following symbols will be used when referring to the eigenvalues derived from the AADP covariance matrix U^C (the `ANISOU` records of the PDB files):

- The largest eigenvalue: λ_{max}^{aniso}
- The "middle" eigenvalue: λ_{mid}^{aniso}
- The smallest eigenvalue: λ_{min}^{aniso}

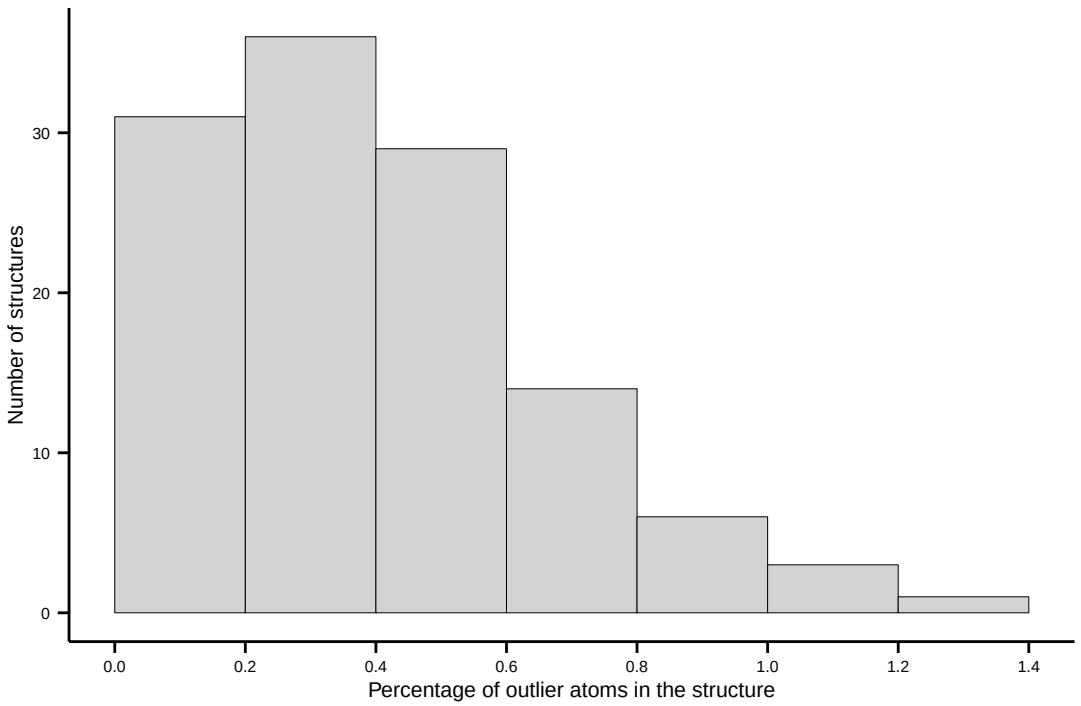
4.4.3 Assessing the quality of the data set

The quality of the structures in the data set was assessed following a similar procedure to that described in chapter 3 for isotropic B-factors. The value of λ_{max}^{aniso} was calculated for every atom in the structure that had been refined with AADPs. The value of the atom's λ_{max}^{aniso} was compared with the values for all the atoms of the same type in a 5 Å radius. The atom was marked as an "outlier" if its λ_{max}^{aniso} differed from the mean value by more than three standard deviations. The percentage of all atoms with atypical "outlier" values for λ_{max}^{aniso} was calculated for each structure and the distribution across the data set is presented in figure 4.1. The maximum eigenvalue was chosen as the discriminatory variable when validating the structures because this quantity represents the maximum anisotropic displacements of atoms and would, therefore, reveal any discrepancies in the precision to which atoms in close proximity have been resolved. As might be expected from a data set of high quality protein structures, very few atom had been assigned atypical AADPs. The percentage of outlier atoms was less than 1% for the majority of the structures in the dataset. Hence, the data set of anisotropically refined structures is of comparable quality and self-consistency to the set of isotropically refined structures.

4.4.4 Normalisation of AADP data

The investigation of isotropic B-factors in chapter 3 suggested that normalisation of B-factors can eliminate some of the inconsistencies in the data. To test whether the same result held

Figure 4.1: Distribution of outliers across the protein data set using λ_{max}^{aniso} as the discriminatory anisotropic atomic displacement parameter.The horizontal axis is the percentage of atoms in a crystal structure that have "atypical" AADPs and is a continuous scale. The data is binned in intervals of 0.2%. For example, the first bar represents the number of structures in the data set with the smallest percentages of "atypical" AADPs (between 0 and 0.2% of all the atoms in these structures have "atypical" AADPs).



for AADP, the effectiveness of different normalisation methods were analysed for alpha-carbon λ_{max}^{aniso} data (tables 4.2 and 4.3). Unexpectedly, the calculations gave very similar results when compared to the isotropic B-factors (3.6). This is surprising because it suggests that the value of λ_{max}^{aniso} is no better than an isotropic B-factor at differentiating between atoms expected to have the greatest and least conformational freedom within the structure. Given the higher quality of anisotropically refined structures, and the more realistic model for atomic displacements, it might have been expected that an atom's λ_{max}^{aniso} value would be a far better indicator of its flexibility than the B-factor.

The similarities in the results also extends to the normalisation methods. As was seen with the isotropic B-factor data, applying median-mad or mean-standard deviation normalisation was beneficial. However, there is little to distinguish between the two normalisation methods. Therefore, in all subsequent work, median-mad normalisation was applied to AADPs values. Median-mad normalisation was selected over mean-standard deviation on the basis that it uses "robust" statistics that are less sensitive to atypical data values. The analysis of isotropic B-factors had showed that ADPs values can be highly variable with a high proportion of "outliers". It was hoped that, unlike mean-standard deviation normalisation, the median-mad method would be less prone to distortion.

Table 4.2: Comparing normalisation methods when applied to the largest eigenvalue of the AADP covariance matrix.

Atom subset criterion	**Transform**	**Normalisation**	**Score**
secondary structure	none	none	0.6545
		mean-sd	0.8358
extended β-structure: 3999 atoms		median-mad	0.8882
unclassified structure : 4908 atoms		min-max	0.8126
	square root	none	0.6773
		mean-sd	0.9136
		median-mad	0.9167
		min-max	0.8806
	natural log	none	0.7083
		mean-sd	0.9772
		median-mad	0.9628
		min-max	0.9430
solvent exposure	none	none	1.7698
		mean-sd	1.9990
SASA > 0.5 : 1178 atoms		median-mad	2.0731
SASA < 0.01 : 4585 atoms		min-max	1.7462
	square root	none	2.0191
		mean-sd	2.3959
		median-mad	2.4225
		min-max	2.0687
	natural log	none	2.3381
		mean-sd	2.8182
		median-mad	2.8320
		min-max	2.5268
coordination number	none	none	1.9191
		mean-sd	2.1322
≥ 15 : 1405 atoms		median-mad	2.0372
≤ 5 : 395 atoms		min-max	2.0588
	square root	none	2.2526
		mean-sd	2.6070
		median-mad	2.4278
		min-max	2.4929
	natural log	none	2.5940
		mean-sd	3.2298
		median-mad	2.8846
		min-max	3.0918
amino acid type	none	none	0.6899
		mean-sd	0.7616
Aspartate : 1219 atoms		median-mad	0.8386
Tryptophan : 283 atoms		min-max	0.6714
	square root	none	0.7175
		mean-sd	0.8139
		median-mad	0.8660
		min-max	0.7588
	natural log	none	0.7568
		mean-sd	0.8475
		median-mad	0.8848
		min-max	0.8087

Table 4.3: Comparing AADP normalisation methods with random selections of atoms. The random selection of 5000 atoms for the two groups was repeated 5 times. The score reported is the mean of these 5 samples and the associated standard deviation.

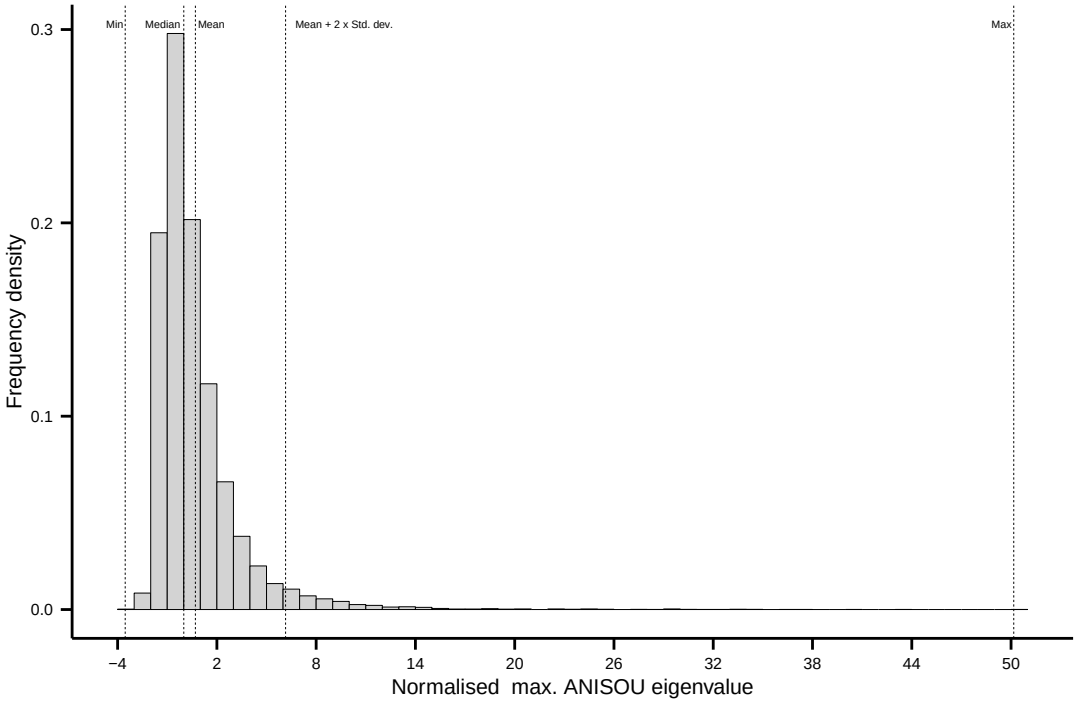
Transform	**Normalisation**	**Score (mean ± sd)**
none	none	0.0198 ± 0.0136
	mean-sd	0.0283 ± 0.0175
	median-mad	0.0135 ± 0.0146
	min-max	0.0154 ± 0.0140
square root	none	0.0445 ± 0.0261
	mean-sd	0.0256 ± 0.0290
	median-mad	0.0173 ± 0.0126
	min-max	0.0322 ± 0.0162
natural log	none	0.0338 ± 0.0161
	mean-sd	0.0490 ± 0.0224
	median-mad	0.0268 ± 0.0209
	min-max	0.0318 ± 0.0380

4.4.5 Distribution of alpha-carbon anisotropic atomic displacement parameters

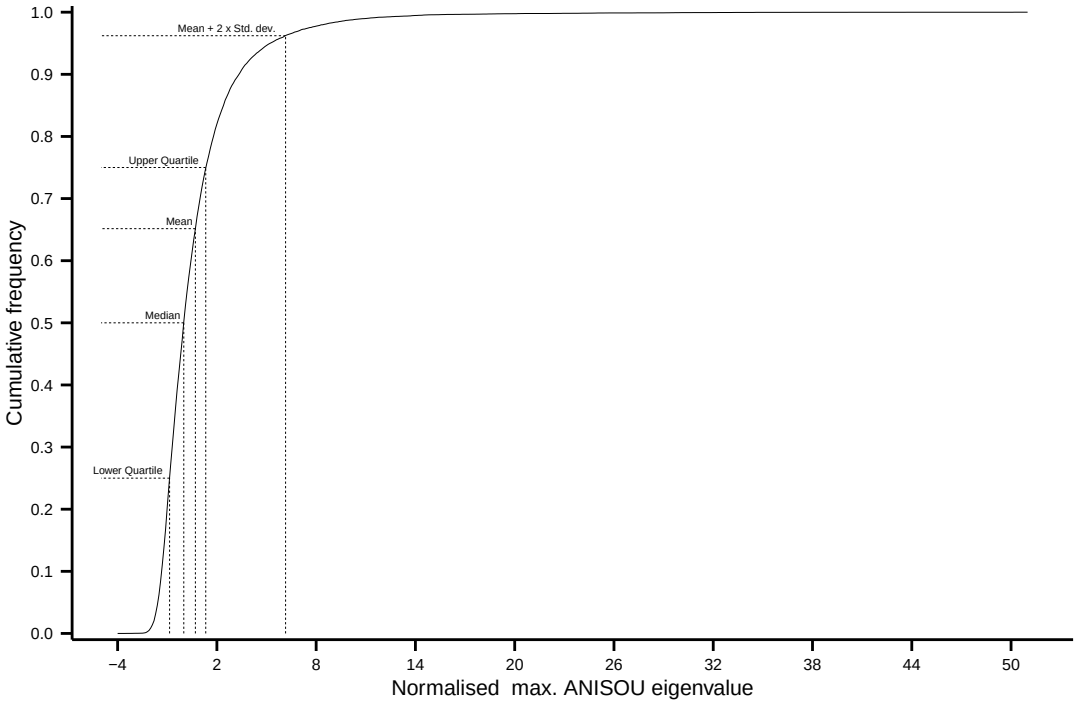
The distribution of the median-mad normalised λ_{max}^{aniso} for the alpha-carbons (figure 4.2) is highly positively skewed (skewness measure 4.443). The high skew of the distribution is partly a consequence of normalisation, which increased the skew from 3.935 for the non-normalised data. The skew measured is higher than that for the isotropic B-factor distribution calculated in chapter 3 (skewness 1.825 non-normalised and 3.021 median-mad normalised), and higher than the "equivalent" B-factors derived from the AADPs (skewness 3.752). The high skew might be explained by the anisotropic model. Atoms with highly asymmetric movements will have far larger λ_{max}^{aniso} values than atoms fluctuating more uniformly about their average positions. If the proportion of highly anisotropic atoms is small, then their high λ_{max}^{aniso} values will enhance the positive skew of the distribution. An attempt to reduce the skewness of the data was made by taking the square root of λ_{max}^{aniso} values before normalisation. Applying the square root transform to the data reduced the degree of skew, but the distribution still remained highly positively skewed (skewness measure 2.359). Furthermore, applying a natural logarithm transform also failed to eliminate the skew (skewness 1.231).

It could be argued that the source of the skew in the distribution is due to the anisotropic displacement data being "incomplete". PDB files are not guaranteed to include anisotropic displacement data for all the atoms in the structure. Furthermore, any atoms added by the modelling software would also be excluded from the analysis. Nonetheless, since the proportion of 'missing" AADP values is small, the data set is likely to be representative of the true distribution. There are 19717 alpha-carbon atoms in the data set of which only 207 (approximately 1%) lacked `ANISOU` PDB records. In addition, of these 207 atoms, 121 (58%) were missing `ANISOU` records due to being unresolved by crystallography. These atoms are likely to be located in the most mobile regions of a protein and would, consequently, be expected lie to the far right of the distribution. Therefore, it is unlikely that the exclusion of a small number of highly flexible atoms would result in significant changes to the shape of the distribution.

Figure 4.2: Distribution of normalised alpha-carbon λ_{max}^{aniso} values for the maximum occupancy structures.



(a) Histogram of the median-mad normalised maximum eigenvalue



(b) Cumulative frequency distribution for the median-mad normalised maximum eigenvalue

4.4.6 Relating AADPs to static structural properties of proteins

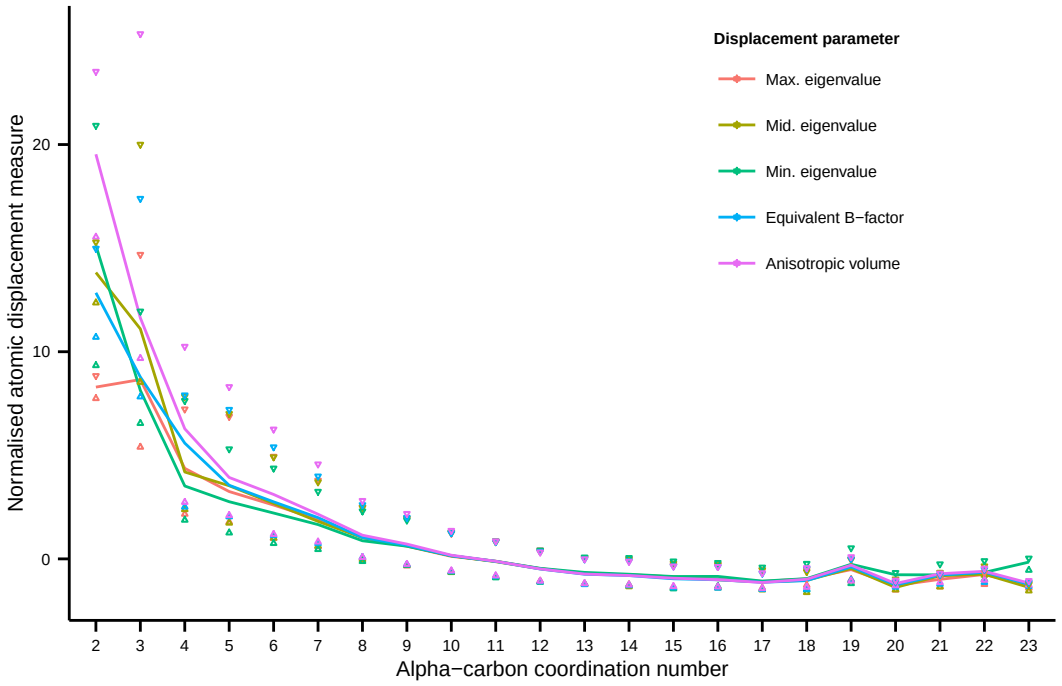
The methodology employed chapter 3 was repeated to generate boxplots relating median-mad normalised λ_{max}^{aniso} values to the static structural properties of protein crystals. The results of the analysis were very similar to those obtained for isotropic B-factors (see appendix A). Alpha-carbons with the highest normalised λ_{max}^{aniso} values were found in regions of the protein crystals that would be expected to have the greatest conformational freedom. Typically, atoms at or close to the surface of proteins; less densely packed together; and not held in extended secondary structure had high λ_{max}^{aniso} values.

The consistency in the results is reassuring and implies that the two data sets are representative of the proteins in general. The distributions of alpha-carbon to alpha-carbon distances in both data sets gave 8 Å as the distance between an alpha-carbon and its immediate neighbours. The occurrence of the same maximum alpha-carbon coordination number (23) and distance to the surface ($\approx$10 Å) in both data sets suggest that these may be structural limits for single chain proteins.

Despite the general consensus in the results, there were some unexpected differences. In the analysis of delta-carbons atom ADPs, amino acids classified with π-helical secondary structure had the highest median normalised λ_{max}^{aniso} values. In contrast, isotropic B-factors of π-helical delta-carbons atoms have one of the lowest median values. This is unlikely to be a statistical anomaly caused by small sample sizes as there are 103 and 85 π-helical delta-carbon atoms in the isotropic and anisotropic data sets respectively. There could, however, be differences between the proteins' tertiary structures that account for this disparity.

Another possibly significant difference can be seen with the variation in ADP values as the distance to the proteins' COMs increases. In the isotropic data set, the trend is an increase in B-factor value as the distance to the COM increases until a distance of approximately 20 Å is reached where the values begin to plateau. In the anisotropic data set, normalised λ_{max}^{aniso} also show the same trend, but after approximately 30 Å the boxplots statistics fluctuate wildly. It is possible that this can accounted for by differences between the structures of the proteins, especially since the proteins of the isotropic data set are, on average, larger than those of the anisotropic data set. In the isotropic data set there are 106 alpha-carbon atoms within the 35 Å bin and 28 at 40 Å. The sample sizes for the anisotropic data set are far smaller, with only 17 atoms at 35 Å and 6 at 40 Å. Hence, with fewer atoms to sample at distances beyond 30 Å, the boxplot statistics calculated for the anisotropic data set may be unrepresentative.

Figure 4.3: Plots of the median normalised alpha-carbon AADPs against amino acid coordination number. The lower and upper quartiles are plotted as upward and downward pointing triangles respectively. The five AADPs are: the maximum, middle and minimum anisotropic eigenvalues; the equivalent B-factor; and the anisotropic volume.

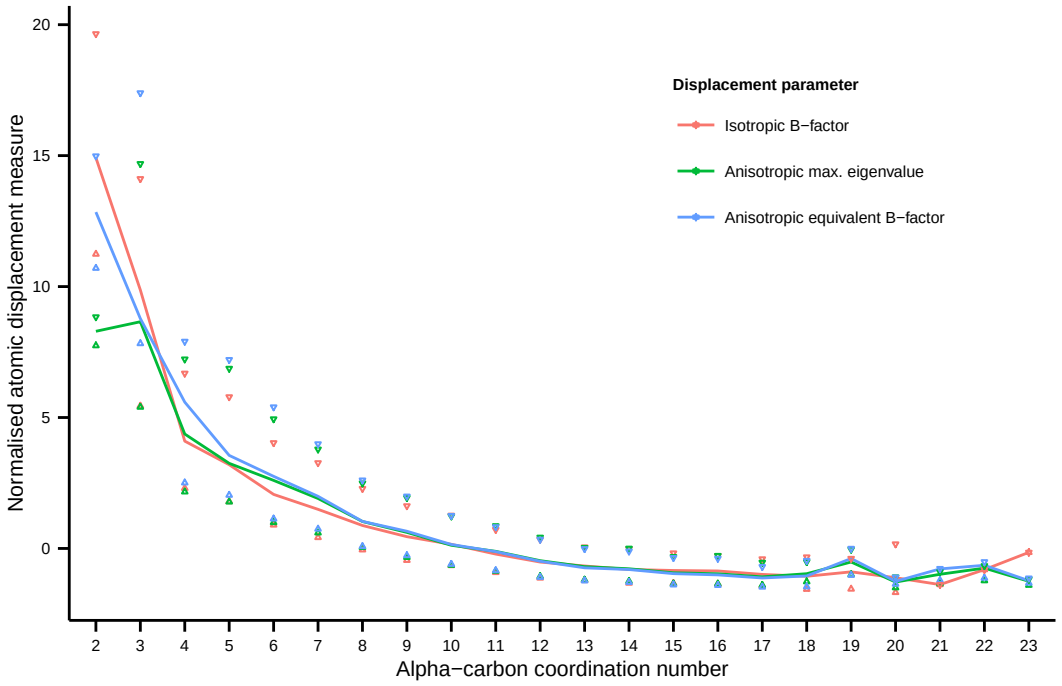


Consideration of other atomic displacement parameters

The analysis was repeated using different metrics for anisotropic atomic displacements. In addition to the eigenvalues λ_{mid}^{aniso} and λ_{min}^{aniso} , the "equivalent" B-factor and the anisotropic volume were also examined. Unexpectedly, the same trends were observed between the protein properties and the normalised AADPs irrespective of how the AADP was defined. As an example, figure 4.3 superimposes the different median-mad normalised AADPs profiles for coordination number. The most striking feature of these plots is the almost identical shape of the different graphs and the similar variability of the data. This result is surprising since, of all the AADPs considered, the maximum anisotropic mean-square displacement (λ_{max}^{aniso}) might have been expected to be a considerably "better" indicator of conformational variability.

Since there was nothing to distinguish between the different AADPs, it was natural to ask whether there was any difference between the AADPs and isotropic B-factors. Figure 4.4 compares the effect of alpha-carbon coordination number on isotropic and anisotropic ADPs. The median-mad normalised isotropic B-factors from the isotropically refined structures of chapter 3 are plotted against normalised λ_{max}^{aniso} and "equivalent" B-factors from the aniso-

Figure 4.4: Plots of the median normalised alpha-carbon AADPs against amino acid coordination number. The lower and upper quartiles are plotted as upward and downward pointing triangles respectively. The three ADPs are: the isotropic B-factor, maximum anisotropic eigenvalue; and the equivalent B-factor.

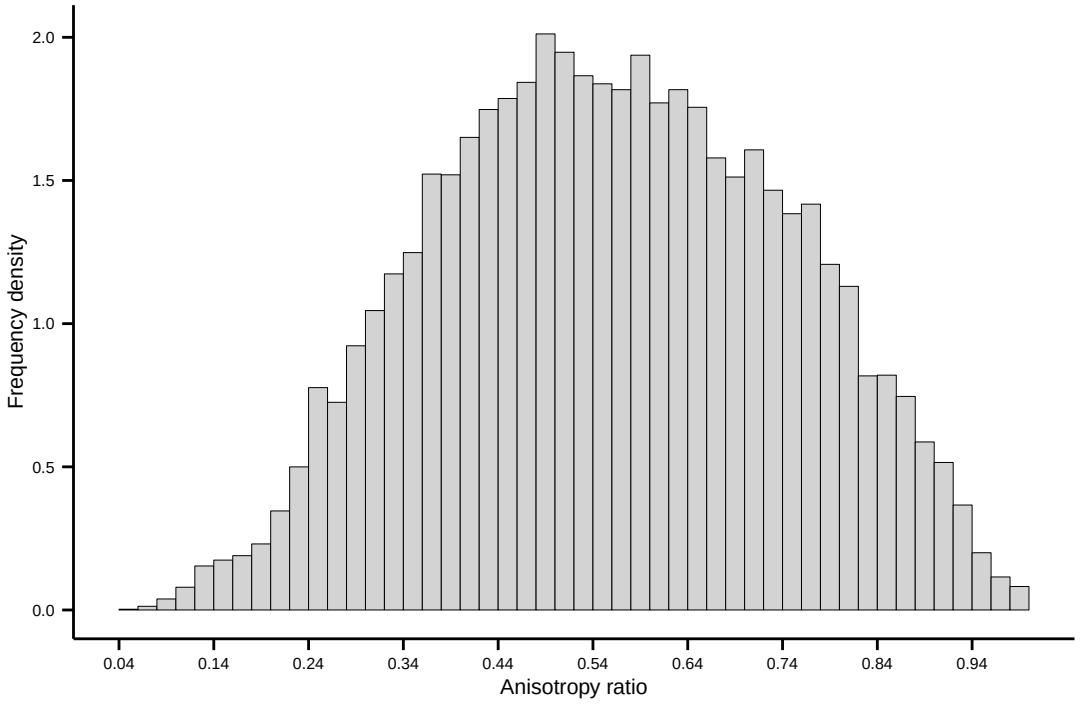


tropically refined structures. Not only are the median ADPs values almost identical, but there is a remarkable degree of consistency in the reduction of the spread of ADP values as coordination number increases. The graph suggests that both anisotropic and isotropic ADPs values reflect the reduction in conformational freedom caused by tighter amino acid packing. Furthermore, contrary to what might have been expected, there is no evidence that AADPs provide a more accurate measure of conformational flexibility than isotropic B-factors.

The simplest explanation to account for the high degree of similarity could be that the displacements of the majority of atoms are not particularly anisotropic. If the majority of atoms refined anisotropically are near-isotropic then this may account for the almost identical set of results. A plot of the distribution of anisotropy ratios (figure 4.5) clearly shows the opposite to be true. Atoms whose fluctuations are approximately isotropic (anisotropy ratios greater than 0.75) are a small proportion of the data set (17%) while the highly ellipsoidal atoms (anisotropy ratios less than 0.5) represent a significant sub-population (39%).

To assess whether the degree of anisotropy is a useful measure in itself, the analysis was repeated using the anisotropy ratio in place of ADPs. Unexpectedly, only extremely weak relationships were observed between the anisotropy ratios and the structural properties of

Figure 4.5: Histogram of the distribution of anisotropy ratios for the alpha-carbon of the maximum occupancy structures.



the proteins. Figure 4.6, relating anisotropy ratio to coordination number, is typical of the results obtained. Looking at the only the median values, there is a trend of the displacements becoming less ellipsoidal as coordination number increases. Presumably, this is due to the greater packing density restricting movement equally in all directions. Nevertheless, when the full distributions are considered, the alpha-carbon atoms are seen to exhibit a near full range of anisotropy ratios irrespective of the coordination number. It could be argued that only at the extremes does the packing density of the protein appear to have a significant influence on the direction of movement. High anisotropy at coordination number two (an unconstrained residue) might be explained by harmonic motion; for example, surface loops or the terminal strands swinging back and forth in the cavities of the lattice. Ellipsoidal displacements at very high packing densities are somewhat unexpected, but it might be a consequence of the atoms being so tightly packed that movement is only possible in certain directions.

A broad distribution of anisotropy ratios was also seen in the analysis by Eyal *et al.* (2007). Interestingly, their study demonstrated a decrease in anisotropy between the alpha-carbons of exposed (normalised SASA greater than 0.5) and buried (normalised SASA less than 0.1) residues. The corresponding results derived for this thesis are given in figure 4.7 and augments their work to show that, with the exception of the most exposed residues, small changes in SASA have negligible effects on anisotropy.

Figure 4.6: Boxplots of alpha-carbon anisotropy ratios grouped according to the coordination number of the amino acid. The proportion of outliers at values 20 and 21 are 25% and 20% respectively.

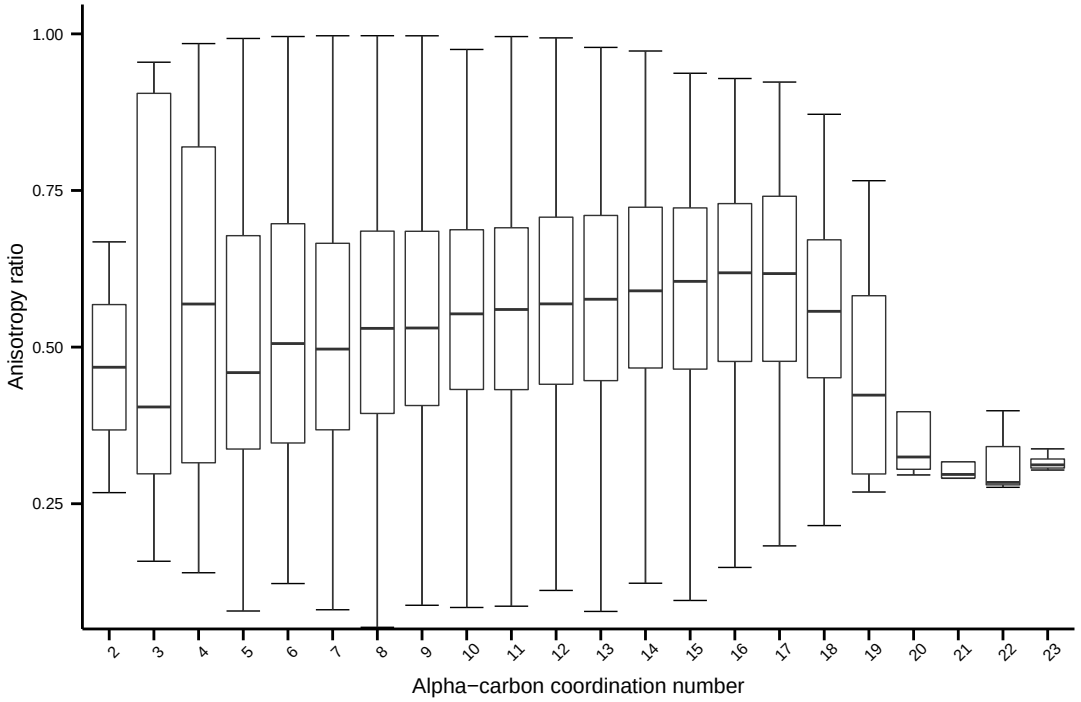


Figure 4.7: Boxplots of alpha-carbon anisotropy ratios grouped according to the coordination number of the amino acid. The proportions of outliers are less than 0.1% in all groupings.

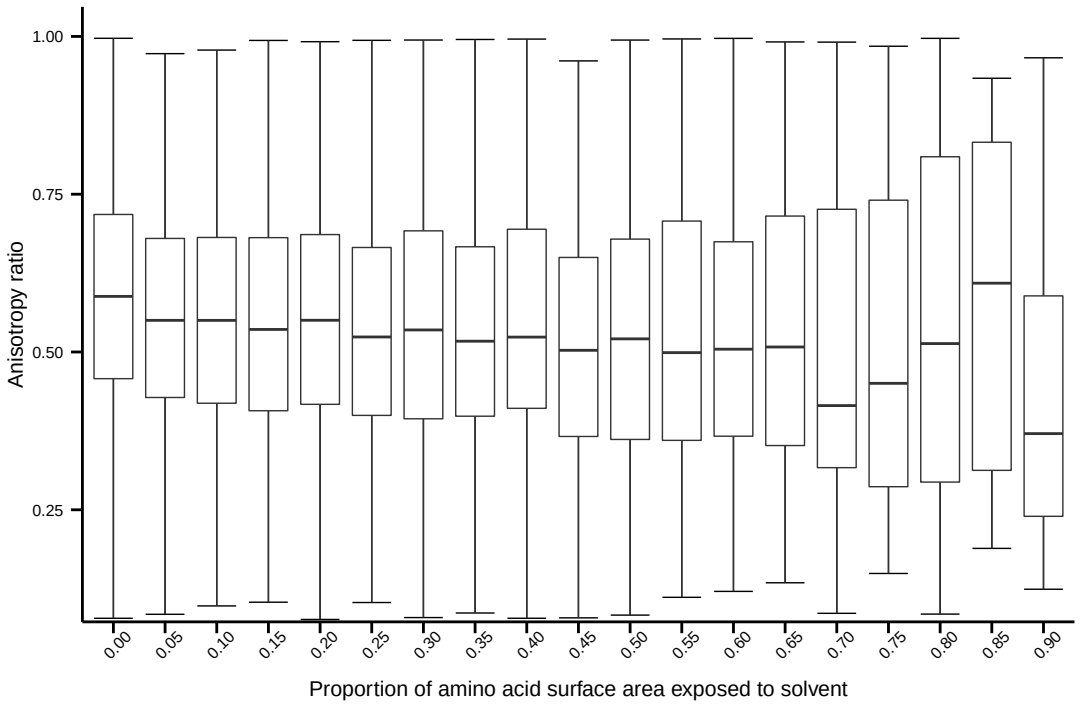
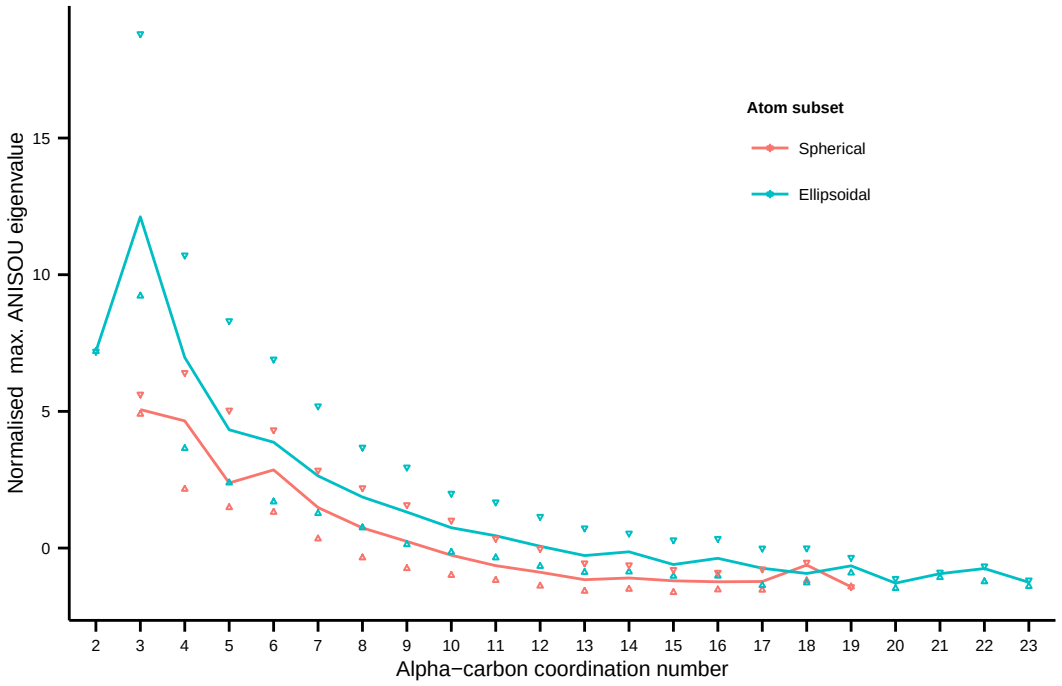


Figure 4.8: Plots of the median normalised maximum eigenvalues against coordination number for spherical and ellipsoidal atomic fluctuations.The lower and upper quartiles are plotted as upward and downward pointing triangles respectively. Spherical atomic fluctuations are defined as an anisotropy ratio greater than 0.75 and for ellipsoidal fluctuations the ratio is less than 0.5.



Although the anisotropy ratio was a poor indicator of conformational variability, it was investigated whether the ratio could be used in conjunction with AADPs. Specifically, to determine whether there was any difference between atoms with approximately spherical displacements and those whose displacements are highly ellipsoidal. The analysis of normalised λ_{max}^{aniso} was repeated ignoring all alpha-carbon atoms except for those with either the most spherical or ellipsoidal fluctuations (anisotropy ratios greater than 0.75 and less than 0.5 respectively). Figure 4.8 superimposes the results for the most spherical and ellipsoidal ADPs with respect to coordination number. Figure 4.8 is typical of the analysis as a whole in that the same trends are observed irrespective of the degree of anisotropy. There are, however, differences in the magnitudes of the λ_{max}^{aniso} for these two subsets of atoms. In general, atoms with the greatest displacements tend to be the most anisotropic.

4.4.7 Combining structural properties

The relationship between AADP values and the structural properties of proteins in combination was investigated with SVM following the same methods discussed in chapter 3. Regression and classification analyses were repeated using both λ_{max}^{aniso} and the equivalent B-factors as the dependent variables. Tables 4.4 and 4.5 are the results of the regression and classification experiments where all protein structural properties have been included as SVM variables. Mann-Whitney tests revealed no statistically significant differences between the Pearson correlation coefficients with λ_{max}^{aniso} and the equivalent B-factor as the AADP. The small improvement in Spearman correlation coefficients for the equivalent B-factors was, however, significant ($p < 0.05$). Classification failed to differentiate between λ_{max}^{aniso} and equivalent B-factors as the ADP that could be most accurately deduced from the static structural properties of a protein (Mann Whitney $p > 0.05$).

The Mann-Whitney tests were repeated to compare the results of regression and classification derived from the anisotropic equivalent B-factors and the median-mad normalised isotropic B-factors of chapter 3 (tables 3.8 and 3.10). There was no statistically significant differences between the Pearson correlation coefficients and results of the atom classification. The difference between the Spearman correlation coefficients were significant ($p < 0.05$). In terms of the hypothesis proposed at the beginning of the chapter, there is no compelling evidence to suggest that "better" predictions of an atom's conformational flexibility can be made with anisotropic atomic displacement parameters compared to isotropic B-factors.

Table 4.4: Correlation coefficients for five independent randomised SVM regression analyses of the anisotropic data set

Correlation coefficients				
Max. eigenvalue		**Equivalent B-factor**		
Pearson	**Spearman**	**Pearson**	**Spearman**	
0.507	0.642	0.613	0.690	
0.602	0.687	0.539	0.697	
0.555	0.644	0.584	0.700	
0.564	0.623	0.586	0.714	
0.531	0.643	0.560	0.686	
0.55 ± 0.04	0.65 ± 0.02	0.58 ± 0.03	0.70 ± 0.01	(mean ± SD)

Table 4.5: Percentages for correct classifications in five independent randomised SVM classification analyses of the anisotropic data set

Percentage correct classification	
Max. eigenvalue	**Equivalent B-factor**
72.6%	75.2%
75.9%	75.3%
72.8%	75.8%
74.8%	77.1%
75.0%	75.1%
74.2 ± 1.5%	75.7 ± 0.8%

(mean ± SD)

The SVM experiments were repeated using only the alpha-carbon coordination number as the independent variable to see if, like isotropic B-factors, AADP values are predominately determined by packing density. The results of the regression and classification analyses are presented in tables 4.6 and 4.7 respectively. Interestingly, the result are almost identical to those presented in chapter 3 for isotropic B-factors. Although not able to predict AADP as accurately as SVM incorporating all structural properties, analyses using only coordination numbers achieve reasonable results.

Table 4.6: Correlation coefficients for five independent randomised SVM regression analyses of the anisotropic data set where only the coordination number was used as an independent variable.

Correlation coefficients			
Max. eigenvalue		**Equivalent B-factor**	
Pearson	**Spearman**	**Pearson**	**Spearman**
0.444	0.498	0.549	0.564
0.483	0.542	0.504	0.554
0.480	0.520	0.527	0.556
0.504	0.516	0.480	0.577
0.458	0.514	0.495	0.561
0.47 ± 0.02	0.52 ± 0.02	0.51 ± 0.03	0.56 ± 0.01

(mean ± SD)

Table 4.7: Percentages for correct classifications in five independent randomised SVM classification analyses of the anisotropic data set where only the coordination number is the independent variable.

Percentage correct classification	
Max. eigenvalue	**Equivalent B-factor**
68.4%	70.5%
67.5%	70.0%
69.2%	68.6%
67.2%	70.8%
68.0%	70.7%
68.1 ± 0.8%	70.1 ± 0.9%

(mean ± SD)

4.5 Methods

4.5.1 Deriving the protein data set

The protein data set was derived following the same procedures described in the methods section of chapter 3 with a few minor adjustments. The initial query to the PDB only retrieved crystal structures of resolution no lower than 1.2 Å rather than the 1.5 Å limit used to derive the isotropic B-factor data set. The second processing that filtered the downloaded PDB data files reduced the maximum permissible R indices for the structures from 0.25 to 0.2. In addition, anistropically refined structures were selected by discarding all PDB files that had no `ANISOU` records associated with the atom coordinate data.

4.5.2 Processing anisotropic atomic displacement parameters

The six integer components: $U11$, $U22$, $U33$, $U12$, $U13$ and $U23$ that define the Cartesian anisotropic displacements of atoms were parsed from the `ANISOU` records of the PDB files. Each component was divided by 10^4 to convert to units of Å^2 and then assembled to give the covariance matrix U^Cas defined by equation 4.1. An example of an `ANISOU` record from the PDB data file for the structure 1BKR (Bañuelos *et al.* 1998) is given in figure 4.9 and equation 4.1.

Figure 4.9: An example of anisotropic atomic displacement parameters in a PDB data file. The `ATOM` and associated `ANISOU` record for the alpha carbon of the second residue of protein 1BKR (a domain from human beta-spectrin) (Bañuelos *et al.* 1998). The `ANISOU` record is annotated with labels to show the ordering of the elements of the covariance matrix U^C.

```
ATOM      2  CA  LYS A   2       9.296  31.931  19.151  1.00 14.98           C
ANISOU    2  CA  LYS A   2     2146   1770   1774   -296   -458   -279       C
                               U11    U22    U33    U12    U13    U23
```

$$U^C = \begin{pmatrix} U11 & U12 & U13 \\ U12 & U22 & U23 \\ U13 & U23 & U33 \end{pmatrix} = 10^{-4} \begin{pmatrix} 2146 & -296 & -458 \\ -296 & 1770 & -279 \\ -458 & -279 & 1774 \end{pmatrix} \quad (4.1)$$

$$= \begin{pmatrix} <(\Delta x)^2> & <\Delta x \Delta y> & <\Delta x \Delta z> \\ <\Delta x \Delta y> & <(\Delta y)^2> & <\Delta y \Delta z> \\ <\Delta x \Delta z> & <\Delta y \Delta z> & <(\Delta z)^2> \end{pmatrix} \text{ (compare with equation 1.2)}$$

All three eigenvalues of U^Cwere calculated and the eigenvalues sorted in descending order of magnitude. The first and last eigenvalues, corresponding to the maximum and minimum

mean-square displacements in the directions of the orthogonal axes, were used to calculate the anisotropy ratio as defined in equation 1.3.

Chapter 5

Simple models of atomic displacement in protein crystals

5.1 Introduction

The analysis of chapters 3 and 4 suggest that ADP values do reflect the conformational variability of atoms within protein crystal structures. However, only very general qualitative relationships could be established between ADPs and static structural properties of proteins. A possible explanation is that, despite using high quality crystallographic data, there was still a high level of "noise" in the ADP data. Atomic displacements may only be a partial determiner of ADP values and factors unrelated to conformational variability, such as crystal defects or experimental and modelling errors, may have a significant influence.

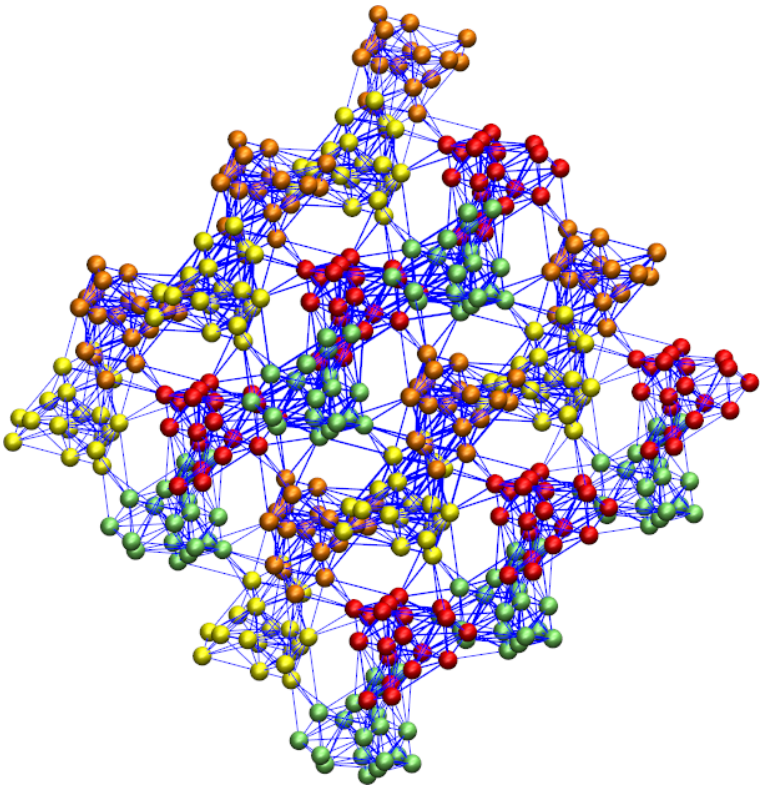
An alternative explanation for the failure to observe clear relationships between an atom's ADP value and its location within a protein's structure could be that no such relationships exist. It may not be possible to formulate a model that adequately predicts ADP values in terms of simple measurements derived from protein crystal structures. If ADPs reflect atomic motion then, perhaps, an approach that explicitly models the equilibrium dynamics of protein crystals may be necessary. The work in chapters 3 and 4 point towards packing density (alpha-carbon coordination number) being a dominant influence on ADP values. In conjunction, the trends, albeit weak, between ADP values and the distances to the protein surface and COM suggest that a more rigorous consideration of the shapes of proteins may be required. As a starting point, the question of whether the ADP values of the protein data sets could be explained by modelling the crystals as a simple ENMs was explored.

ENMs were considered to be appropriate models for protein crystals because the theory underlying ENMs is based on the same assumptions that are employed in crystallographic refinement. An ENM assumes that the protein's conformation has achieved a state of equi-

librium. The protein's tertiary structure is essentially fixed, and all atomic motion is reduced to simple harmonic oscillations about the atoms' average positions. Conceptually, an ENM can be visualised as a coarse "bead and spring" model of a protein. Structural elements of the protein are represented by beads and are connected by springs modelling the combined inter and intra-molecular interactions. Typically, ENMs model proteins with one bead per amino acid. However, coarser representations are possible with individual domains represented by one or more beads. At the other extreme, ENMs can model the interactions between each individual atom where the ENM becomes equivalent to the application of NMA to an all atom MD force field.

Figure 5.1 is a visualisation of an elastic network for the crystal of the "lasso" peptide 3NJW (Nar *et al.* 2010). The peptide is not one of the proteins analysed in this thesis, but is used as an example because its small size makes it clearer to see the individual "beads" and "springs". Each amino acid is represented by a spherical bead centred on the amino acids' alpha-carbons. The springs are represented by blue lines and two beads are connected by a spring if their centres are within 8 Å of one another. The lattice contains multiple repeats of the unit cell and corresponding chains within the unit cells are shaded with the same colour to emphasise the symmetry of the crystal. For clarity, the whole structure has been rotated to view the crystal along one lattice axis. Figure 5.1 illustrates both the inter and intra-molecular interactions modelled by the ENM. Although the majority of the interactions are between beads within the same peptide, there are also interactions between adjacent peptides that hold the lattice together.

Figure 5.1: Visualisation of the elastic network model for the "lasso" peptide 3NJW (Nar *et al.* 2010).



ENM were constructed at the amino acid scale because atomic scale models would have been too demanding in terms of both time and computational resources. Furthermore, using amino acid "beads" in an ENM allows for a straight-forward comparison with the alpha-carbon ADP data derived in chapters 3 and 4. Following convention, an ENM is constructed with amino acid beads centred at the coordinates of the alpha-carbons in the crystal structure. As the structure moves, the mean-square displacements of the beads about their equilibrium positions will be proportional to the values of the alpha-carbon ADPs in the original crystal structure. In addition, constraints can be applied to the ENM that correspond to the assumptions made during crystallographic refinement. A Gaussian Network Model (GNM) (Tirion 1996; Bahar *et al.* 1997; Haliloglu *et al.* 1997) is a form of ENM where the displacements of the beads are isotropic and, therefore, the mean-square displacements correspond to isotropic B-factors. The Anisotropic Network Model (ANM) (Doruker *et al.* 2000; Atilgan *et al.* 2001) lifts the restriction of isotropic displacements allowing free movement in all directions. The mean-square displacements of the beads of an ANM correspond to the AADPs of crystal structures refined anisotropically.

5.2 Aim

The aim of this study is to compare the ADP of high resolution crystal structures to the predictions of ENMs. GNMs will be used to calculate mean-square displacements that can be compared to the isotropic B-factors of the proteins analysed in chapter 3. The anisotropic "equivalent" B-factors (B_{iso}^{equiv}) of the proteins analysed in chapter 4 will be compared to the mean-square displacements calculated by ANMs.

5.3 Hypothesis

An ENM is expected to be a reasonable approximation for the equilibrium dynamics of a protein crystal. Isotropic B-factors should correlate with the mean-square displacements predicted by GNMs of protein crystals. Similarly, anisotropic equivalent B-factors would be expected to correlate well with the predictions of ANMs.

5.4 Results and discussion

5.4.1 Gaussian network models

GNMs were constructed and isotropic B-factors estimated for all the proteins analysed in chapter 3. Two GNMs were derived for each protein: a model of the isolated protein and a model of the crystal lattice. The models of the single proteins were derived using the structures of the crystallographic asymmetric units. The GNM of the crystal lattices were derived by reconstructing the crystallographic unit cell and assigning connections between amino acids consistent with the periodicity of the crystal lattice. A cutoff distance of 8 Å was used when assigning connections between amino acids. This distance was chosen because, as confirmed by the work in chapters 3 and 4, the distance between an alpha-carbon and its immediate neighbours is within a range of 3–8 Å. For simplicity, the interactions between connected amino acids were all modelled as springs with identical elasticity constants.

The agreement between the GNMs and the experimental data was quantified by calculating Spearman correlation coefficients between the alpha-carbon isotropic B-factors published in the PDB files and the mean-square displacements for the amino acid beads predicted by the model. The non-parametric Spearman method for calculating correlation coefficients was used in preference over the Pearson method because, from the analysis of B-factor data in previous chapters, the data is unlikely to meet the criteria of being both normally distributed and homoscedastic. The calculations of the correlation coefficients did not include alpha-carbons for which there was no experimental data. All mean-square displacements derived from atoms that had been added by modelling software were omitted.

Figure 5.2: Distribution of the Spearman correlation coefficients between the predictions of GNMs and experimentally determined isotropic B-factors.

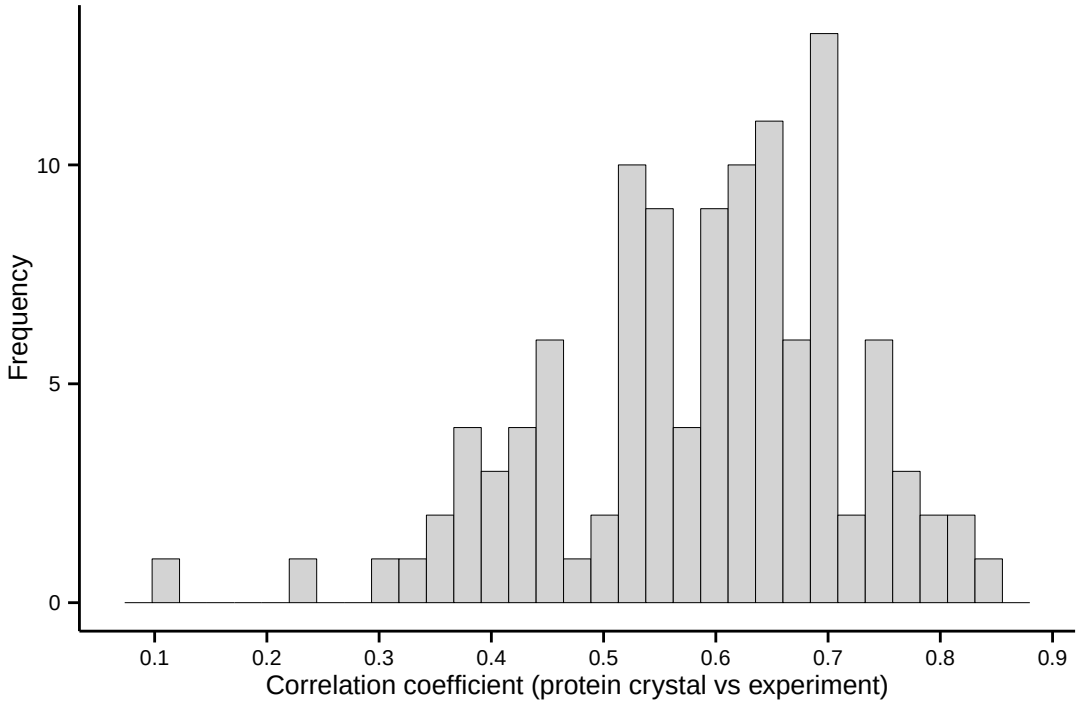
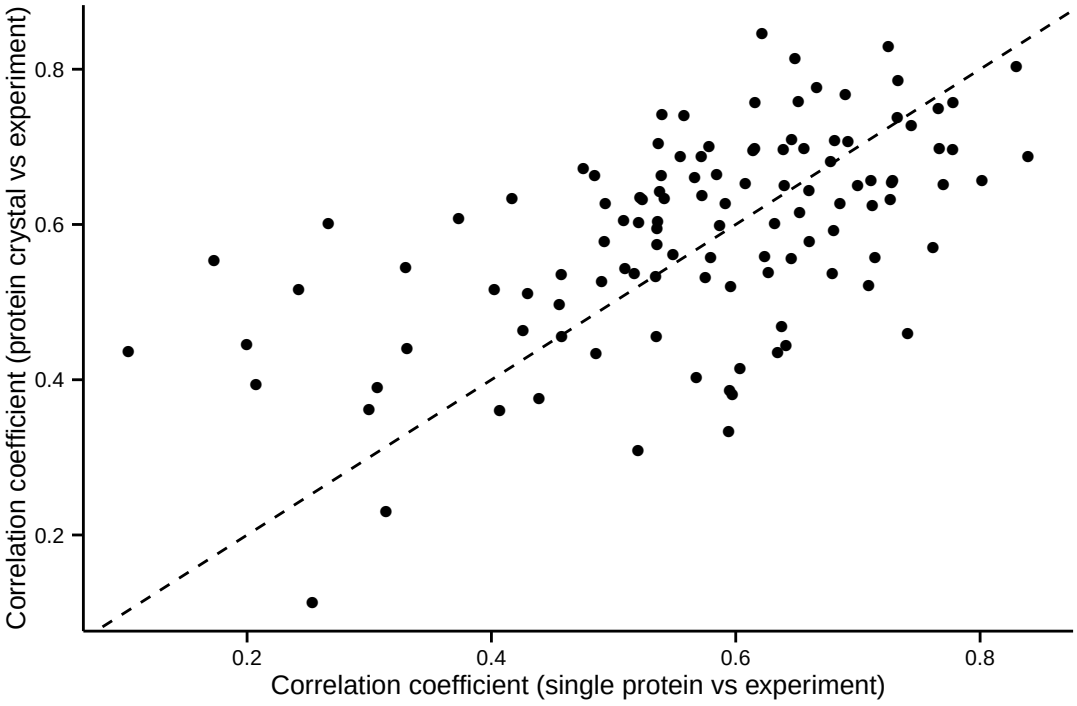


Figure 5.2 plots the distribution of Spearman correlation coefficients between experimental isotropic B-factors and the predictions of the GNM for the protein crystal lattices. The distribution shows a broad range of correlation values. The GNMs predict the isotropic B-factors of some proteins almost perfectly while others show a very poor correspondence. On average, there is a weak correlation between the experimental data and the predictions of the GNMs. The mean Spearman correlation coefficient is 0.589 with a standard deviation of 0.131.

Unexpectedly, the results for the GNMs derived for the asymmetric units of the proteins are almost identical to the results for the crystal lattices. The mean correlation coefficient for the single proteins is 0.569 with a standard deviation of 0.149. A paired Mann-Whitney test showed no statistically significant differences between the mean-square displacements calculated for the crystal lattices and those derived for single proteins ($p > 0.1$). Therefore, for the GNM, the influence of lattice structure and the contacts between proteins has a negligible influence on equilibrium dynamics. The comparison between GNMs for single and lattice proteins is visualised in figure 5.3. Most points in the scatter plot follow the central dashed line, highlighting that the correlation coefficient calculated for a protein in isolation is usually very similar to the calculation when the protein is a component of a crystal lattice.

Figure 5.3: Scatter plot of the Spearman correlation coefficients between the predictions of GNMs and experimentally determined isotropic B-factors for GNMs of single proteins and proteins in the unit cell. The dotted line indicates the points where the correlation coefficients are equal



5.4.2 Anisotropic network models

ANMs were constructed analogously to GNMs using the data set of crystallographic structures refined anisotropically described in chapter 4. ANMs for both single proteins and crystal lattices were constructed using the same 8 Å cutoff distance between amino acids and an elasticity constant of unity. Whilst it is usually recommended to apply a longer cutoff distance when building an ANM, the ANMs were constructed using the same cutoff as for the GNMs to allow for a direct comparison between the two types of ENM. It was not possible to derive an ANM model for one of the proteins of the data set because the size of the resulting Kirchhoff matrix for the unit cell was too large to diagonalise. This protein was excluded from the analysis. The predictions of the ANMs were compared to experiment by calculating Spearman correlation coefficients between B_{iso}^{equiv} derived from the ANMs to the alpha-carbon B-factors published in the structures' PDB files. As discussed in chapter 4, the majority of alpha-carbon atoms in the data set had been refined anisotropically and, therefore, most B-factors published in the PDB files would be B_{iso}^{equiv} values derived from AADPs. This assertion was tested and it was confirmed that, with the exception of two structures, all the B-factors agreed with the their corresponding B_{iso}^{equiv} value to at least one decimal place. In the case of the two exceptions, the B-factors and B_{iso}^{equiv} values agreed but with a lower degree of precision (all B-factors and B_{iso}^{equiv} values differed by less than 0.5).

Figure 5.4 plots the distribution of Spearman correlation coefficients between experimental B-factors and the "equivalent" B-factors derived from ANMs of crystal lattices. The mean correlation coefficient is 0.640 with a standard deviation of 0.147. In comparison, the mean correlation coefficient between predicted and experimental B-factors for single proteins is lower at 0.564 (standard deviation of 0.157). Interestingly, unlike GNMs, a paired Mann-Whitney test showed that the differences between the correlation coefficients for single proteins and crystal lattices were statistically significant ($p < 10^{-5}$). This difference is apparent from the scatter plot of the correlation coefficients for single proteins against those calculated for crystal lattices (figure 5.5). In figure 5.5, many points lie above the dashed line indicating that the predictions of the ANM are more accurate when the structure of the crystal lattice is accounted for.

Figure 5.4: Distribution of the Spearman correlation coefficients between the predictions of ANMs and experimentally determined isotropic B-factors.

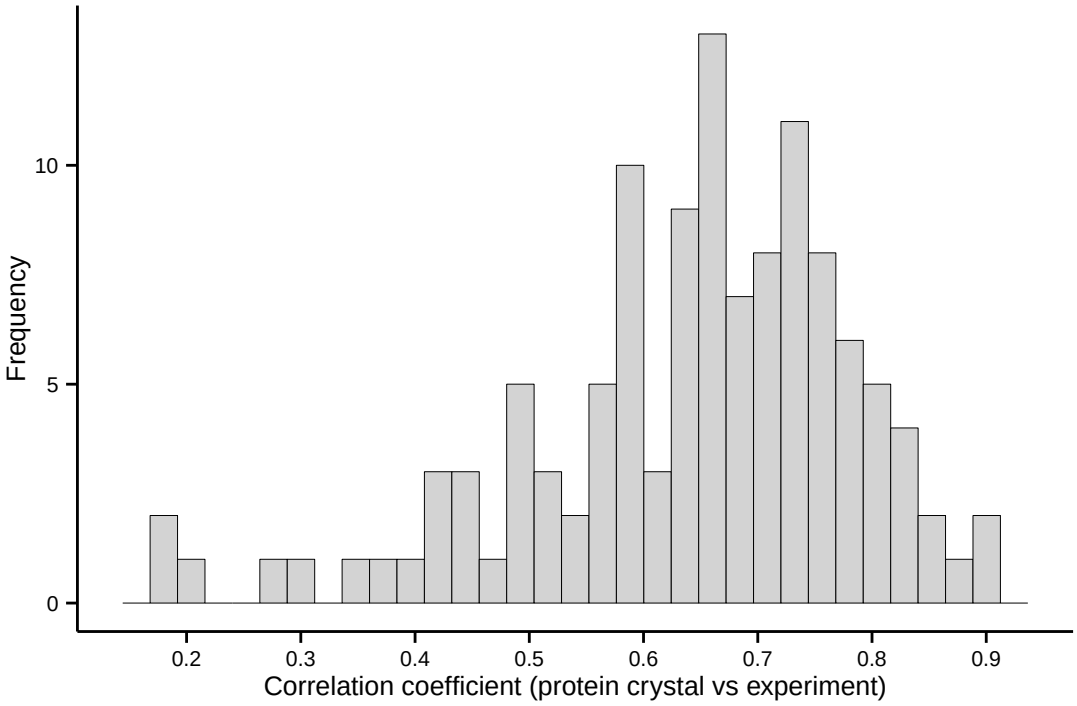


Figure 5.5: Scatter plot of the Spearman correlation coefficients between the predictions of ANMs and experimentally determined isotropic B-factors for ANMs of single proteins and proteins in the unit cell. The dotted line indicates the points where the correlation coefficients are equal

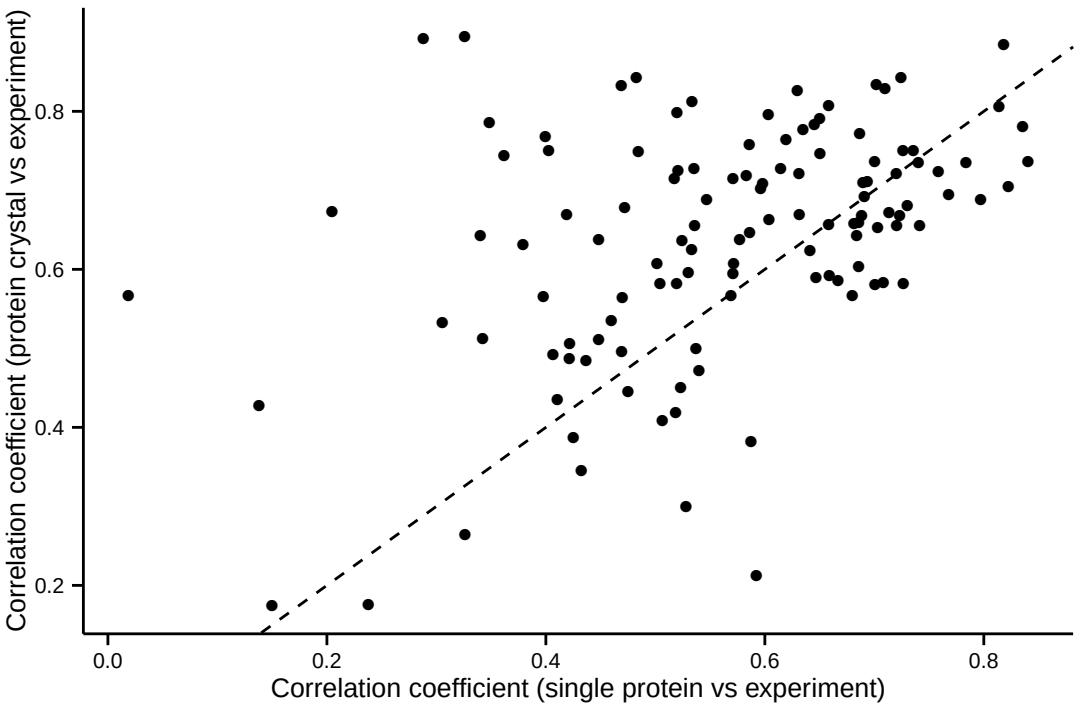


Table 5.1: Summary statistics for the Spearman correlation coefficients calculated between experimental B-factors and the predictions of Gaussian and anisotropic elastic network models for the data set of anisotropically refined structures.

	Single proteins		**Crystal lattices**	
Elastic Network	mean±sd	median±mad	mean±sd	median±mad
ANM	0.564±0.157	0.583±0.108	0.640±0.147	0.663±0.082
GNM	0.610±0.151	0.641±0.077	0.634±0.116	0.658±0.073

Paired Mann-Whitney tests with a significance level $p < 0.05$:

1 Statistically significant differences between single proteins and crystal lattices for the ANM.

2 No statistically significant differences between single proteins and crystal lattices for the GNM.

3 Statistically significant differences between GNM and ANM for single proteins.

4 No statistically significant differences between GNM and ANM for crystal lattices proteins.

The mean correlation coefficient for the ANMs of crystal lattices is higher than the mean correlation coefficient achieved with GNMs of the data set of isotropically refined proteins (0.640 compared to 0.589). To investigate whether an ANM is a better model of equilibrium dynamics than a GNM, GNMs were derived for the data set of anisotropically refined structures. Unexpectedly, the predictions of the GNM were comparable to those of the ANMs. The mean Spearman correlation coefficient between experiment and B-factors derived from GNMs of crystal lattices is 0.634 with a standard deviation of 0.116. A paired Mann-Whitney test between the predictions of the GNM and ANM for the crystal lattices did not identify any statistically significant differences ($p = 0.0741$). The GNM of the single proteins, however, performed significantly better than the corresponding ANMs with a mean correlation coefficient of 0.610 and standard deviation of 0.151 (Mann-Whitney $p < 10^{-5}$). These statistics are summarised in table 5.1.

The results of the comparison between the Gaussian and anisotropic elastic network models for the anisotropically refined structures are interesting for two reasons. Firstly, for crystal lattices, the more sophisticated ANM does not generate significantly better predictions for B-factors when compared to the GNM. Secondly, the structure of the crystal lattice appears to be important for an ANM but less so in the case of the GNM. The insensitivity of a GNM to lattice structure could be explained by the fact that it is a more constrained model than an ANM. A GNM imposes restrictions on both the extent and direction of an amino acid's displacement from its equilibrium position. An ANM relaxes the limits on the direction of movement, allowing for asymmetrical displacements. Therefore, because a GNM is a more constrained system, the sparse protein-protein interactions across the crystal lattice may have a lesser influence on the dynamics compared to the more numerous interactions between an amino acid and its immediate neighbours. In contrast, due to fewer constraints, the dynamics of an ANM may be more sensitive to the effects of lattice structure. This may explain why longer cutofff distances are usually employed when modelling proteins with ANMs compared to GNM (Atilgan *et al.* 2001; Riccardi *et al.* 2009). Extending cutoff distances generates a

network with more extensive interactions and, in a sense, imposes more structural "order" on the system. Perhaps, in the case of crystal lattices, long cutoff distances for ANMs are not necessary because the periodicity and symmetry of the lattice will restrict the dynamics of the proteins.

5.5 Methods

5.5.1 Creating Gaussian network models

GNMs were derived for the set of crystal structures analysed in chapter 3. All structures were complete and the coordinates for all atoms were in maximum occupancy locations. All atoms were removed from the structures except for the alpha-carbons. The coordinates of the alpha-carbons were used as the locations of the amino acid "beads" in the GNMs. The structure of the crystallographic unit cells were reconstructed by applying the symmetry operations defined in the PDB files. A virtual "spring" connected two amino acid beads if the distance between the alpha-carbons was less than 8Å.

The structure of a GNM was encoded as a Kirchhoff matrix (Bahar *et al.* 1997; Haliloglu *et al.* 1997):

$$K = \begin{pmatrix} k_{1,1} & k_{1,2} & \dots & k_{1,N} \\ k_{2,1} & k_{2,2} & \dots & k_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ k_{N,1} & k_{N,2} & \dots & k_{N,N} \end{pmatrix} \tag{5.1}$$

The elements $k_{i,j}$ encode the connectivity between amino acids i and j. Amino acids are numbered sequentially following the order of the protein primary structure. The numbering is contiguous across the multiple chains of the unit cells. Therefore, for a unit cell containing m chains of length n, the amino acids indices run consecutively from 1 to $N = m \times n$.

The values of the elements $k_{i,j}$ are set using the following rule:

$$k_{i,j} = \begin{cases} -1 & \text{if } i \neq j \text{ and the amino acids are within the cutoff distance} \\ 0 & \text{if } i \neq j \text{ and the amino acids are outside the cutoff distance} \\ s_i & \text{if } i = j \end{cases} \tag{5.2}$$

where s_i is a positive integer such that all rows and columns of the Kirchhoff matrix sum to zero i.e.,

$$s_i = -\sum_{j=1, j\neq i}^{N} k_{i,j}$$

The periodicity of the crystal lattices was taken into account when determining the connectivity between amino acids.

Isotropic B-factors were estimated by calculating the mean-square displacements for each amino acid within the unit cells. The Kirchhoff matrix of a GNM is proportional to the Hessian matrix modelling the simple harmonic dynamics of network. The mean-square

displacements of the amino acids were calculated exactly through diagonalisation of the Kirchhoff matrix and its inversion to obtain a covariance matrix for atomic displacements (Bahar *et al.* 1997; Haliloglu *et al.* 1997).

Let Kirchhoff matrix K be diagonalised in the form $K = U\Lambda U^T$ where U is the orthogonal matrix of eigenvectors and Λ a diagonal matrix of eigenvalues. Then, the inverse matrix was calculated as $K^{-1} = U\Lambda^{-1}U^T$. Strictly speaking, the Kirchhoff matrix is non-invertible since it has at least one zero eigenvalue due to the degeneracy of the GNM; however, diagonalisation allows for the calculation of the elements of the "inverse" Kichhoff matrix.

The mean-square displacements of the amino acids in the unit cell were calculated using the relationship:

$$< |\mathbf{\Delta r_k}|^2 > \propto \frac{1}{\gamma} \sum_{i=1}^{N} \frac{1}{\lambda_i} [\mathbf{u_i u_i}^T]_{k,k} \quad \text{for all } \lambda_i \neq 0 \tag{5.3}$$

where	$< \|\mathbf{\Delta r_k}\|^2 >$	is the mean-square displacement of the kth amino acid.
	γ	is the spring constant for the network.
	λ_i	is the ith eigenvalue of the Kirchhoff matrix.
	$\mathbf{u_i}$	is the eigenvector corresponding to eigenvalue λ_i.
	$[\mathbf{u_i u_i}^T]_{k,k}$	is the kth element on the leading diagonal of the matrix resulting from the product of the ith eigenvector with its transpose.

For convenience, the spring constant γ was set to 1 for all GNMs

5.5.2 Creating anisotropic network models

ANMs were derived for the set of crystal structures analysed in chapter 4. The same preprocessing steps used for GNMs were followed to create unit cell structures of alpha-carbons. The virtual "springs" connecting two amino acid beads used the same 8Å cutoff distance. However, unlike GNMs, ANMs define multiple springs between amino acids. In combination, the springs model the direction dependent oscillations of the atoms about their equilibrium positions.

For a unit cell of N amino acids, the Hessian matrix modelling the dynamics is a $3N \times 3N$ matrix. Amino acids are indexed following the same convention used for GNMs, but the rows and columns of the Hessian matrix now account for three Cartesian variables for each amino acid.

The Hessian matrix of an anisotropic ENM can be visualised as the concatenation of 3 sub-matrices each of which describe the interactions between two amino acids. If $H_{i,j}$ is a sub-matrix for the interactions between amino acids i and j, then the Hessian matrix for the network, $\mathcal{H}$, is given by (Atilgan *et al.* 2001):

$$\mathcal{H} = \begin{pmatrix} H_{1,1} & H_{1,2} & \dots & H_{1,N} \\ H_{2,1} & H_{2,2} & \dots & H_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N,1} & H_{N,2} & \dots & H_{N,N} \end{pmatrix} \tag{5.4}$$

For $i \neq j$, the elements of the sub-matrices $H_{i,j}$ are calculated from the Cartesian coordinates for the two amino acids. For two amino acids separated by a distance greater than the 8Å cutoff, the sub-matrix $H_{i,j}$ is a zero matrix representing no interaction. Otherwise, if the coordinates of amino acids i and j are (x_i, y_i, z_i) and (x_j, y_j, z_j) respectively, then

$$H_{i,j} = \frac{-\gamma}{d^2} \begin{pmatrix} h_{x,x} & h_{x,y} & h_{x,z} \\ h_{y,x} & h_{y,y} & h_{y,z} \\ h_{z,x} & h_{z,y} & h_{z,z} \end{pmatrix} \tag{5.5}$$

where

$$\begin{aligned} h_{x,x} &= (x_i - x_j)^2 \\ h_{y,y} &= (y_i - y_j)^2 \\ h_{z,z} &= (z_i - z_j)^2 \\ h_{x,y} &= h_{y,x} = (x_i - x_j)(y_i - y_j) \\ h_{x,z} &= h_{z,x} = (x_i - x_j)(z_i - z_j) \\ h_{y,z} &= h_{z,y} = (y_i - y_j)(z_i - z_j) \\ d^2 &= (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \\ \gamma &\quad \text{is the spring constant for the network (set to 1)} \end{aligned}$$

The elements of the sub-matrices $H_{i,j}$ where $i = j$ are set, analogous to a GNM, to values that ensure that all the sub-matrices of each row and column of the Hessian matrix $\mathcal{H}$ (equation 5.4) sum to give zero matrices. Thus,

$$H_{i,i} = -\sum_{j=1, j\neq i}^{N} H_{i,j} \tag{5.6}$$

The Cartesian covariance matrix describing mean-square anisotropic displacements of each amino acid is derived by inverting the Hessian matrix $\mathcal{H}$ following a similar methodology to that used for GNMs. Hence, the elements of the inverse Hessian matrix $\mathcal{H}^{-1}$ are given by

the equation:

$$\mathcal{H}^{-1}[a,b] \;=\; \sum_{c=1}^{3N} \frac{1}{\lambda_c}[\mathbf{u_c}\mathbf{u_c}^T]_{a,b} \quad \text{for all } \lambda_c \neq 0 \tag{5.7}$$

where	$\mathcal{H}^{-1}[a,b]$	is the element at row a and column b of the inverse Hessian matrix $\mathcal{H}^{-1}$.
	λ_c	is the cth eigenvalue of the Hessian matrix $\mathcal{H}$.
	$\mathbf{u_c}$	is the eigenvector corresponding to eigenvalue λ_c.
	$[\mathbf{u_c}\mathbf{u_c}^T]_{a,b}$	is the the element at row a and column b of the matrix resulting from the product of the cth eigenvector with its transpose.

Like the original Hessian matrix $\mathcal{H}$, the inverse Hessian $\mathcal{H}^{-1}$ can be viewed as the concatenation of 3 sub-matrices. Each of these sub-matrices is a covariance matrix for the anisotropic displacements of the amino acids in the unit cell. The Cartesian covariance matrices giving the anisotropic mean-square displacements for each amino acid about its equilibrium position are the sub-matrices running along the leading diagonal of the inverse Hessian.

$$\mathcal{H}^{-1} = \begin{pmatrix} H_{1,1}^{-1} & H_{1,2}^{-1} & \dots & H_{1,N}^{-1} \\ H_{2,1}^{-1} & H_{2,2}^{-1} & \dots & H_{2,N}^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ H_{N,1}^{-1} & H_{N,2}^{-1} & \dots & H_{N,N}^{-1} \end{pmatrix} \tag{5.8}$$

Thus, the Cartesian covariance matrix for the anisotropic displacements of the kth amino acid of the unit cell, U_k^C, is given by the sub-matrix $H_{k,k}^{-1}$ of the inverse Hessian. The equivalent B-factor for the amino acid is calculated from the trace of the matrix U_k^C as given in equation 1.4 in chapter 1.

Chapter 6

Using isotropic B-factor data to validate molecular dynamics force fields

6.1 Introduction

The accumulated evidence from previous chapters points towards ADPs being unreliable indicators of proteins conformational dynamics. One possible cause for the this inaccuracy may be a high level of background "noise" associated with ADP data. As discussed previously, ADPs can be influenced by factors unrelated to the movements of atoms. It was, perhaps, incorrect to assume that, by only considering high quality atomic resolution structures, ADP values would be predominantly determined by conformational dynamics. The high degree of variability in the data suggests otherwise. Even in the highest quality crystal structures, the dynamical component to ADP values appears to be buried under a layer of noise.

It was considered whether it might be possible to eliminate the noise associated with ADP data. It would be reasonable to assume that the noise would be random and unlikely to be the result of systemic errors common to all crystallographic experiments. Therefore, by averaging the ADP data over multiple determinations of the same crystal structure, it may be possible to derive consensus ADP values that better reflect protein dynamics. In contrast to the strategy of previous chapters, where diverse collections of *single* crystal structures were analysed, the focus of the research switched to the small number of proteins for which multiple crystal structures have been published.

A equally valid criticism of the previous analysis is whether the methods chosen to measure conformational variability are sufficiently accurate. In chapters 3 and 4, static structural features were assumed to be correlates of protein flexibility while, in chapter 5, proteins were

reduced to very simplistic elastic network models. It could be argued that these approaches are too coarse to adequately account for all the subtle effects that may influences protein dynamics. In particular, accurate modelling of the chemical structure of proteins and the interactions between the proteins and the water and ions that permeate the crystal. An all-atomistic MD simulation is one method for modelling the dynamics of a protein crystal that could account for all these different factors. Therefore, by running MD simulations of protein crystals it may be possible to generate estimates for ADPs that align closely with the protein's consensus (averaged) ADP profile. Furthermore, if MD simulations are found to be consistent with consensus ADP profiles, then this suggests a new method to validate MD force fields. Assuming a protein's consensus ADP profile is a true reflection of its dynamics, the MD force field that most closely reproduces the profile is the force field that most realistically models protein molecules.

6.2 Aim

The initial aim will be to identify those proteins in the PDB whose structures have been determined multiple times. Averaging the ADPs data for a protein should reveal whether a consensus exists or if ADPs are simply complex error terms with little relation to the protein's dynamics. If consensus ADP profiles can be established for the proteins, then these profiles will be compared to atomic fluctuation measurements derived from MD simulations of crystals. Repeating the simulations using different MD force fields may reveal which force fields most accurately model protein dynamics.

6.3 Hypothesis

Averaging ADP data over multiple independent crystal structures of the same protein will eliminate much of the "error" that obscures the relationship between protein conformational flexibility and ADP values. Consensus ADP profiles should correlate with the atomic fluctuations predicted by MD simulations of protein crystals. The strength of the correlation will reflect how well the force field models protein chemistry. ADP estimates from simulations using modern force fields such as CHARMM and AMBER would be expected to correlate more strongly than those that use older force fields such as All Atom Optimised Potentials for Liquid Simulations (OPLS-AA).

Table 6.1: Details of the clusters of X-ray structures sharing more than 95% sequence homology.

Number of structures	Example PDB Id	Description
521	1VSP	Lysozyme C (hen egg white lysozyme)
494	164L	T4 lysozyme
469	3MGO	Beta-2-microglobulin and its complexes
265	1HRO	Prokaryotic ribosomal subunits[1]
239	2SPO	Myoglobin
238	4J67	Pancreatic ribonuclease
225	3HV5	Mitogen-activated protein (MAP) kinase
218	2PWR	HIV-1 protease[1]
215	1YHR	Haemoglobin[1]
209	1STN	Staphylococcal nuclease
203	3JYT	HIV-1 reverse transcriptase
200	1LHL	Human lysozyme
195	4NVC	Cytochrome c peroxidase
181	1XW7	Insulin[1]
137	3T10	Heat shock protein HSP 90-alpha
129	4L79	Calmodulin and its complexes
125	3TMP	Ubiquitin and its complexes
123	3TLI	Thermolysin
120	2I4J	Peroxisome proliferator-activated receptor gamma
114	2VH5	H-ras GTPase and its complexes
113	3MVD	Histone-DNA complexes
111	5CPP	Cytochrome P450
108	3F3V	Proto-oncogene tyrosine-protein kinase Src
100	3K7A	Eukaryotic RNA polymerase II subunits

1 Multiple clusters identified for these proteins. Only the largest cluster is reported.

6.4 Results and discussion

6.4.1 Analysis of PDB clusters

The PDB generates statistics on the sequence homology of all protein structures deposited. These statistics are published in the form of a report listing all clusters of NMR and X-ray structures sharing 95% or more sequence homology. This cluster data was used to identify proteins for which a large number of X-ray structures were available. A summary of all the clusters identified containing 100 or more X-ray structures is given in table 6.1.

Subsequent analysis of these structures assessed the quality and the degree of similarity between the structures in each cluster. Crystals of single proteins were favoured over large macromolecular assemblies to ensure consistency between the protein-protein and protein-nucleic acid interactions across the cluster. Ribosomal subunits, calmodulin, ubiquitin, beta-2-microglobulin, histone complexes and RNA polymerase were, therefore, discounted from

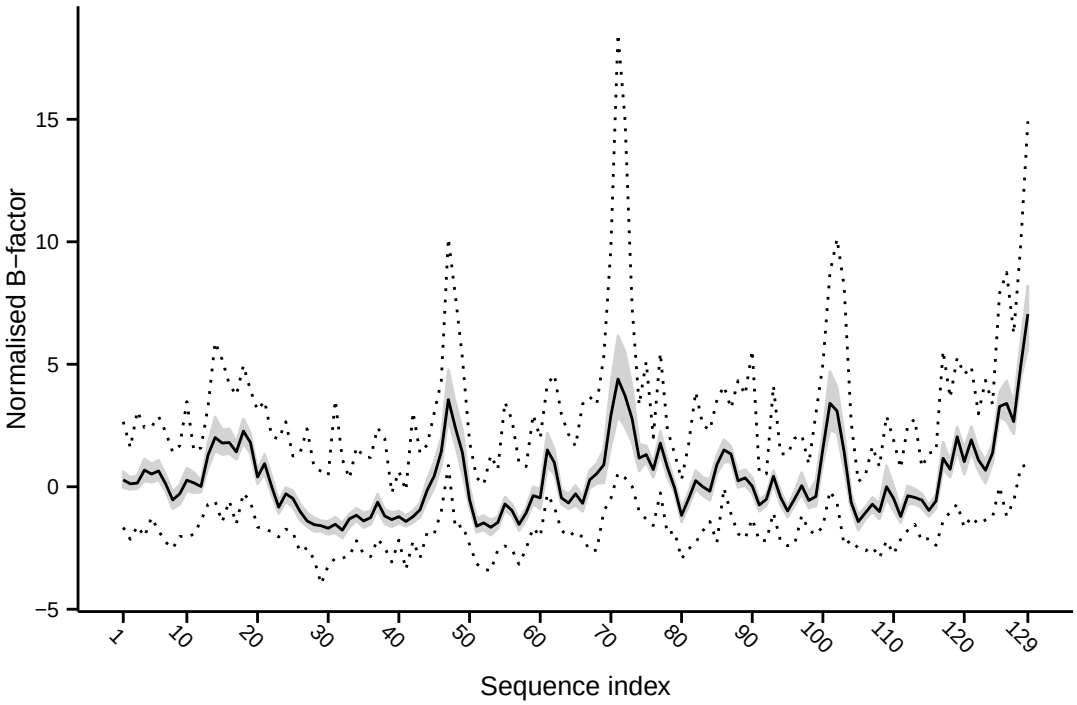
subsequent analysis. For the same reason, clusters were subdivided according to the crystals' space groups to eliminate variability in lattice structures. Differences between the sequences of the proteins in each cluster were determined through comparisons with a reference structure of the wild-type protein. Structures were discarded from the clusters if the sequences deviated from the reference by more than a few engineered substitution mutations. Where possible, structures containing large ligands were excluded from the analysis. However, in certain cases, ligands were permitted where there was consistency across the structures; for example, all myoglobin structures contained the same haem prosthetic group.

It was not possible to establish a global set of criteria to assess the degree of consistency between structures. Each cluster varied in the number, degree of sequence similarity and quality of its structures. Slightly different filtering criteria were, therefore, applied to each cluster to ensure a sufficient number of structures were retained. Nonetheless, it was not always possible to find the required level of agreement between the structures in every cluster. Clusters were discounted if fewer than ten similar structures could be found. The lack of consistency between structures is the reason that HIV-1 reverse transcriptase; peroxisome proliferator-activated receptor gamma; mitogen-activated protein kinase and proto-oncogene tyrosine-protein kinase Src were excluded from further analysis. Table 6.2 summarises the analysis of the clusters of crystal structures and lists the proteins for which reasonable numbers of similar structures could be identified.

Table 6.2: Summary of the clusters of X-ray structures after selecting structures sharing a high degree of similarity.

Description	Space group	Number of structures	Reference PDB Id
Hen egg white lysozyme (lysozyme C) *Max. 2 substitutions*	P 4_3 2_1 2	252	1HEL
T4 Lysozyme *Max. 2 substitutions*	P 3_2 2 1	76	3LZM
Human Lysozyme *Max. 2 substitutions*	P 2_1 2_1 2_1	150	1LZ1
Staphylococcal nuclease *Max. 2 substitutions*	P 4_1	30	1STN
Pancreatic ribonuclease (*Bos taurus*) *Max. 2 substitutions*	P 1 2_1 1	21	1KF5
Pancreatic ribonuclease (*Bos taurus*) *Max. 2 substitutions*	P 3_2 2 1	54	1KF5
Cytochrome c peroxidase (*Saccharomyces cerevisiae*) *Max. 5 substitutions. Includes haem.*	P 2_1 2_1 2_1	22	2CYP
Cytochrome P450 (*Pseudomonas putida*) *Max. 2 substitutions. Includes haem and camphor.*	P 2_1 2_1 2_1	14	2ZAX
Sperm whale myoglobin *Max. 2 substitutions. Includes haem.*	P 1 2_1 1	55	1SWM
Sperm whale myoglobin *Max. 4 substitutions. Includes haem.*	P 6	100	1MLO
Human haemoglobin *Max. 4 substitutions. Includes haem.*	P 2_1 2_1 2	54	3HHB
Human insulin *Max. 2 substitutions.*	H 3	35	1TRZ
Thermolysin (*Bacillus thermoproteolyticus*) *Max. 2 substitutions. Includes large ligands.*	P 6_1 2 2	82	3FB0
Human heat shock protein 90 *Max. 2 substitutions. Includes large ligands.*	I 2 2 2	31	3T0H
HIV-1 protease *Max. 6 substitutions. Includes large ligands.*	P 2_1 2_1 2_1	31	1HPX
Human HRas GTPase *Max. 2 substitutions. Includes large ligands.*	H 3 2	35	2RGE

Figure 6.1: Consensus alpha-carbon B-factor profile for hen egg white lysozyme



6.4.2 Consensus B-factor profiles for PDB clusters

The structures in each of the clusters listed in table 6.2 vary in resolution and crystallographic quality and, consequently, only a small proportion of the structures are resolved with AADPs. Therefore, only isotropic B-factors, or the "equivalent" B-factors in the case of anisotropically refined atoms, are considered when averaging ADPs over the cluster. Figure 6.1 plots the profile of the median-mad normalised B-factors for the alpha-carbon atoms of the hen egg white lysozyme structures. The median normalised B-factor values are plotted as a solid line; the interquartile range as a grey ribbon and the minimum and maximum values are represented by dotted lines. The figure shows that, although there is a high degree of variability in B-factor values across the set of structures, a consensus exists. The median normalised B-factors trace a definite profile within a narrow interquartile range. This confirms that the B-factors are consistent across all the structures in the cluster. If there was no agreement between the structures, and B-factors were error terms independent of the protein's structure, then the graph plotted in 6.1 would be very different. Instead of plotting a defined profile, the median B-factors would be close to a straight line, flanked either side by a broad interquartile range.

Graphs of normalised alpha-carbon B-factor profiles for the other clusters of table 6.2 are included in appendix B. Although containing fewer crystal structures than hen egg white lysozyme, all the other proteins give similar results. When averaged over a number of

crystal structures, consensus B-factor profiles exist. Therefore, B-factor values appear to be structure dependent and not catch-all error terms of the crystallographic refinement process.

Despite being dictated by the composition of the PDB, there is diversity across the clusters investigated. A consideration of the differences between the proteins and their crystal structures hints at some of the factors that might influence the B-factors. Six of the clusters represent the crystal structures of single chain proteins in the absence of large ligands. As might be expected, clear consensus B-factor profiles emerge for the three lysozymes, staphylococcal nuclease and ribonuclease. However, even for these "ideal" proteins, there are outlier alpha-carbon B-factors that deviate significantly from the consensus. A logical explanation for these outliers is that they are caused by perturbations to the protein's structure arising from protein engineering or changes to the phyiscochemical properties of the crystals. This is to be expected, since the primary motivation for a researcher to repeat a crystallographic experiment is to explore how protein conformation is changed by mutation or physicochemical effects. In addition, as has already been discussed in previous chapters, B-factor data is inherently "noisy" and the presence of outliers is to be expected. Thus, only by averaging B-factor data over a number of independently determined crystal structure can atypical B-factor values be eliminated to reveal the underlying trends.

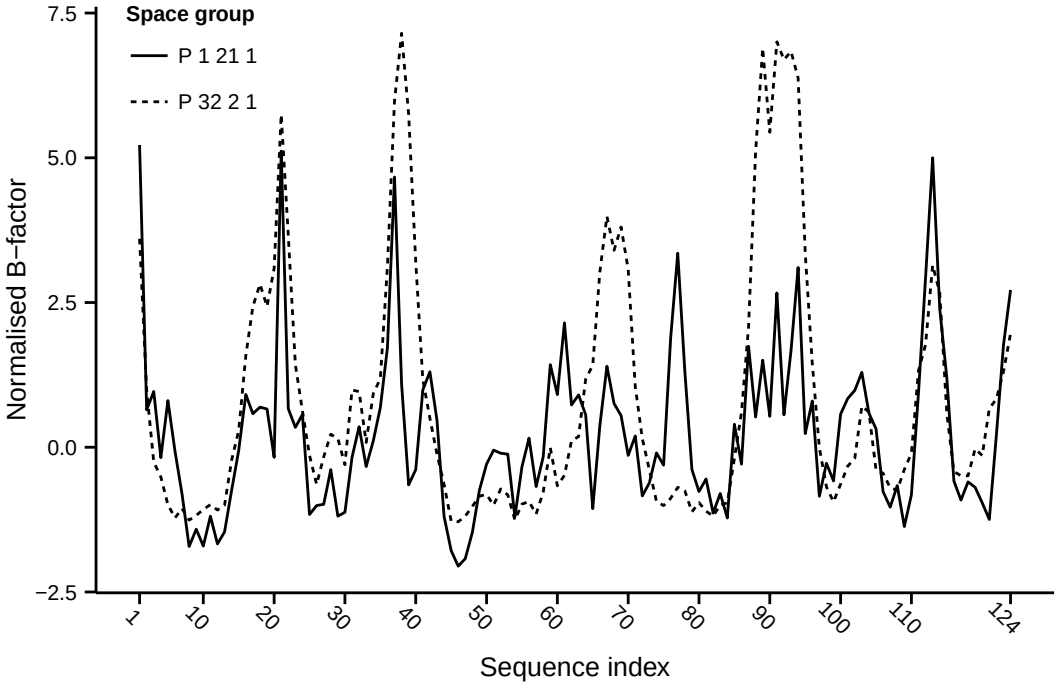
The majority of the crystal structures studied contain cofactors or other large ligands bound to the protein. Interestingly, the presence of these large organic molecules does not appear to cause major perturbations to the consensus B-factor profiles. In the case of haemoglobin, myoglobin and cytochrome c, all the structures within the cluster share common prosthetic groups bound at the same sites to the proteins. It is, therefore, not surprising to observe a high degree of consistency in the B-factor profiles. In contrast, in crystal structures such as HIV-1 protease, HRas GTPase and thermolysin, the proteins are bound to a disparate collection of ligands. It is somewhat unexpected that these proteins also show consensus B-factor profiles. A possible explanation is that, despite the ligands being structurally different they may be functionally equivalent; that is, they bind to the same target sites in the proteins and interact with the proteins in similar ways. For example, many of the ligands in the structures of HIV-1 protease are potential drug candidates that target the enzyme's active site.

Another notable feature of the data set is the effect of space group on the B-factor profile. Whilst most clusters fall into single space groups, the crystal structures of ribonuclease and myoglobin are split across two space groups. In the case of both ribonuclease and myoglobin, there are differences between the consensus B-factor profiles for the two space groups. Figure 6.2 superimposes the consensus B-factors profiles for the two space groups of both proteins to illustrate the extent to which the profiles deviate. In figure 6.2, most of the peaks and troughs occur at the same locations in the structure indicating a broad level of agreement between the profiles for the two space groups. However, there are differences in the relative heights of the peaks between the profiles. These deviations suggests that

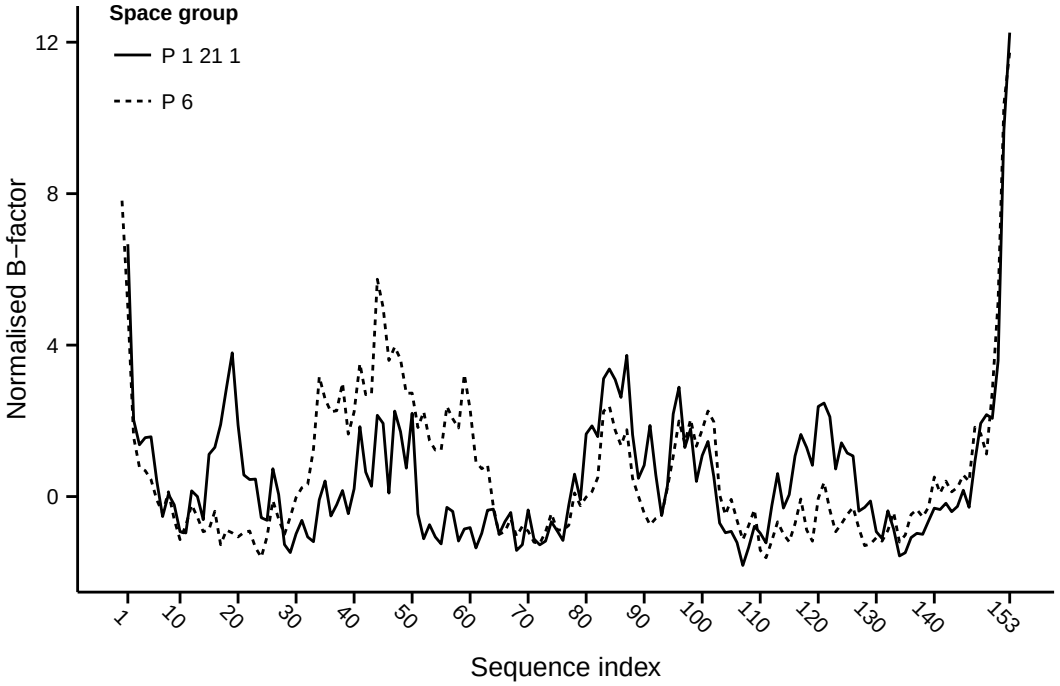
the structure of the crystal lattice is an important factor in determining B-factor values. The effect of space group could be attributed to differing protein-protein interactions or, perhaps, steric effects that give certain regions of the protein more or less conformational freedom depending on how the proteins are arranged in the crystal. These observations are supported by the weak positive correlation coefficients calculated between the space group profiles (table 6.3a). Interestingly, the differences observed for the myoglobin profiles are similar to those discussed by Kondrashov *et al.* (2008) who compared single high resolution myoglobin strucures in different space groups.

The consensus profiles of insulin, haemoglobin and HIV-1 protease also point towards the crystal lattice having a subtle influence on B-factor values. The asymmetric units for all of these proteins are multimeric, composed of multiple copies of chains with the same primary sequence. However, despite having identical sequences and very similar tertiary structures, the consensus profiles for crystallographically different copies of the same chain show deviations from one another. As an example, figure 6.3 superimposes the consensus profiles for the two copies of the alpha and beta chains of haemoglobin. Table 6.3b presents the correlation coefficients calculated for the two copies of the same chain in each cluster of structures. As seen previously when comparing different space groups, the two profiles for the same chain do not coincide exactly. This suggests that identical chains may exhibit the same underlying conformational dynamics modulated by the structure of the crystal lattice. The only exception is insulin's "A" chain where there is no agreement between the B-factor profiles. The small size of the "A" chain (21 amino acids) may offer an explanation for this discrepancy. The "A" chain is essentially a peptide and, therefore, its conformation and dynamics may be far more susceptible to influence by the local crystal environment. The two copies of the "A" chain are crystallographically different and, consequently, experience different intermolecular interactions.

Figure 6.2: Superimposition of the consensus alpha carbon B-factor profiles for crystals of the same protein in two different space groups.



(a) Space groups of ribonuclease



(b) Space groups of myoglobin

Table 6.3: Correlation coefficients for the B-factor profiles of protein chains with the same primary structure but crystallographically different structures

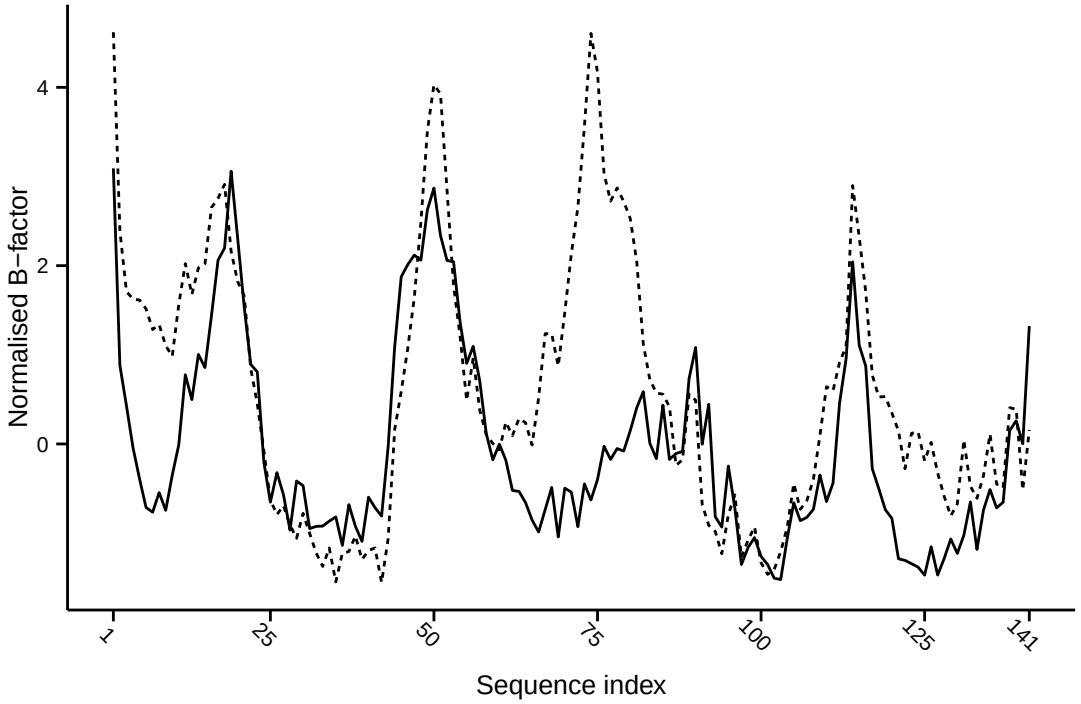
(a) Profiles for the same protein in different space groups

Protein profiles	Correlation coefficient	
	Pearson	**Spearman**
Ribonuclease (P 1 2_1 1 and P 3_2 2 1)	0.575	0.618
Myoglobin (P 1 2_1 1 and P 6)	0.663	0.417

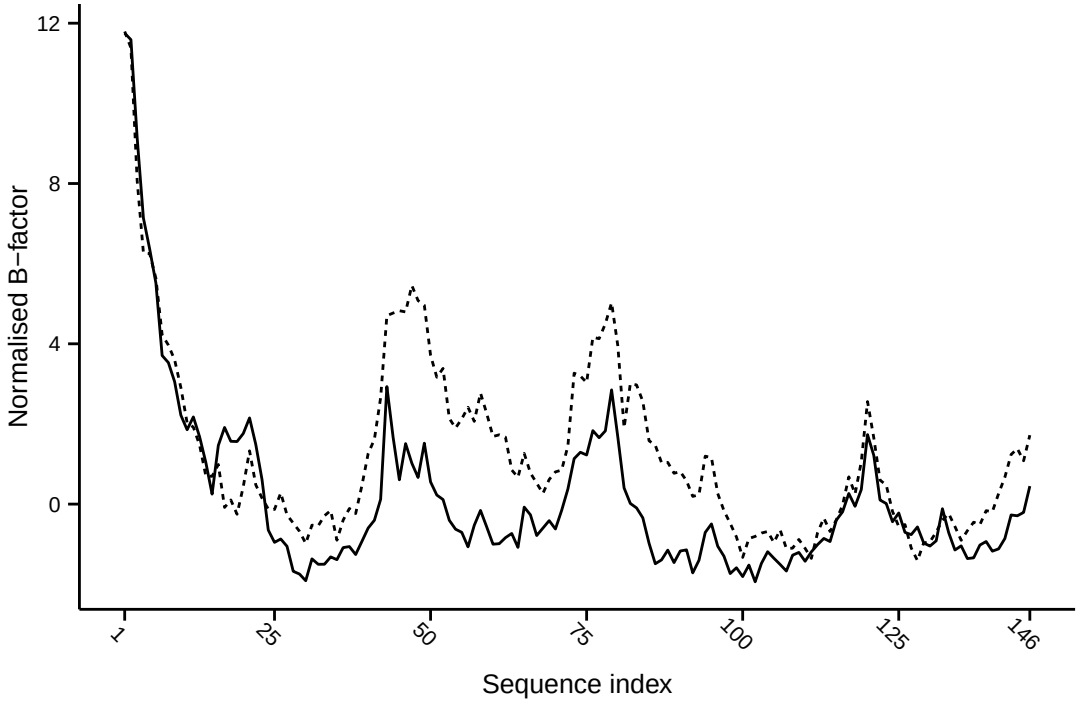
(b) Profiles for crystallographically different versions of the same chain in the asymmetric unit

Protein profiles	Correlation coefficient	
	Pearson	**Spearman**
HIV-1 protease	0.698	0.736
Insulin "A" chain	-0.120	-0.165
Insulin "B" chain	0.690	0.900
Haemoglobin alpha chain	0.643	0.665
Haemoglobin beta chain	0.848	0.743

Figure 6.3: Superimposition of the consensus alpha-carbon B-factor profiles for the two alpha and beta chains of haemoglobin.



(a) Haemoglobin alpha chains



(b) Haemoglobin beta chains

6.4.3 Molecular dynamics simulations

Having established that consensus B-factors can be derived for crystal structures, the next logical step was to investigate whether these profiles reflect the conformational dynamics of the proteins within the crystals. A decision was made to model the crystals with MD simulations since MD is generally accepted as one of the most accurate computational techniques for investigating protein motion at the atomic scale. Rather than attempt to simulate all the crystal structures discussed previously, the research focused on the simplest crystal structures. These are the crystals where the unit cells are primarily composed of water and multiple copies of a single protein chain. The five proteins modelled were: hen egg white lysozyme, human lysozyme, T4 lysozyme, staphylococcal nuclease and pancreatic ribonuclease. From the two most common space groups for ribonuclease crystals, P 12_11 was selected because the unit cell contained fewer proteins and was, therefore, less computationally intensive to model.

Proteins bound to cofactors or large ligands were not modelled because deriving correct MD topologies and force field parameterisations for the ligands would be an additional layer of complexity. The development of topologies and parameter sets for proteins and water molecules is more mature than those derived for other organic molecules and metal ions. As a result, it was assumed that models of protein-only crystals would be more reliable than models of crystals where proteins are complexed with ligands. Insulin was not chosen for a combination of reasons. Despite being a small protein, the hexagonal symmetry of insulin's space group, H 3, generates a larger unit cell compared to the other proteins, and would, therefore, be far more computationally demanding to simulate. In addition, a key feature of the structure of insulin crystals is the coordination with zinc ions (Smith *et al.* 1984; Smith, Pangborn *et al.* 2003; Dunn 2005). Organometallic interactions are not typically parameterised by MD force fields and, therefore, there is no guarantee that the effects of zinc coordination would be correctly incorporated into the model.

Crystals were simulated by reconstructing the arrangement of proteins in the unit cell and setting the simulation box to be equivalent to the parallelepiped that describes the unit cell. With periodic boundary conditions duplicating the unit cell in all directions, the simulation system models an infinitely large crystal lattice. The only other components of the simulation system were water molecules and chloride ions. The water and ions were added randomly to fill the cavities in the crystal and were not positioned according to known locations in the reference crystal structures. It could be argued that this would affect the accuracy of the simulation, but the solvent is rarely fully resolved by crystallography, so there would be insufficient information to model these molecules exactly. Furthermore, the salt composition of the simulations differs from the real crystals since MD can only model a protein in a fixed ionisation state and, by default, assumes the protein is in an aqueous environment at neutral pH. Ions were only added to neutralise the net charge of the system rather than attempt

Table 6.4: Composition of the simulation "boxes" representing the unit cells of protein crystals. The amount of water is an approximation since the number of water molecules varied for each simulation. The number of water molecules used with united atom force fields are higher compared to all-atom force fields and are, therefore, listed separately.

	Number of molecules			
		Water		Chloride
Protein crystal	**Proteins**	**all-atom**	**united-atom**	**ions**
Hen egg white lysozyme	8	2800	3300	64
Human lysozyme	4	1200	1500	32
T4 lysozyme	6	5400	6000	48
Staphylococcal nuclease	4	1950	2300	32
Ribonuclease (P 1 2_1 1)	2	770	880	8

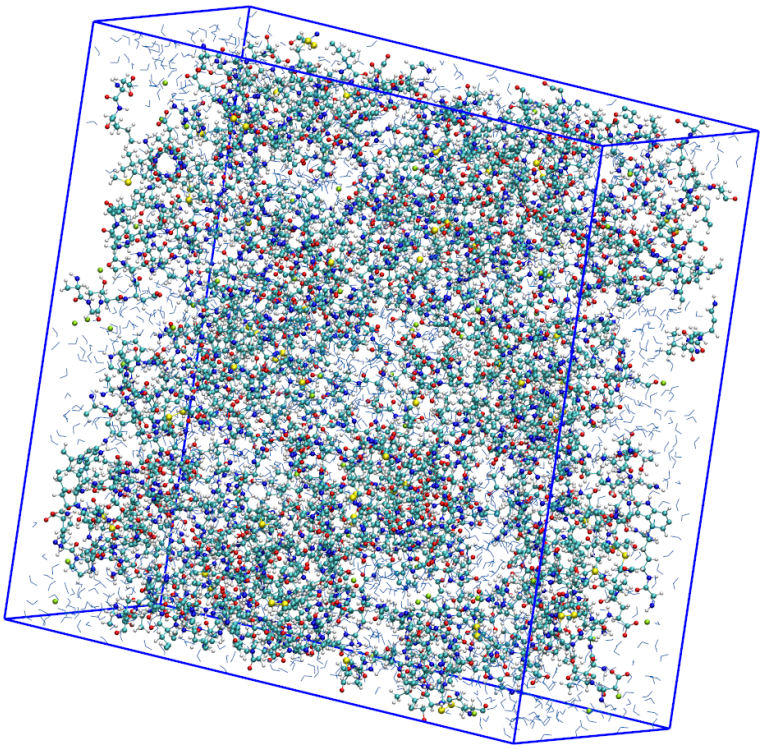
to model the ionic strength of the solvent. While the models of the crystals are arguably simplistic, the arrangement of proteins across the crystal lattice was accurately represented. Table 6.4 outlines the composition of the unit cells simulated for each protein crystal. As an example, figure 6.4 is a visualisation of the periodic unit cell box constructed for a hen egg white lysozyme simulation.

All simulations were run under the NVT ensemble at a constant temperature of 300 K. Pressure coupling was not applied in the simulations because this would have scaled the dimensions of the simulation boxes and, consequently, altered the structure of the crystal lattice. It was felt that it was more important to maintain a fixed lattice structure than attempt to account for atmospheric pressure. Although some crystallographic experiments are performed at room temperature, many are performed at cryogenic temperatures. A simulation temperature of 300 K would, therefore, seem unrealistic but is necessary when running standard MD simulations. MD force fields and topologies are parameterised to model molecules under standard physicochemical conditions and are, therefore, unlikely to reproduce cryogenic behaviour accurately.

Comparing MD simulations to B-factor profiles

Mean Square Fluctuation (MSF) was calculated for the movements of all the alpha-carbon atoms in the MD simulations of the crystal structures. The MSF of an atom should be directly proportional to its isotropic B-factor, assuming that isotropic B-factors measure the fluctuations of atoms about their average positions in a crystal structure. Prior to calculating MSFs, the protein chains were aligned to eliminate the effects of rigid body movements. Thus, the MSF values calculated for the alpha-carbons should only measure protein conformational flexibility as simulated by MD. Although it has been argued that B-factors may reflect the rigid body movements of the protein in a crystal, this thesis did not to attempt to incorporate any rigid body motion in the MD calculations. Simply calculating MSF without alignment would not be appropriate. The process of crystallographic refinement is more

Figure 6.4: Visualisation of the periodic simulation box used to model the unit cell of a hen egg white lysozyme crystal

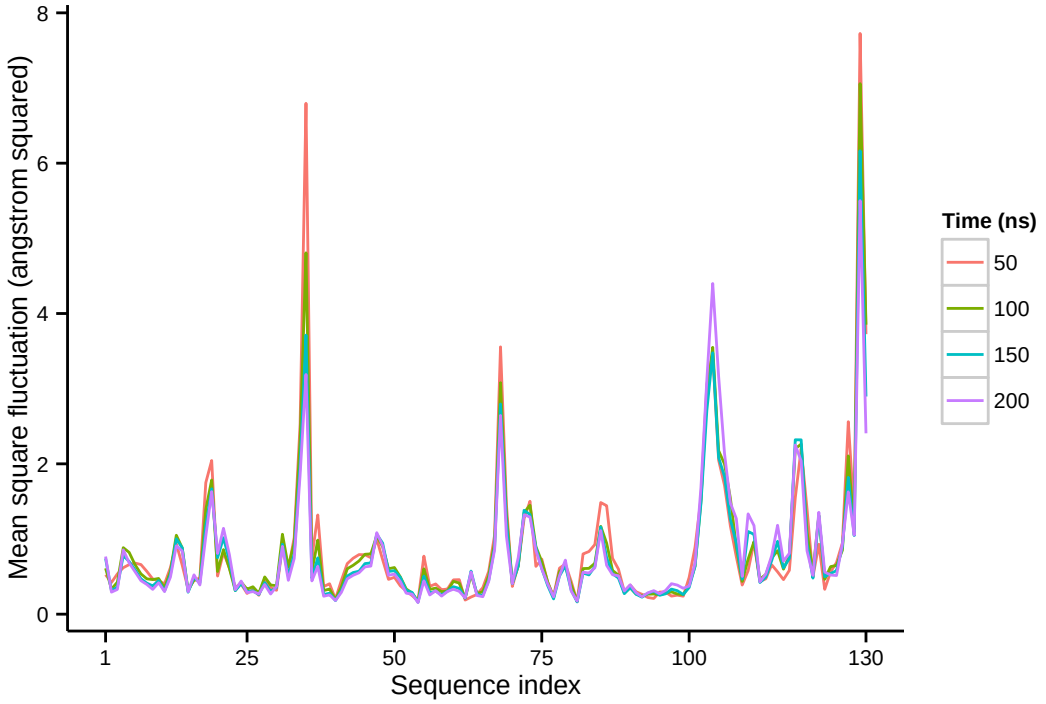


sophisticated than obtaining a set of average coordinates for atoms and their mean square deviations. Refinement generates feasible models for a structure under the constraints of protein chemistry and bond geometry. Furthermore, in the case of proteins whose structures have already been determined, existing structures may be used as a basis for formulating the the model. Therefore, atomic MSFs from an aligned MD trajectory are, perhaps, not so dissimilar from the isotropic B-factors derived from modern crystallographic refinement techniques.

In calculating MSF, it was assumed that the duration of the MD trajectory was sufficient to sample the equilibrium dynamics of the proteins. Preliminary experiments varied the number of simulations steps in order to determine an appropriate value to use for the production simulations. Too few steps would give inaccurate results while too many steps would be unnecessarily wasteful of computational resources. Figure 6.5 illustrates the reasoning behind the decision to limit the production simulations to 200 ns. MSF profiles for the alpha-carbon of one chain in a simulation of a human lysozyme crystal are plotted after 50 ns, 100 ns, 150 ns and 200 ns of simulation time. The plots show convergence in the shapes of the profiles after 150 ns suggesting that 200 ns is a reasonable end point for the simulations.

Calculation of the alpha-carbon MSF profiles revealed differences between the chains in the models of the unit cells. Figure 6.6a superimposes the MSF profiles for the four chains in the unit cell of human lysozyme over a 200 ns MD simulation. The models of the unit cells were

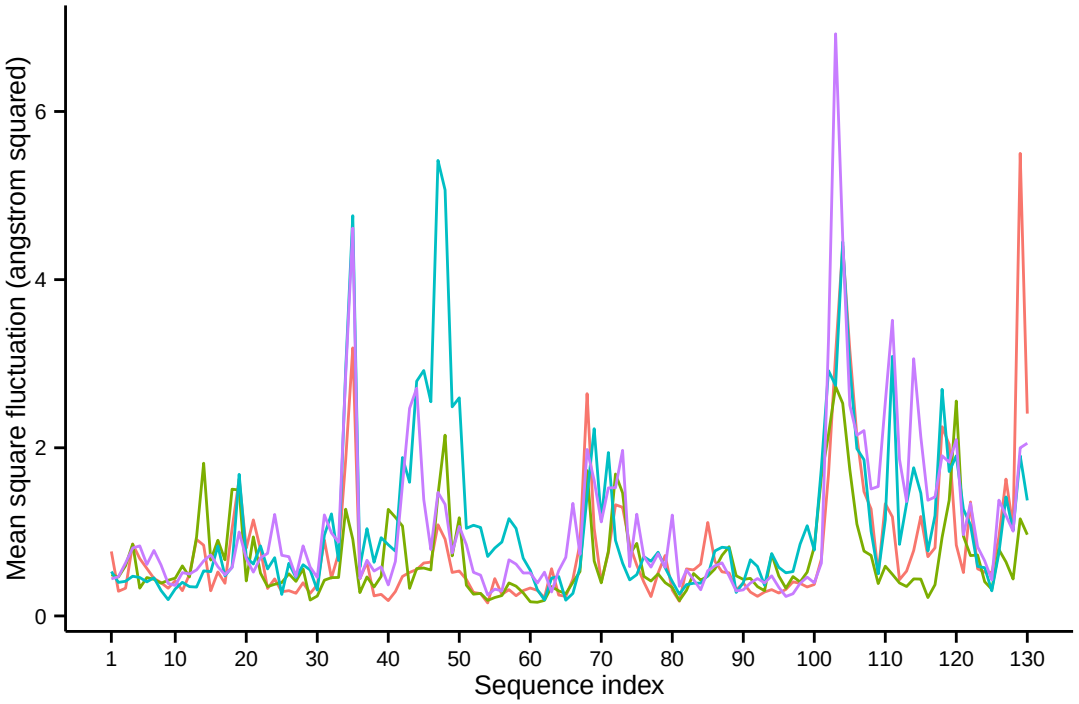
Figure 6.5: Alpha carbon MSF plots for one chain in an Amber99SB-ILDN MD simulation of a human lysozyme crystal. Profiles are calculated for simulations of the protein's dynamics over 50 ns, 100 ns, 150 ns and 200 ns.



constructed so that all the proteins made identical contacts with one another. Asymmetry was, however, introduced with the addition of water and chloride ions into the system. Thus, each protein will experience slightly different intermolecular interactions which may account for the differences in the dynamics. It is interesting to note that there is agreement in the shapes of the MSF profiles; that is, all show peaks and troughs in the same regions of the structure. All chains may share the same intrinsic dynamics as a consequence of their near identical tertiary structures. The differences between the MSF profiles could be the result of slightly different local environments in which the chains are situated. Figure 6.6b combines the MSF profiles over the four chains to highlight the common features.

Figure 6.6: Alpha-carbon MSF plots for the four chains in an Amber99SB-ILDN MD simulation of a human lysozyme crystal.

(a) MSF profiles for the four individual chains.



(b) Summary of the MSF profiles. The grey ribbon represents the full range of MSF values. The solid and dotted lines plot the median and mean MSF values respectively.

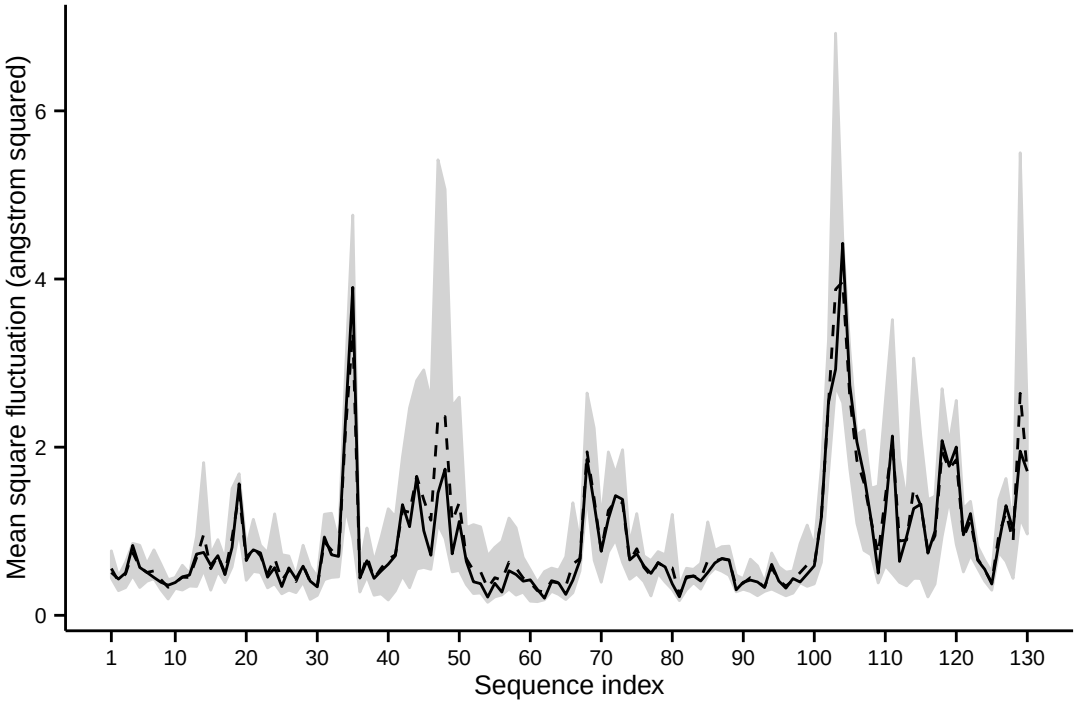


Table 6.5: Pearson and Spearman correlation coefficients between alpha-carbon MSF values and crystallographic B-factors. The MSF value of an alpha-carbon is the median value over all the chains in a 200 ns simulation of the model of the unit cell. The B-factor of an alpha-carbon is the median median-mad normalised B-factor over the cluster of crystal structures collated for the protein.

		Correlation coefficient	
Protein	**Force field**	**Pearson**	**Spearman**
Hen egg white lysozyme	Amber99SB-ILDN	0.638	0.663
	OPLS-AA	0.486	0.595
	CHARMM27	0.512	0.564
	GROMOS54a7	0.467	0.679
Human lysozyme	Amber99SB-ILDN	0.496	0.643
	OPLS-AA	0.408	0.559
	CHARMM27	0.270	0.595
	GROMOS54a7	0.307	0.583
T4 lysozyme	Amber99SB-ILDN	0.597	0.329
	OPLS-AA	0.625	0.502
	CHARMM27	0.477	0.550
	GROMOS54a7	0.589	0.507
Pancreatic ribonuclease	Amber99SB-ILDN	0.157	0.663
	OPLS-AA	0.351	0.475
	CHARMM27	0.250	0.473
	GROMOS54a7	0.496	0.556
Staphylococcal nuclease	Amber99SB-ILDN	0.803	0.740
	OPLS-AA	0.717	0.687
	CHARMM27	0.703	0.735
	GROMOS54a7	0.802	0.634

In order to compare the results of the MD simulations to the consensus B-factor profiles, it was necessary to summarise both sets of data. The median values of both the normalised B-factors and MD MSF data were chosen. Agreement between the median MSF values and median normalised B-factors was calculated as a linear correlation coefficient. Table 6.5 presents the correlation coefficient data for the five crystal structures and four MD force fields investigated.

The most notable feature of the data in table 6.5 is how poorly the MSF profiles agree with the consensus B-factor profiles across all five proteins. Simulations of staphylococcal nuclease appear to be the most accurate in reproducing crystallographic B-factors, but there is a caveat to these results. The N and C-termini of staphylococcal nuclease are not resolved by crystallography but were modelled by the MD simulations. The most unstructured and conformationally dynamic regions of this protein were, therefore, excluded from the calculation of the correlation coefficients. All the other proteins were completely resolved by crystallography and their terminal regions were included in the correlation calculations. It is feasible that inaccuracies in modelling the dynamics of the N and C-termini may be responsible for the low correlation coefficients. However, repeating the calculations with five residues

excluded from both ends did not improve the results.

One of the aims of this study was to try to use the consensus B-factor profiles to validate MD force fields. The analysis is, unfortunately, unable to differentiate between the MD force fields. All force fields appear to be roughly equivalent in their agreement with the consensus B-factor profiles, and there is no force field that clearly outperforms the others across all the crystal structures modelled. This is, perhaps, to be expected because MD force fields are not parametrised to reproduce the conformational dynamics of proteins. Maintaining expected bond geometries and correctly simulating protein folding are the main criteria against which force fields are assessed (Krieger *et al.* 2004; Soares *et al.* 2004; Zagrovic and van Gunsteren 2006; Best *et al.* 2008; Aliev and Courtier-Murias 2010; Hayre *et al.* 2011; Beauchamp *et al.* 2012; Cino *et al.* 2012; Lindorff-Larsen *et al.* 2012). Furthermore, the results of this study are consistent with current opinion amongst structural bioinformaticians. Although certain force fields may be optimised for specific applications, there is no general agreement as to which force field most accurately models the conformational dynamics of protein molecules.

The derivation of B-factors assume the harmonic fluctuations of atoms in proteins with a fixed conformation, so short MD simulations should, in theory, be adequate to predict B-factor values. However, the weak correlations between simulation and experiment suggest that the simulated proteins may have failed to exhibit the same degree of conformational variability as a real crystal. A 200 ns simulation of a crystal, while a reasonable time scale for MD, is brief in comparison to the time scale over which some protein conformational transitions occur (Henzler-Wildman and Kern 2007). In addition, the classical mechanics of MD can result in limited exploration of the dynamics across all of the conformational space available to the proteins (Tai 2004; Lei and Duan 2007). By averaging the MSF values over all the proteins in the unit cell, the models do account for some of diversity in the conformations and dynamics of the proteins in a crystal. However, the numbers of proteins in the simulations are minuscule in comparison to the trillions of proteins that comprise the crystals used in diffraction experiments. Furthermore, the high degree of symmetry imposed by the periodic boundary conditions of the MD simulation box may also mean that the movements of the proteins are too tightly coupled to one another. Thus, the conformations and dynamics of the simulated proteins may be biased and not representative of the population of proteins in a real crystal.

The data presented in table 6.5 are the results from single simulations of protein crystals. It was not feasible to repeat each simulation multiple times for every force field due to time limitations. MD simulations are chaotic systems; that is, the trajectories are deterministic but extremely sensitive to initial conditions (Braxenthaler *et al.* 1997). To assess whether the results were consistent, the simulations of human lysozyme were repeated three times for both the Amber99SB-ILDN and OPLS-AA force fields. Each repeat of the simulation started with the same protein crystal structure, but all other steps, including the addition of water and ions, were completely rerun. Thermodynamically, each production simulation was

Table 6.6: Pearson and Spearman correlation coefficients between alpha-carbon MSF values and crystallographic B-factors for independent repeated simulations of human lysozyme. The MSF value of an alpha-carbon is the median value over all the chains in a 200 ns simulation of the model of the unit cell. The B-factor of an alpha-carbon is the median media-mad normalised B-factor over the cluster of crystal structures collated for the protein.

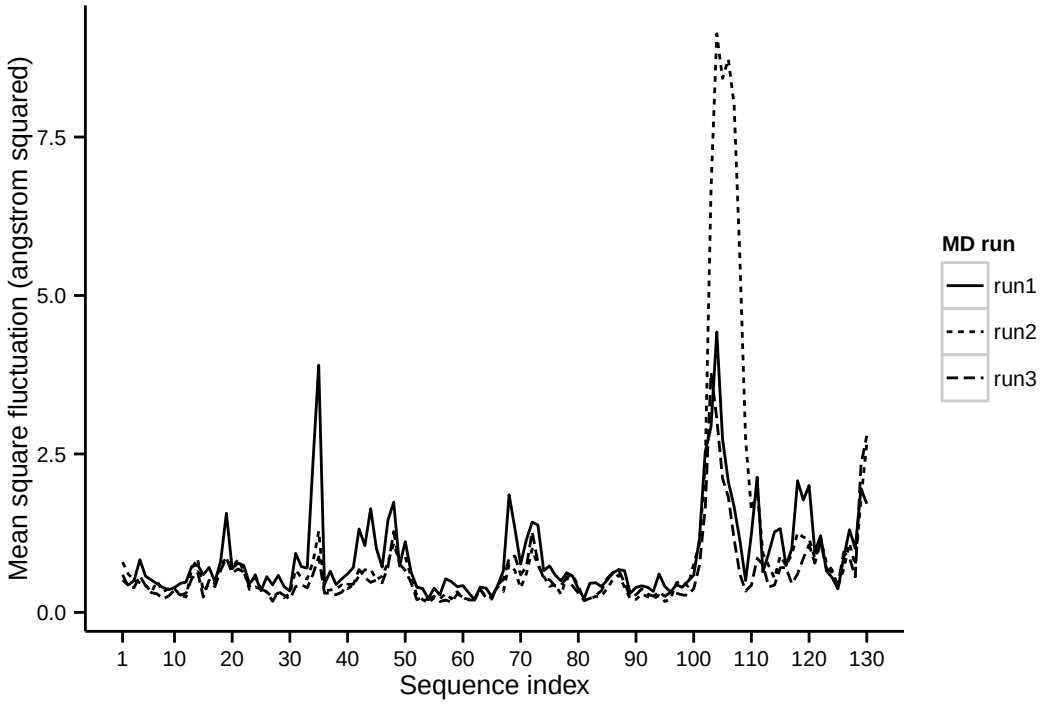
		Correlation coefficient	
Force field	**Run**	**Pearson**	**Spearman**
Amber99SB-ILDN	1	0.496	0.643
	2	0.318	0.674
	3	0.558	0.699
	mean±sd	0.457 ± 0.124	0.672 ± 0.028
OPLS-AA	1	0.408	0.559
	2	0.505	0.597
	3	0.514	0.668
	mean±sd	0.475 ± 0.059	0.608 ± 0.055

identical, but in terms of the initial positions and velocities of all the atoms in the system, each simulation run was different. The correlation coefficients for the repeated simulations are presented in table 6.6.

The correlation coefficients calculated in table 6.6 show that repeated simulations give variable but broadly consistent results. Comparing the correlation coefficients with a Mann Whitney test found no statistically significant differences between the two MD force fields tested. The median MSF profiles for the three simulations for both force fields are plotted in figure 6.7. Superimposing the median profiles shows that each simulation gives very similar median MSF profiles. This was confirmed by calculating the correlation coefficients between the median profiles for each repeated simulation (table 6.7). There is a surprisingly high level of agreement between the results of independent simulations. Each run traces a near identical MSF profile with the only difference being the extent of the fluctuations. This is reflected in the values of correlation coefficients with the Spearman coefficients being generally much higher than the Pearson. Therefore, the weak correlations between experimental B-factors and the predictions of MD simulations cannot be attributed entirely to the chaotic nature of MD trajectories.

Figure 6.7: Median alpha-carbon MSF plots for three independent simulation of a human lysozyme crystal (labelled run1, run2 and run3).

(a) Simulations using the Amber99SB-ILDN force field.



(b) Simulations using the OPLS-AA force field.

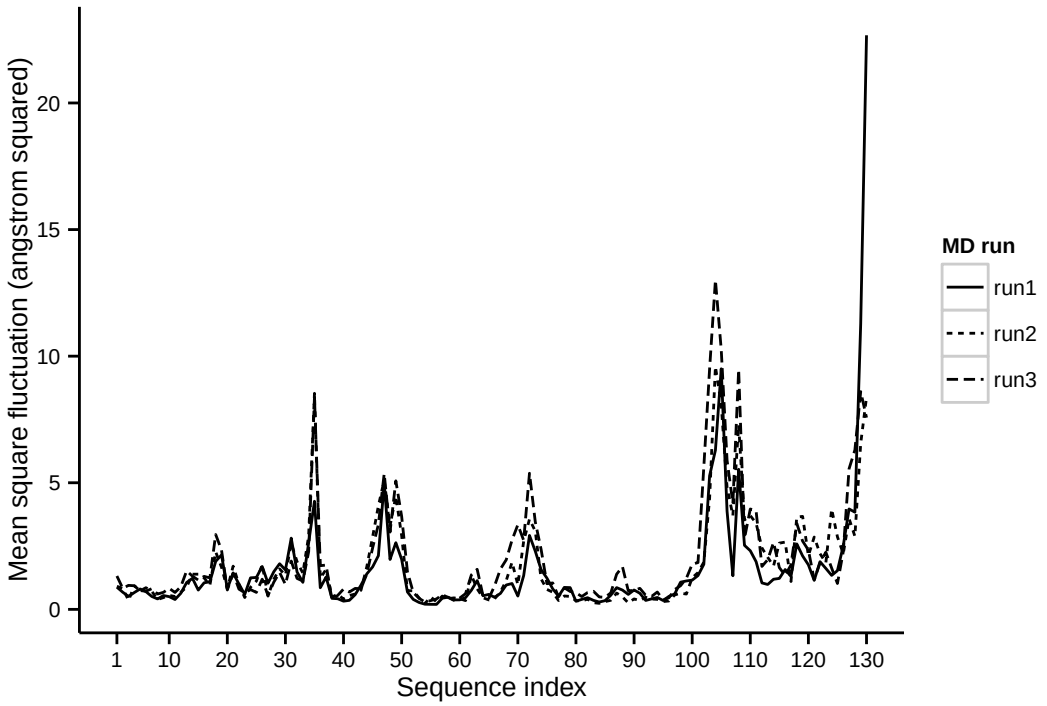


Table 6.7: Pearson and Spearman correlation coefficients between median alpha-carbon MSF values for three independent MD simulations. The Spearman correlation coefficients are printed in italic text.

(a) Amber99SB-ILDN force field

	run 1	**run 2**	**run 3**
run 1	1.000	0.669	0.777
	1.000	*0.853*	*0.867*
run 2		1.000	0.738
		1.000	*0.869*
run 3			1.000
			1.000

(b) OPLS-AA force field

	run 1	**run 2**	**run 3**
run 1	1.000	0.771	0.721
	1.000	*0.912*	*0.876*
run 2		1.000	0.911
		1.000	*0.892*
run 3			1.000
			1.000

Table 6.8: Pearson and Spearman correlation coefficients between alpha-carbon MSF values of proteins in solution and both the MSF values of simulations of the crystal lattices and the consensus B-factor profiles. The MSF value of an alpha-carbon in the crystal lattice is the median value over all the chains in a 200 ns simulation of the model of the unit cell. The B-factor of an alpha-carbon is the median media-mad normalised B-factor over the cluster of crystal structures collated for the protein.

	Correlation coefficient			
	Lattice simulations		**B-factors**	
Protein	**Pearson**	**Spearman**	**Pearson**	**Spearman**
Hen egg white lysozyme	0.615	0.805	0.418	0.606
Human lysozyme	0.619	0.813	0.268	0.677
T4 lysozyme	0.356	0.660	0.470	0.430
Pancreatic ribonuclease	0.736	0.758	0.366	0.485
Staphylococcal nuclease	0.874	0.615	0.795	0.639

6.4.4 Protein dynamics in solution

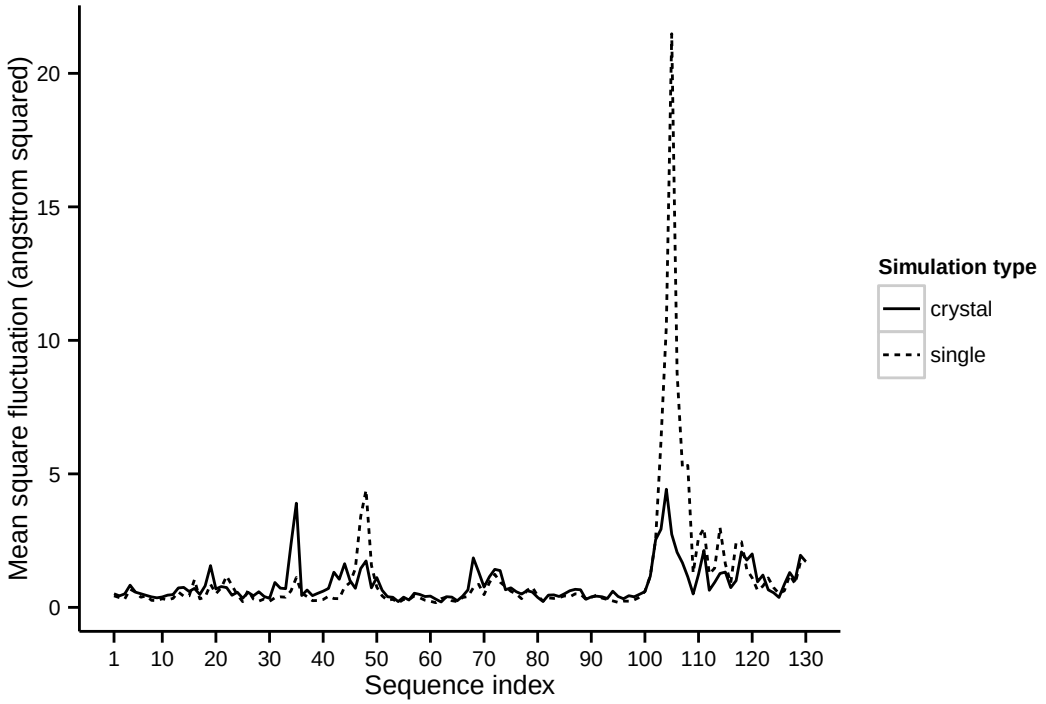
MD simulations of single proteins in solution were run to assess the effect of the crystal lattice on the simulations. The Amber99SB-ILDN force field was used for all the simulations. Single protein simulations are very computationally intensive and, therefore, repeating the simulations using all four force fields was not deemed to be practical. Furthermore, since the previous work had not revealed any marked differences between the force fields, a full comparison was considered to be unnecessary. Table 6.8 presents the calculations of the correlation coefficients between alpha-carbon MSF of the single proteins in solution and the corresponding MSF values for the simulations of the crystal lattices and the consensus B-factor profiles.

Comparing tables 6.8 and 6.5 is surprising in that the level of agreement between B-factors and MSF values is the same irrespective of whether the simulations are of single proteins or crystal lattices. Interestingly, the MSF profiles of single proteins are generally in agreement with the corresponding profiles in the crystal lattice. The stronger Spearman correlation coefficients suggest that the correspondence is with respect to the shapes of the MSF profiles rather than the MSF values. This is consistent with proteins having the same intrinsic dynamics both in solution and the crystal, but with the more sterically restrictive environment of the crystal modulating these movements. It is, perhaps, not surprising that the simulations of the proteins in crystals and solution are so similar. MD force fields are parametrised with constraints on permissible torsion angles for both the protein backbone and the amino acid side chains. Since both types of simulations start with the same protein structure, these restraints may bias the movements of the proteins so that they never deviate too far from the initial structure. Figure 6.8 superimposes the MSF profiles for the Amber99SB-ILDN MD simulations of human and T4 lysozyme both in solution and in the crystal. As expected from the correlation coefficients, the profiles of human lysozyme in solution and in the crystal are very similar and only differ in the extent of the fluctuations. In contrast, the profiles of T4

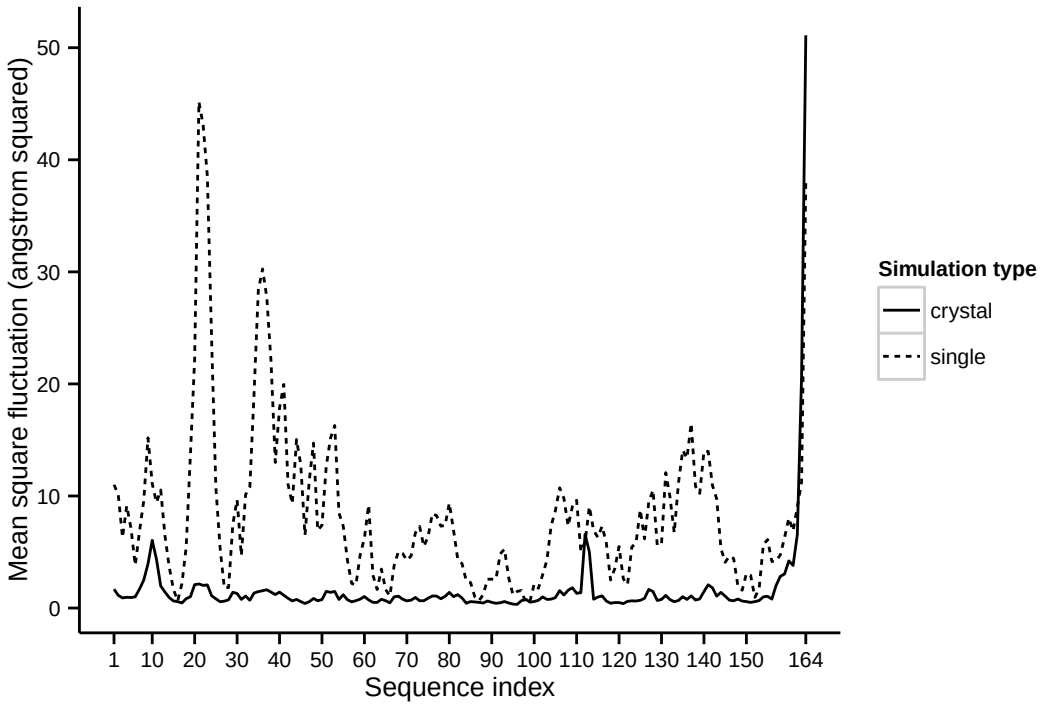
lysozyme are very different which is reflected in the poor correlation coefficients. The crystal environment appears to suppress the movements of T4 lysozyme to a much greater extent with only the N-terminal region showing any conformational variability in the crystal.

Figure 6.8: Superimposing the alpha-carbon MSF plots for protein in solution and in a crystal. The profile of eth crystal is the median profile over all proteins in the unit cell

(a) Simulations of human lysozyme.



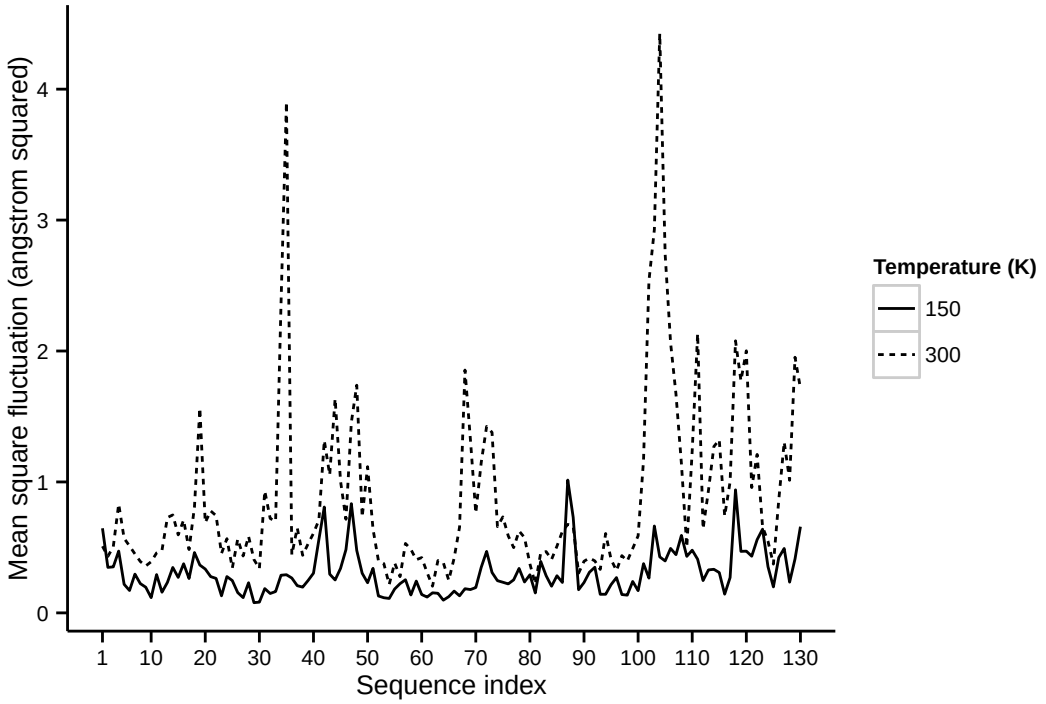
(b) Simulations of T4 lysozyme.



6.4.5 Low temperature simulations

The poor correlation between experimental B-factors and the MSF values derived from MD simulations could be attributed to the choice of 300 K as the simulation temperature. Although some crystallographic experiments are undertaken at ambient temperatures, many crystals are cooled to cryogenic or near cryogenic temperatures to limit damage caused by X-ray exposure. Unfortunately, it is not obligatory for PDB files to publish the temperature under which the X-ray diffraction pattern was recorded. Of the 150 PDB structures analysed for human lysozyme, 50 state crystal temperatures of 100 K; 48 at 283 K; and 33 give no temperature information. To test whether lowering the temperature of the simulation would give more representative dynamics for the crystals, a simulation of the human lysozyme crystal was repeated at 150 K. The median MSF profiles for the simulations at 150 K and 300 K are plotted in figure 6.9. The plot clearly shows that there is very little movement at 150 K with the extent of most fluctuations below 1 Å. As might be expected, there is some agreement between the regions exhibiting the most and least flexibility at the two temperatures. However, low temperature simulations did not improve the agreement with experimental B-factors. The correlation between the median alpha-carbon MSF values for the simulation at 150 K and the median normalised B-factor values for human lysozyme are 0.540 (Pearson) and 0.603 (Spearman). There are two possible reasons why the low temperature simulation offered no improvement when modelling the dynamics of the crystal. Firstly, assuming atomic fluctuations are proportional to temperature, normalisation of the B-factor data should correct for differences in the temperatures of the crystallographic experiments. Secondly, MD force fields are not parametrised to model the dynamics of the proteins or water far below standard conditions. In addition, MD simulation "temperature" is very different to the macroscopic temperature that would be recorded in a crystallographic experiment. It is more accurate to describe the simulation temperature as an indication of the total kinetic energy. Thus, there is also uncertainty concerning how representative the dynamics at a simulation temperature of 300 K are to the dynamics of the ensemble of molecules within a crystal at the equivalent macroscopic temperature.

Figure 6.9: Median alpha carbon MSF plots for the chain in an Amber99SB-ILDN MD simulation of a human lysozyme crystal. Profiles are plotted for 200 ns simulation at temperatures of 150 K and 300 K.



6.4.6 Alternative measures of conformational variability

An alternative argument that could account for the poor agreement between B-factor data and the MSF values derived from MD simulations is that the two quantities may not be directly comparable. An improved level of agreement may be possible by deriving a MSF value from the experimentally determined crystal structures rather than using B-factor data. By aligning all the independently determined crystal structures of the same protein, the MSF of the alpha-carbon atoms can be calculated in exactly the same way as for the MD trajectory. Figure 6.10 plots the profile for the deviations in the alpha-carbons coordinates across all the structures of human lysozyme. The plots of the median and interquartile range on the graph shows that, across the majority of structures, there is very little difference in protein conformation. Interestingly, the plot of the mean (equivalent to MSF) differs from the median and suggests that there are a small number of structures exhibiting alternate conformations in certain regions of the protein. The most likely explanation is that the mean has been distorted by a small number of structures showing "extreme" deviations in the positions of certain alpha-carbons through substitution mutations or non-standard crystallisation conditions. It might, therefore, be expected that, while both the mean and median deviations would correlate with B-factors and MD MSF measurements, the median deviations would be the more reliable measure of conformational variability. On the contrary, neither the mean nor median deviations show a particularly convincing correlation with the MD simulation data (table 6.9). There is a better correlation between B-factors and the deviations between aligned structure (table 6.10). Nonetheless, across all five proteins considered, the correlation between the structural deviation measurements and B-factors is weak.

Measuring structural deviations between aligned structures may not be the most appropriate measure of conformational variability. The method requires aligning structures, which is not only a computationally expensive operation, but can also give a distorted picture of conformational change. The Kabsch alignment algorithm superimposes two structures through the operations of translation and rotation to minimise the Euclidean distances between the atomic coordinates. The implicit assumption of the algorithm being that the proteins have near identical conformations and the translation and rotation will correct for any rigid body displacement. There is, therefore, no guarantee that the difference between two aligned structures will be an accurate reflection of protein conformational change. Consequently, an alternative approach was considered that could measure conformational variability without the reliance on structural alignment. The methodology employed was to examine the variability of the phi and psi torsion angles of the protein backbone. The rationale behind this decision was that these torsion angles describe rotations about bonds involving alpha-carbon atoms and measuring torsion angle variability might, therefore, give a flexibility measure comparable with alpha-carbon B-factors. The degree of variability in the alpha-carbon torsion angles was quantified by calculating angular dispersion. Consistent with

Figure 6.10: Profile of the square deviations in the positions of the alpha-carbon for human lysozyme with respect to the centroid of the cluster (structure `1C43`). The solid line plots the median deviation and the grey ribbon plots the interquartile range. The dotted line plots the mean deviations which are equivalent to the MSF measurements of a MD simulation.

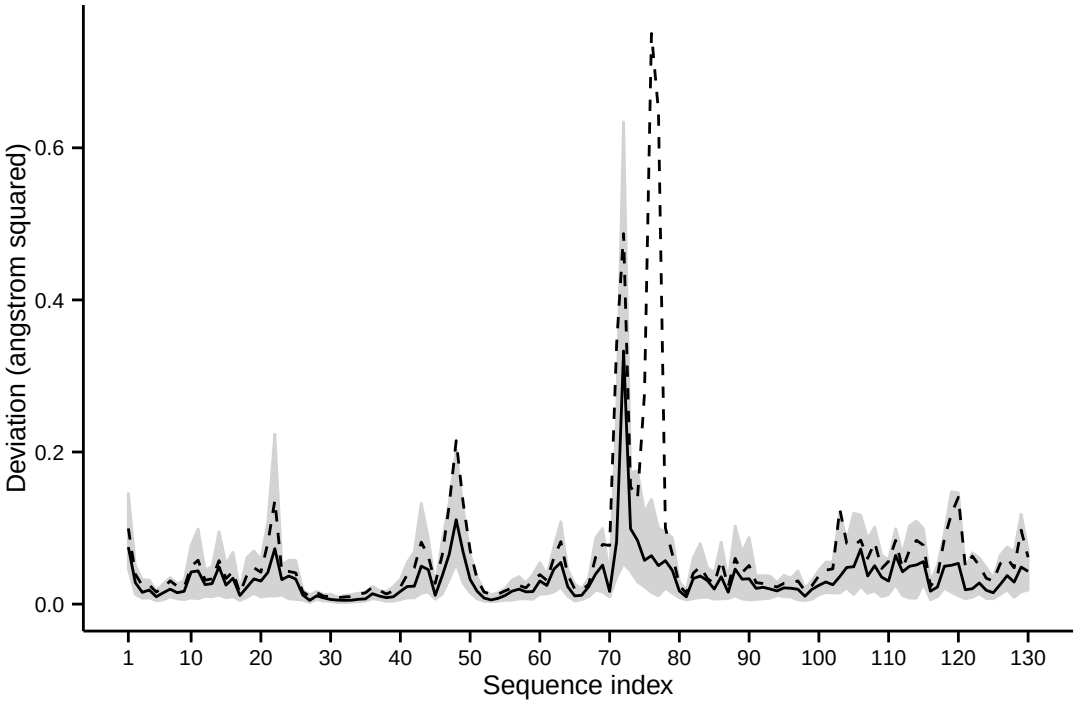


Table 6.9: Pearson and Spearman correlation coefficients between alpha-carbon MD MSF values and deviations measured between aligned PDB files. The MSF value of an alpha-carbon is the median value over all the chains in a 200 ns simulation of the model of the unit cell. The deviations between the alpha-carbon atoms of aligned structures is summarised as either the mean or median value.

		Correlation coefficient			
		Mean deviations		**Median deviations**	
Protein	**Force field**	**Pearson**	**Spearman**	**Pearson**	**Spearman**
Hen egg white lysozyme	Amber99SB-ILDN	0.470	0.613	0.583	0.528
	OPLS-AA	0.401	0.567	0.409	0.509
	CHARMM27	0.296	0.478	0.282	0.411
	GROMOS54a7	0.243	0.614	0.225	0.520
Human lysozyme	Amber99SB-ILDN	0.135	0.539	0.251	0.453
	OPLS-AA	0.070	0.414	0.172	0.351
	CHARMM27	0.059	0.416	0.148	0.333
	GROMOS54a7	0.077	0.546	0.215	0.499
T4 lysozyme	Amber99SB-ILDN	0.502	0.274	0.829	0.309
	OPLS-AA	0.359	0.498	0.626	0.385
	CHARMM27	0.231	0.349	0.447	0.359
	GROMOS54a7	0.428	0.265	0.551	0.276
Pancreatic ribonuclease	Amber99SB-ILDN	0.110	0.592	0.184	0.524
	OPLS-AA	0.377	0.492	0.299	0.411
	CHARMM27	0.053	0.258	0.156	0.443
	GROMOS54a7	0.499	0.634	0.472	0.522
Staphylococcal nuclease	Amber99SB-ILDN	0.674	0.726	0.730	0.688
	OPLS-AA	0.620	0.717	0.626	0.659
	CHARMM27	0.704	0.746	0.580	0.728
	GROMOS54a7	0.550	0.727	0.705	0.663

Table 6.10: Pearson and Spearman correlation coefficients between alpha-carbon consensus B-factor profiles and deviations measured between aligned PDB files. The B-factor of an alpha-carbon is the median media-mad normalised B-factor over the cluster of crystal structures collated for the protein. The deviations between the alpha-carbon atoms of aligned structures is summarised as either the mean or median value.

	Correlation coefficient			
	Mean deviations		**Median deviations**	
Protein	**Pearson**	**Spearman**	**Pearson**	**Spearman**
Hen egg white lysozyme	0.689	0.779	0.626	0.696
Human lysozyme	0.408	0.735	0.534	0.603
T4 lysozyme	0.615	0.340	0.742	0.460
Pancreatic ribonuclease	0.626	0.709	0.649	0.769
Staphylococcal nuclease	0.555	0.742	0.793	0.759

standard measures of variance, angular dispersion measures the spread of torsion angles. But, instead of measuring the differences between the magnitudes of the torsion angles, dispersion quantifies the agreement in the angles' directions. Thus, angular dispersion avoids the problems that can arise when standard descriptive statistics are applied to circular data.

In the absence of any experimental data, the angular dispersions of the phi and psi torsion angles were calculated across the clusters of PDB structures to establish a baseline measure of torsion angle variability in the crystals. Although not a direct measurement of how the backbone torsion angles twist in a protein crystal, this calculation offered a reasonable approximation. The most flexible regions of the structure would show the greatest variability in torsion angles across the published crystal structures and this would be reflected in the angular dispersions calculated. A comparison with the torsion angle dispersions calculated for the MD trajectories may, therefore, reveal which MD force-field most accurately models the flexibility of the protein backbone. The agreement between experiment and simulation was quantified by calculating the correlation coefficients between the angular dispersions calculated across the PDB structures and the angular dispersions of the chains over the duration of the MD simulations. The angular dispersions of the multiple chains simulated by MD were summarised by taking the median value.

The torsion angle profiles for the five proteins analysed are included in appendix C. The correlation coefficients calculated between the torsion angle dispersions for the PDB structures and the MD trajectories are presented in table 6.11. The results of the torsion angle analysis are very similar to the other methods that attempted to validate MD simulations. The correspondence between the experimental torsion angle dispersions and those derived from simulation are very weak as seen from the low Pearson correlation coefficients. The slightly higher Spearman correlation coefficients suggest that there is a degree of agreement between experiment and simulation in locating the most and least flexible regions of the structure. It could be argued, however, that the poor correlations are due to incorrectly assuming that the calculation of torsion angle dispersion across multiple PDB structures is representative of how the proteins would move. It is, of course, impossible to test this assumption by measuring torsion angle dispersions directly, but it is possible to test whether the torsion angle dispersions are consistent with backbone flexibility. To counter this criticism, the correlation between the torsion angle dispersion and B-factors was calculated. Interestingly, although not a very strong correlation, the correlation between torsion angle dispersion and B-factors is more convincing than that seen between the torsion angle dispersions derived from the MD simulations (table 6.12). Thus, the torsion angle dispersions derived from the clusters of PDB files are a reflection of the proteins' dynamics. It is conceivable that the poor agreement between torsion angle dispersions and the predictions of MD simulations may highlight inaccuracies in how MD simulations model torsion angle rotations.

Table 6.11: Psi and phi torsion angle dispersions correlation coefficients between PDB structures and MD simulation. Both Pearson and Spearman correlation coefficients are calculated for both torsion angles. The torsion angle dispersions of the MD trajectories are summarised as the median over all the chains in a 200 ns simulation of the model of the unit cell

		Correlation coefficient			
		Phi dispersion		**Psi dispersion**	
Protein	**Force field**	**Pearson**	**Spearman**	**Pearson**	**Spearman**
Hen egg white lysozyme	Amber99SB-ILDN	0.061	0.492	0.072	0.561
	OPLS-AA	0.040	0.421	0.068	0.538
	CHARMM27	-0.004	0.530	0.079	0.543
	GROMOS54a7	0.411	0.577	0.421	0.489
Human lysozyme	Amber99SB-ILDN	0.752	0.617	0.011	0.469
	OPLS-AA	0.011	0.469	0.062	0.526
	CHARMM27	0.011	0.610	0.031	0.603
	GROMOS54a7	0.149	0.596	0.186	0.599
T4 lysozyme	Amber99SB-ILDN	0.305	0.454	0.892	0.525
	OPLS-AA	0.159	0.404	0.391	0.551
	CHARMM27	0.214	0.480	0.402	0.502
	GROMOS54a7	0.075	0.512	0.306	0.534
Pancreatic ribonuclease	Amber99SB-ILDN	0.120	0.332	0.148	0.329
	OPLS-AA	0.188	0.298	0.166	0.383
	CHARMM27	0.215	0.353	0.143	0.287
	GROMOS54a7	0.201	0.357	0.325	0.329
Staphylococcal nuclease	Amber99SB-ILDN	0.153	0.461	0.201	0.509
	OPLS-AA	0.183	0.453	0.279	0.420
	CHARMM27	0.164	0.501	0.182	0.516
	GROMOS54a7	0.198	0.512	0.179	0.398

Table 6.12: Pearson and Spearman correlation coefficients between alpha-carbon consensus B-factor profiles and torsion angle dispersions derived from the clusters of PDB files. The B-factor of an alpha-carbon is the median media-mad normalised B-factor over the cluster of crystal structures collated for the protein.

	Correlation coefficient			
	Phi dispersion		**Psi dispersion**	
Protein	**Pearson**	**Spearman**	**Pearson**	**Spearman**
Hen egg white lysozyme	-0.149	-0.583	-0.238	-0.622
Human lysozyme	-0.212	-0.729	-0.292	-0.708
T4 lysozyme	-0.570	-0.484	-0.565	-0.574
Pancreatic ribonuclease	-0.530	-0.524	-0.336	-0.353
Staphylococcal nuclease	-0.601	-0.628	-0.699	-0.650

6.4.7 Qualitative analysis

Approaches to quantitatively assess how closely different measures of conformational variability reflect one another all converge at the same conclusion: there is no substantive evidence that MD simulation can reliably measure conformational dynamics in a crystal. Instead, the level of agreement between simulation and experimental data is weak. Nonetheless, inspection of the "flexibility" profiles for the proteins investigated (appendix C) suggests that there may be some consistency between simulation and experiment in the regions of the proteins exhibiting the most conformational variability. To test this further, a more qualitative approach was taken to compare the predictions of simulation and experiment. For each protein, the five most flexible residues were identified as determined by metrics derived from both MD and PDB data. The reasoning behind this was to establish whether there was any agreement in the regions of the proteins that were deemed to be the most dynamic. The results are presented in appendix D in tabular form.

Even with this qualitative analysis, there is poor level of agreement between the flexibility measures derived from PDB structures and MD simulations. Across the five proteins considered, there are only a small number of incidences where the most flexible residues coincide. Of the four MD force fields considered, Amber99SB-ILDN and OPLS-AA appear to out perform CHARMM27 and GROMOS54a7. However, since none of the force fields performed particularly well, it would be inappropriate to draw any conclusions about the validity of a particular force field based on these results. An interesting observation is the consistency between the four force fields. The MD simulations do, in general, agree with one another in the regions of the proteins that exhibit the greatest conformational variability. It may not necessarily be exactly the same residue, but there is agreement within the same local region of the protein.

For both MD and PDB data, the residues identified as having the greatest conformational flexibility are those that would be expected to be the most dynamic given their amino acid type and position within the structure. The most dynamic regions of the structures are all residues in surface exposed turns or "random coil". Often these residues are at the transition points between regular secondary structure and a stretch of residues forming a turn or an extended loop. Interestingly, 3-10 helices are often adjacent to these highly dynamic regions of the protein. This observation is consistent with the analyses of chapters 3 and 4 which showed that, of all the regular secondary structure types, residues within 3-10 helices appeared to be the most dynamic. In terms of amino acid composition, glycine and proline feature prominently in the regions identified due to their highly dynamic character and potential disruptive effects on the formation of regular structure in their vicinity.

Despite not entirely agreeing with PDB data, MD simulations always identify regions of a protein that would be expected to show conformational variability. In a sense, the simulations are "correct", but fail to reproduce the dynamics of the proteins that are entirely

consistent with the experimental data. There are two possible explanations to account for these differences. Firstly, flexibility metrics derived from PDB data, and B-factors in particular, may not accurately reflect the dynamics of the protein. Much of the information about the protein's dynamics that could be discovered by crystallography may have been lost or obscured by the refinement process. MD simulations might, therefore, provide a truer picture of the dynamics of the proteins in the crystal lattice. The alternative explanation is that the inconsistencies between PDB data and MD are due to deficiencies in the construction of the models and the simulation methods. The models employed are greatly simplified representations of a crystal lattice and do not account for all the physicochemical properties of the crystals. In addition MD force fields are not parametrised to model protein chemistry under the extreme conditions of a crystallographic experiment. MD simulations may, therefore, offer nothing better than a very coarse approximation and are unable to account for all the subtle effects that modulate protein dynamics in a crystalline environment.

6.5 Methods

6.5.1 Identifying PDB clusters

The file `XrayAndNMR.txt` listing X-ray and NMR structures sharing 95% or more sequence homology was downloaded from the Research Collaboratory for Structural Bioinformatics (RCSB) FTP repository: `ftp://resources.rcsb.org/sequence/clusters/` (October 2014). The file was parsed to identify the proteins and to count the number of structures that had been determined by X-ray crystallography. Clusters were sorted in descending order of size to identify those proteins which contained 100 or more X-ray structures. The PDB identifiers of each X-ray structure in the chosen clusters were extracted and used to download the individual PDB structure files for further analysis.

6.5.2 Screening structure files within a PDB cluster

The structure files in each cluster were parsed to find subsets of the most structurally similar crystals. All structures with R indices above 0.3 were deemed poor quality and immediately eliminated. Structures were also eliminated if the PDB file could not be completely parsed or of there were inconsistencies between the structural and sequence records. Structures were grouped according to the space group of the crystals, and only those space groups representing a significant proportion of the cluster were considered further. The presence of large ligands (molecules of 10 or more atoms) was used as a criteria for selection depending on whether it was typical for the protein to be crystallised with bound ligands. For example, lysozyme structures were selected in the absence of ligands; haemoglobin and myoglobin

were selected only with haem cofactors; and HIV-1 protease was selected irrespective of the bound ligands.

Sequence similarity between the structures in a cluster was established by comparison to a reference structure. Where possible, a high quality structure of the wild type protein was chosen as the reference. Using PDB files for sequence reference as opposed to a sequence database was necessary to account for features such as the presence of engineered expression tags or runs of unresolved residues that are common to all structures. Typically, only structures that deviated from the reference structure by no more than two amino acid substitutions were included in the cluster subset. However, this criterion was relaxed, as in the case of cytochrome c, if too many structures would be excluded. Insertions and deletions were not permitted as these engineered mutations were considered to be more disruptive than substitutions. Furthermore, although occupying equivalent positions in the protein's sequence, residues may not be spatially equivalent in the vicinity of an insertion or deletion.

6.5.3 Molecular dynamics simulations

Molecular dynamics simulations were prepared and run using the GROningen MAchine for Chemical Simulation (GROMACS) software suite (version 5.0.2) (Berendsen *et al.* 1995; Lindahl *et al.* 2001; Van Der Spoel *et al.* 2005; Hess *et al.* 2008; Pronk *et al.* 2013). The four MD force fields tested were: OPLS-AA (Jorgensen *et al.* 1996; Kaminski *et al.* 2001), Amber99SB-ILDN (Hornak *et al.* 2006; Lindorff-Larsen *et al.* 2010), CHARMM27 (MacKerell *et al.* 1998; Mackerell *et al.* 2004) and GROMOS54a7 (Schmid *et al.* 2011). Three point water models were used for all simulations. The TIP3P model (Jorgensen *et al.* 1983) was used for the all-atom force fields: OPLS-AA, Amber99SB-ILDN and CHARMM27, and the Simple Point Charge (SPC) model (Berendsen *et al.* 1981) was was used with the united atom GROMOS54a7 force field. GPU acceleration was enabled by compiling GROMACS with support for NVIDIA's CUDA library (version 5.5). GROMACS was compiled and used with single floating point precision.

All MD simulations of the crystals were run under the NVT ensemble. Temperature equilibration was achieved with the velocity scaling thermostat and a coupling constant of 0.1. The temperature of the proteins in the simulations was maintained separately to the solvent through the use of two temperature coupling groups. No pressure coupling was applied to ensure the dimensions of the unit cell and structure of the crystal lattice remained constant throughout the simulation. The molecular dynamics integrator used a step size of 2 fs and bond elongation was corrected by setting all bonds as constraints and applying the one iteration of the LINCS algorithm with an order parameter of 4. Periodic boundary conditions were applied in every direction to approximate the structure of the crystal lattice. Molecules were defined as periodic to allow proteins to interact with themselves across the boundaries of the simulation box. The linear velocity of the system's centre of mass was corrected every

100 steps.

The potential function for short-range VdW and electrostatic interactions applied an initial cutoff value of 1 nm. Energy and pressure corrections were applied for long range dispersion interactions. Particle Mesh Ewald (PME) was used for long range electrostatic interactions with an initial Fourier grid spacing of 0.12 nm and cubic interpolation of the PME calculations. Lists of neighbouring atoms were maintained with a grid search and the initial update frequency was set to every 20 steps. The buffered Verlet algorithm dynamically adjusted the update frequency for the lists of interacting atoms and VdW and electrostatics parameters to improve performance whilst maintaining accuracy.

Preparing the simulation system

Reference PDB files were chosen to build the simulation systems. The structures chosen were: `1HEL` (hen egg white lysozyme) (Wilson *et al.* 1992), `1LZ1` (human lysozyme) (Artymiuk and Blake 1981), `3LZM` (T4 lysozyme) (Matsumura *et al.* 1989), `1KF5` (ribonuclease) (Berisio *et al.* 2002), `1STN` (staphylococcal nuclease) (Hynes and Fox 1991). The PDB files were processed to eliminate all molecules other than the proteins and to set all atoms in their highest occupancy locations. In the case of staphylococcal nuclease, the missing N and C-terminal residues were added using the modeller (Sali and Blundell 1993) software suite following the same methodology described in chapter 3. The `pdb2gmx` tool was used to generate the protein topology, position restraint and coordinate files for a given force field and water model combination. The crystal symmetries of the original PDB file were applied to the coordinate files to recreate the arrangement of proteins in the unit cell. The `editconf` tool was used to set the dimensions of the MD simulation box to coincide with the unit cell by reading the lengths and angles that define the unit cell from the `CRYST1` record of the PDB file. In addition, in the case of T4 lysozyme, it was necessary to apply the rotation defined by the `ORIGX` records to align the lattice basis vector **a** with the basis vector **x** of the left-handed Cartesian coordinate system assumed by GROMACS.

To eliminate steric clashes arising from the creation of the unit cell, steepest descent energy minimisation of the unit cell in a vacuum was run for a maximum of 10 steps or until all forces between atoms fell below $1000\,\mathrm{kJ\,nm^{-1}\,mol^{-1}}$. Water molecules were added to fill the cavities between the proteins using the `solvate` tool and the GROMACS default three point water coordinate file `spc216.gro`. The system was made electrically neutral with the `genions` tool which substituted water molecules for either sodium or chloride ions. Unfavourable interactions inadvertently introduced during the solvation of the unit cell were corrected through two passes of energy minimisation. The first round used the steepest descent algorithm to quickly bring all forces under $1000\,\mathrm{kJ\,nm^{-1}\,mol^{-1}}$. Energy minimisation then switched to use the conjugate gradient algorithm with the maximum force set to $100\,\mathrm{kJ\,nm^{-1}\,mol^{-1}}$. The solvent was relaxed and the system equilibrated to a temperature

of 300 K with a restrained MD simulation. Positional restraints were applied to the protein molecules and the temperature of the system was raised from zero to 300 K over 100 ps. With the restraints still in place, the system was equilibrated for a further 100 ps at 300 K.

The production MD simulation was run as a continuation of the equilibration simulation with the removal of the restraints on the protein molecules. The system was maintained at 300 K and run for a maximum of 200 ns. Coordinates were recorded every 2 ps to a precision of 10^{-5} nm.

Simulations of proteins in solution

Simulations of single proteins in solutions were prepared and run following a similar methodology to that described above. The structure of the asymmetric unit was used as the reference structure for all simulations. The proteins were placed at the centre of a dodecahedral simulation box that was sized so that the minimum distance between the farthest extent of the protein and the edges of the box was 1.3 nm in all directions. Empty space was filled with water molecules and the minimum number of chloride ions to neutralise the overall charge of the system. Energy minimisation and equilibration was undertaken identically to the simulations of the crystal lattices. The production simulation was run for 200 ns under the NVT ensemble to allow direct comparison with the simulations of the crystal lattices. Despite modelling "simpler" systems, the simulations of the single proteins contained more atoms than the corresponding crystal simulations (5.8 times the number in the case of ribonuclease). Simulations were, therefore, more computationally intensive in terms of both processor time and storage requirements for the trajectory coordinate data.

Analysing MD trajectories

The molecules of the MD trajectories were made whole and abrupt translations across the simulation box were corrected using two passes of the `trjconv` tool. All molecules except for the proteins were discarded and the trajectory was subsampled to record the coordinates every 20 ps, resulting in trajectories of 10000 frames. The trajectories were then split to create separate files for each protein chain. Each chain was then aligned by rotation and translation to minimise the mean square deviations in the positions of the alpha-carbon atoms with respect to a reference structure. The reference structures used were the coordinates of the chains at the beginning of the production simulation (the first frames of the trajectory files). Python scripts were written to analyse the aligned trajectories of the protein chains. The python library `MDAnalysis` (Michaud-Agrawal *et al.* 2011) was used to parse the GROMACS trajectory files.

The MSF of the alpha-carbon atoms for each chain were calculated using a python script

that calculated the mean square displacement in coordinates of the atoms with respect to the reference structure (equation 6.1). MSF was calculated using coordinates every 10 frames in the trajectory; that is, sampling 1000 frames at intervals of 200 ps.

$$MSF_i = \frac{1}{N} \sum_{n=1}^{N} |\mathbf{R}_i(n) - \mathbf{R}_i(0)|^2 \tag{6.1}$$

where MSF_i MSF of the ith atom
N total number of frames in the trajectory
$\mathbf{R}_i(n)$ coordinates of the ith atom at frame number n in the trajectory

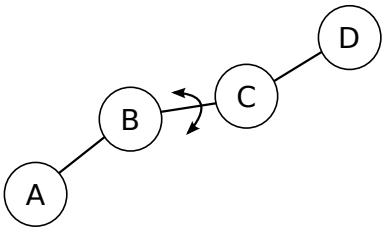
6.5.4 Calculation of MSF for crystal structures

Crystal structures were aligned to minimise the sum of the square deviations in the positions of the alpha-carbon atoms by the Kabsch algorithm (Kabsch 1976, 1978). MSF was calculated analogously to MD trajectories with equation 6.1. The reference structure used to align all other structures was the centroid of the cluster with respect to the square deviations in the positions of the alpha-carbons. The centroid was identified by first calculating the sum of the square distances between alpha-carbons coordinates after alignment for every pair of structures in the cluster. For each structure, the sums were added together across all the structure's pairings and the structure with the lowest total was defined as the centroid of the cluster.

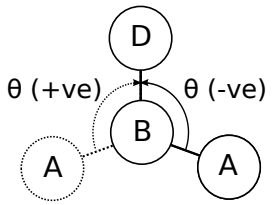
6.5.5 Torsion angle calculations

Two torsion angles, phi and psi, were calculated from the coordinates of the atoms of a protein's backbone. All calculations followed the definition of torsion angle recommended by International Union of Pure and Applied Chemistry (IUPAC) (Moss 1996) as presented in the on-line "Gold Book" (Nic *et al.* 2014) based on the original "Gold Book" compendium (McNaught and Wilkinson 1997) (see figure 6.11). The phi and psi torsion angles were defined following convention and are illustrated in figure 6.12. Calculation of the phi torsion angle of an amino acid requires knowledge of the coordinates of the carbonyl carbon of the previous amino acid. Thus, it is not possible to calculate the phi torsion angle for the first amino acid in a chain. Similarly, because the calculation of the psi torsion angle of an amino acid requires the coordinates of the amine nitrogen of the following amino acid, there is no psi torsion angle for the last amino acid in a chain. For convenience, the first phi and last psi torsion angles were always set to zero.

Figure 6.11: IUPAC definition of torsion angle. Atoms A, B, C and D are atoms bonded in a linear sequence. The torsion angle θ for the bond B-C is the angle of rotation that would result in the alignment of bond A-B with bond C-D when viewed along bond B-C. The absolute value of the torsion angle is restricted to the range -180–$180°$ with a positive and negative values corresponding to clockwise and anticlockwise rotations respectively.

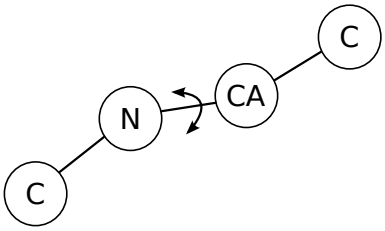


(a) Generalised chain of bonded atoms.

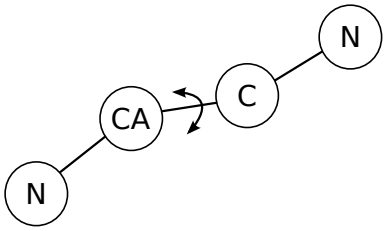


(b) Torsion angle for bond B-C (dashed atom indicates a positive value).

Figure 6.12: Definitions of torsion angles using a schematic representation of the protein backbone. The labels are as follows: CA : alpha carbon, N : amine nitrogen and C : carbonyl carbon. Amino acid side chains, all hydrogens and the carbonyl oxygen have been omitted for clarity.



(a) Phi torsion angle.



(b) Psi torsion angle.

Variability of torsion angles across PDB structures and frames of MD simulations were quantified by using the metric of angular dispersion as defined in equation 6.2 adapted from Zar (2010) chapter 26, Circular Distributions: Descriptive Statistics.

Let $\theta_1, \theta_2, \ldots, \theta_N$ be N torsion angles. Then, the angular dispersion r is defined as:

$$\begin{aligned} r &= \sqrt{A^2 + B^2} \\ \text{where} \quad A &= \frac{1}{N}\sum_{i=1}^{N} \sin\theta_i \\ \text{and} \quad B &= \frac{1}{N}\sum_{i=1}^{N} \cos\theta_i \end{aligned} \tag{6.2}$$

Angular dispersion is restricted to values in the range zero to one. A value of zero corresponds to a uniform distribution of torsion angles over the full range of angles (-180–$180°$). A value of one corresponds to no variation; that is, a sequence of equal torsion angles.

Chapter 7

Conclusions

The overall aim of this thesis was to re-evaluate the usefulness of crystallographic atomic displacement parameters (ADPs) as measures of protein conformational dynamics since current opinion is divided on this issue. Many structural bioinformaticians continue to make use of the ADP data deposited in the PDB whilst others question the value of these crystallographic data and use alternative methods to quantify protein flexibility. Irrespective of personal opinion on this issue, there are two fundamental questions that need to be answered in order to assess whether ADPs give a true reflection of the dynamics of a protein in a crystal:

1. Is there a correspondence between ADP values and those regions of a protein's structure that are known to be flexible or rigid?

2. What metrics of protein flexibility are the most suitable to quantitatively validate or discredit ADPs?

Although these two questions are closely interlinked and difficult to separate, both have been addressed by research presented in this thesis. Firstly, the question of whether ADPs are a true reflection of conformational variability was examined by attempting to establish relationships between ADP values and static structural features that are widely accepted to be correlates of protein flexibility. The scope of the investigation then broadened to assess computer modelling as an alternative approach to validating ADPs. It was hoped that, by not relying on one particular method of validating ADP values, this would give a more balanced assessment of ADP data. The conclusions from these two strands of research are discussed below.

7.1 ADPs as measures of protein flexibility

Chapter 3 re-examined the usefulness of isotropic B-factors as measures of conformational variability in crystal structures. The results were broadly consistent with what might have been expected from previous studies in this area (Sheriff *et al.* 1985; Carugo and Argos 1997; Parthasarathy and Murthy 1997; Smith, Radivojac *et al.* 2003; Radivojac *et al.* 2004; Yuan *et al.* 2003; Zhang *et al.* 2009; Sonavane *et al.* 2013). The atoms in a crystal structure resolved with the highest isotropic B-factors are typically located within regions of the structure that would be assumed to be highly flexible; whilst atoms with the lowest B-factors typically coincide with the most conformationally restrained elements of the structure. What was surprising, however, was that there were many exceptions to this general principle, making it impossible to establish reliable qualitative relationships between structural features of a protein and isotropic B-factor values.

It is conceivable that there may be errors in published PDB structures and this may be reason why the B-factor data failed to give convincing results. Recently, Touw and Vriend (2014), have developed a crystallographic database, the BBD, to check and correct any inconsistencies in the B-factor data published by the PDB. Whilst it would be interesting to repeat the analysis using structural data derived from the BDB, it is unlikely that the results would change significantly. The proportion of structures corrected by the BDB is small (less than 10%) and, because the B-factor data was normalised, some of corrections made by the BDB would have little or no effect on the outcome of the analysis. In conducting the research, only the highest quality structures were selected and the periodic nature of the crystal lattice was fully accounted for in all structural analyses. The absence of any clear correlations between B-factor values and structural conformational variability cannot, therefore, be attributed to anomalies in the crystallographic data or flaws in the research methodology. Instead, this research suggests that it may now be time to question the widely-held belief that B-factors can be assumed to be an indirect quantitative measure of protein flexibility.

Although a clear link between isotropic B-factors and protein flexibility/rigidity could not be established, this does not necessarily mean that crystallographic data cannot be used to measure conformational dynamics. Chapter 4 explored the possibility that, perhaps, it was the assumption that underpins the derivation of isotropic B-factor values that was at fault. It is conceivable that many atoms in a crystal structure do not vibrate with perfect spherical symmetry about their average positions. Instead, a more realistic anisotropic model of atomic motion might reveal clearer relationships between conformational flexibility and crystallographic data. Consequently, chapter 4 repeated the analysis of chapter 3 using anisotropic atomic displacement parameters (AADPs) in place of isotropic B-factors .

Contrary to what might have been expected, the results from the analysis of AADPs were

no better than those of isotropic B-factors. Despite the structures refined with AADPs being of higher quality and near atomic resolution, the correlation between AADP values and protein conformation flexibility remained tenuous. What was surprising, however, was that the choice of AADP used in the analysis was irrelevant. The same weak correlations were observed irrespective of which AADP was chosen to measure conformational flexibility. The inability of AADP to probe molecular motion with any greater precision than that currently offered by isotropic B-factors may tell us something about the limitations of X-ray crystallography to measure conformational dynamics. In theory, the AADP of high resolution crystal structures should give a far more reliable and accurate picture of the movements of atoms than what was observed. One possible explanation for this disparity is that current crystallographic methods have reached the limit of what can be measured experimentally. Thus, there is no advantage of measuring conformational flexibility using AADPs over isotropic B-factors since the majority of X-ray diffraction experiments are not yet sufficiently sensitive to distinguish between these two ADPs. Although there is only the work in this thesis to support these claims, it is interesting to note that research on AADPs to measure protein flexibility is in a minority in comparison to isotropic B-factors. This may be partly due isotropic B-factors being a more familiar measure of flexibility and there being more software tools available for analysis. Nevertheless, it is also feasible that the lack of published research on AADPs is due to the fact that previous research in this area has also failed to break new ground. Unfortunately, since negative results are rarely published in the literature, it is very difficult to test this hypothesis.

Despite a scarcity of research, the literature does give some indirect evidence that there is little to gain by using AADPs in place of isotropic B-factors as measures of conformational flexibility. New research in this field is moving away from classical X-ray crystallographic methods and is exploring the possibilities offered by new technologies. In particular, the Fraser lab have been advocating new ways to undertake and interpret crystallographic experiments to capture protein dynamics at the atomic scale. Fraser and colleagues argue that traditional crystallographic experiments may give us a misleading picture of protein conformational dynamics having demonstrated that the process of cryocooling alters the conformational distributions of approximately 35% of side chains in a sample of 30 proteins (Fraser *et al.* 2011). Ambient temperature crystallography and new computational methods to extract previously "hidden" conformational states in electron density maps are the two main avenues the group have followed in order to address the limitations of classical crystallography (Fraser *et al.* 2009, 2011; Lang *et al.* 2010, 2014; Woldeyes *et al.* 2014).

Although not a primary objective, one positive result obtained from this research was to demonstrate the benefit of normalising ADP values as opposed to using raw data in analyses. Normalisation of ADP values is widely accepted as necessary when comparing crystal structures, but there is an absence of any quantitative evidence in the literature to justify its use. By devising a simple metric to compare normalisation methods, this thesis has estab-

lished that normalisation does significantly reduce level of the "noise" inherent in ADP data. Furthermore, this thesis provides evidence to justify the choice of using particular normalisation methods over others. It was found that median-mad and z-normalisation consistently outperformed min-max scaling. This could potentially be an area for further research since, to date, there has been no work to systematically compare the relative effectiveness of all the different normalisation methods currently employed.

7.2 Validating ADPs using computer modelling

This thesis has explored various computational methods for modelling the conformational dynamics of proteins within a crystal lattice. Investigations began, in chapter 5, with the application of ENMs. Similar to the classical analyses of chapters 3 and 4, using high quality crystallographic data and fully accounting for the symmetry of the crystal lattice did not result in particularly strong correlations between ADP values and the predictions of the ENMs. Although the ENM were simplistic in comparison to some of the more sophisticated models that have been developed, refining the ENMs is highly unlikely to yield better results. The correlation coefficients between ADPs and the predictions of ENMs from the literature are typically in the range 0.5-0.6 so the consensus view is of a weak correlation (Kundu *et al.* 2002; Eyal *et al.* 2007; Kondrashov *et al.* 2007; Xia and Wei 2013; Opron *et al.* 2014). This is not to say that the computer models are incorrect, but as suggested previously, the limiting factor may be the imprecision of current ADP data. Thus, attempting to develop ENMs that better reproduce current ADP datasets may be a futile exercise.

The conclusion from chapters 3 to 5 initially suggested that the possibilities for new research into ADP datasets have been exhausted. However, chapter 6 considered whether it might be possible to derive high quality ADP data by attempting to eliminate the experimental "noise" in the data. The approach taken was to consider proteins whose crystal structures had been derived multiple times. By averaging the ADP data across these structures, it was hoped that experimental errors would be eliminated and the resulting consensus profile would be an accurate reflection of protein dynamics. Although limited by the data available, the results do support this hypothesis. There is consistency in the ADP profiles between independent crystallographic experiments meaning that ADP values are related to the protein's structure and are not simply artefacts of the refinement process. Establishing whether these consensus ADP profiles are a true reflection of protein dynamics proved to be a more difficult question to answer.

The numbers of protein structures for which consensus ADP profiles could be derived was limited and could not be considered to be a representative sample of all the structures deposited in the PDB. Therefore, repeating the classical analysis of chapters 3 and 4 on such a small sample of proteins was not considered to be worthwhile. Furthermore, assessing

the simplistic ENMs of chapter 5 using such a small sample size would not be particularly insightful. Consequently, the more sophisticated, but computationally more demanding, molecular dynamic simulations were chosen as a means of assessing the consensus ADP data. In addition, the opportunity was taken to compare the ADP data to the predictions of different MD force-fields. It was envisaged that this would reveal which, of the most commonly used MD force-fields, would best reproduce the ADP data. This could have potentially have been a new method of evaluating MD force-fields.

There has been some previous work validating MD force fields by simulating protein crystals and comparing to B-factor profiles. However, previous studies had drawbacks in that they only looked at one protein and compared the simulations to limited B-factor data. Eastman *et al.* (1999) investigated bovine pancreatic trypsin inhibitor; Cerutti *et al.* (2010) a scorpion toxin; Hu and Jiang (2010) lysozyme; Xue and Skrynnikov (2014) ubiquitin; and Kuzmanic *et al.* (2014) the villin headpiece. With the exception of Xue and Skrynnikov, these previous studies were all based on simulations shorter than the 200 ns simulations presented in this thesis. The approach of using multiple proteins and validating the against consensus ADP data is, therefore, novel and not simply reproducing the work of others.

Surprisingly, the MD simulations showed only a partial agreement with the consensus ADP profiles. There was little to differentiate between the different force fields so, on the basis of this research, there is nothing to suggest that one force field is significantly better at modelling protein dynamics. This is consistent with the outcomes of the work by Cerutti *et al.* (2010) and Hu and Jiang (2010) which come to different conclusions about which are the superior force fields for simulating protein crystals. Failure to clearly identify the "best" MD force field suggests that there is still some way to go before bioinformaticians can be entirely confident in the accuracy of MD simulations. This assessment, of course, is purely based on the criterion of how well the force fields model dynamics in of the crystal lattice. Nevertheless, it is not unreasonable to assume that, if the simulations do not adequately model dynamics in the crystal, there may be deficiencies when simulating proteins in other environments. The discrepancies between crystallographic data and MD simulations has been commented on previously. Studies have found that the atomic fluctuations observed in MD simulations are far greater than would be expected from ADP data (Kuzmanic *et al.* 2014; Xue and Skrynnikov 2014). Nonetheless, at this point in time, it is difficult to know whether it is the MD simulations or the crystallographic data that is closer to reality.

There is the possibility, however, that the weak relationship between the consensus ADP profiles and the MD simulations may be due to the incorrect assumption that the ADP profiles are a true reflection of the protein dynamics. Averaging ADP profiles will suppress the level of noise in the data sets but it may also act as a filter to eliminate some of the essential features of protein's dynamics. In effect, averaging could result in misleading picture of protein dynamics by only preserving the general trends that are present across multiple crystal structures. Consequently, the consensus ADP profiles might only be describing atomic

motion at a very low resolution. Thus, the consensus ADP profiles and the predictions of the MD simulations may not be directly comparable. It would be interesting to investigate the consensus ADP profiles in more detail in order to determine how they compare to the different types of dynamics exhibited by proteins. The consensus profiles may represent the global "slow" dynamics of the protein molecules rather than the fast atomic fluctuations captured by the nanosecond time scales of the MD simulations. It may be possible to test this hypothesis by comparing the consensus ADP profiles to the dynamics modelled by MD simulations over millisecond time scales. Whilst this would be difficult to achieve with all-atom simulations, it may be possible to develop coarse-grained simulations of protein crystals to model the slower global dynamics.

An alternative approach to validating MD simulations using crystallographic data has been used by Wall *et al.* (2014). Instead of ADP data, diffuse X-ray scattering was used as an indicator of conformational dynamics in a crystal. Interestingly, the research used the same protein, staphylococcal nuclease, as used in this thesis but achieved near perfect agreement between and experiment and simulation. The strategy differed, through, in that the MD simulations were used to generate expected scattering intensities which were compared to the experimental data. Thus, Wall *et al.* were using "primary" crystallographic data which may account for the better results. Yet, before rejecting ADPs in favour of raw diffraction data, the Wall *et al.* study only focused on a single protein and it is unknown whether experiments on other protein crystals would be equally successful. Nevertheless, validating MD data with diffraction signals is under-explored and could prove to be a fruitful area for further research.

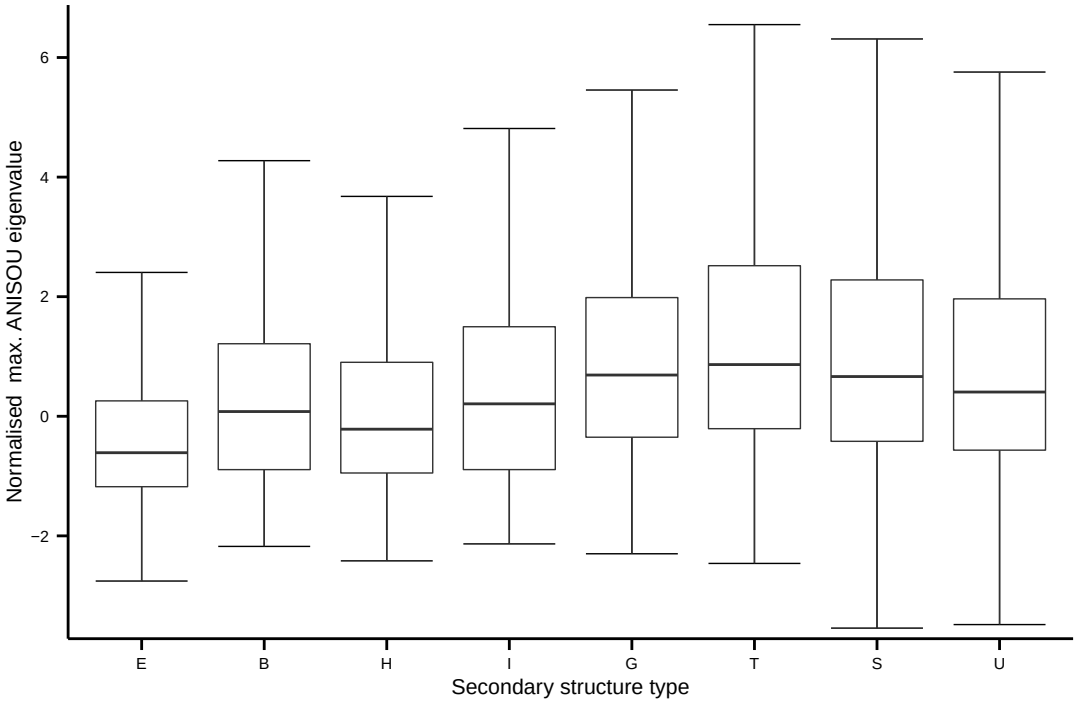
7.3 Summary

Crystallographic ADP data can tell us something about the dynamics of proteins within crystals. However, the extent to which ADP data, in general, can be trusted as a reliable indicator of protein flexibility or rigidity remains unclear. Molecular biologists should, therefore, be highly sceptical of any inferences made about protein dynamics based on ADP data alone. The limitations of ADP data does not mean that crystallographic data has no value in measuring protein dynamics. On the contrary, it can be reveal a great deal about the molecular mechanisms that drive conformational changes when used in conjunction with other experimental methods and computer modelling. In a recent review of the current state of measuring protein conformational dynamics, van den Bedem and Fraser (2015) optimistically describe a synergistic relationship between NMR, crystallography and computer simulation. The research presented in this thesis suggests that we may have reached the limit of current crystallographic methods; however, advances in technology can only improve the resolution at which the movements of proteins can be probed by crystallography.

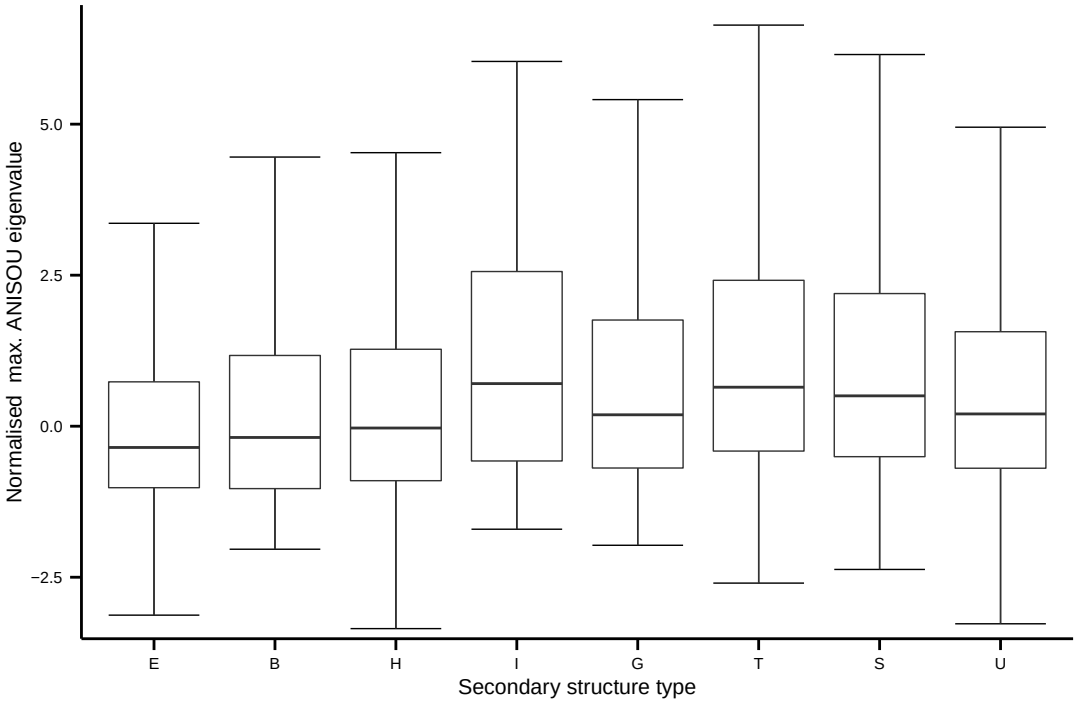
Appendix A

Anisotropic atomic displacement parameter data

Figure A.1: B-factors grouped according to secondary structure. Secondary structure labels are the DSSP classifications: E : β (extended); B : β (bridge); H : α-helix; I : π-helix; G : 3-10 helix; T : turn; S : bend; and U : unclassified ("coil").

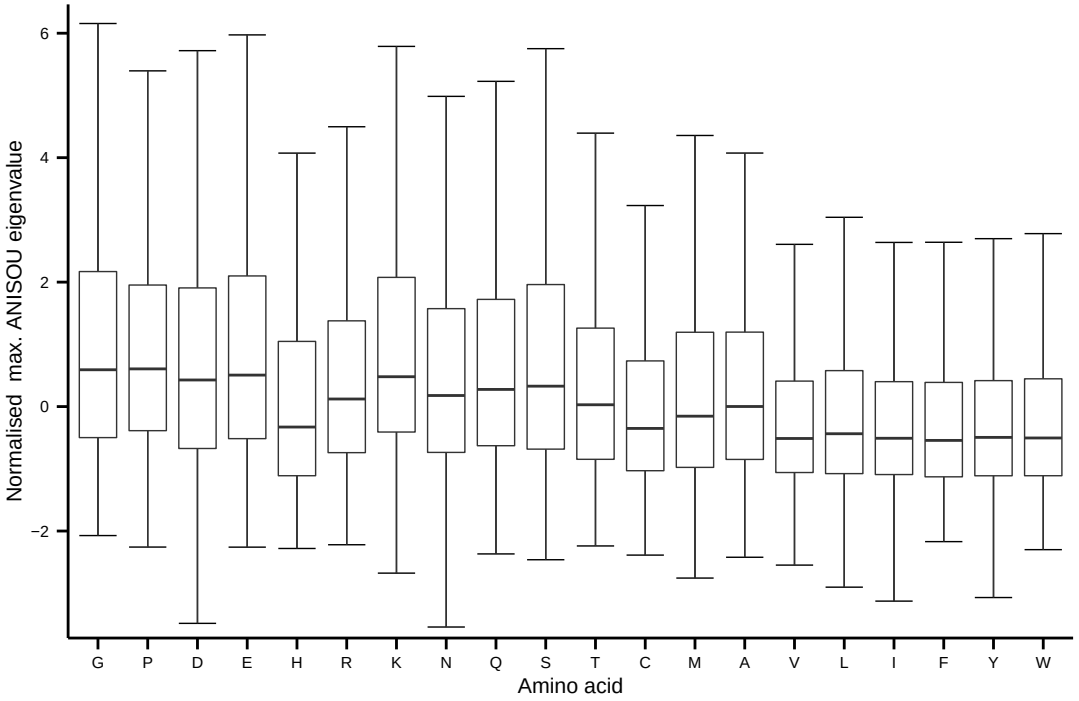


(a) Alpha-carbon AADPs. The proportion of outliers was less than 8% for all groupings except for π-helix (12.6%)

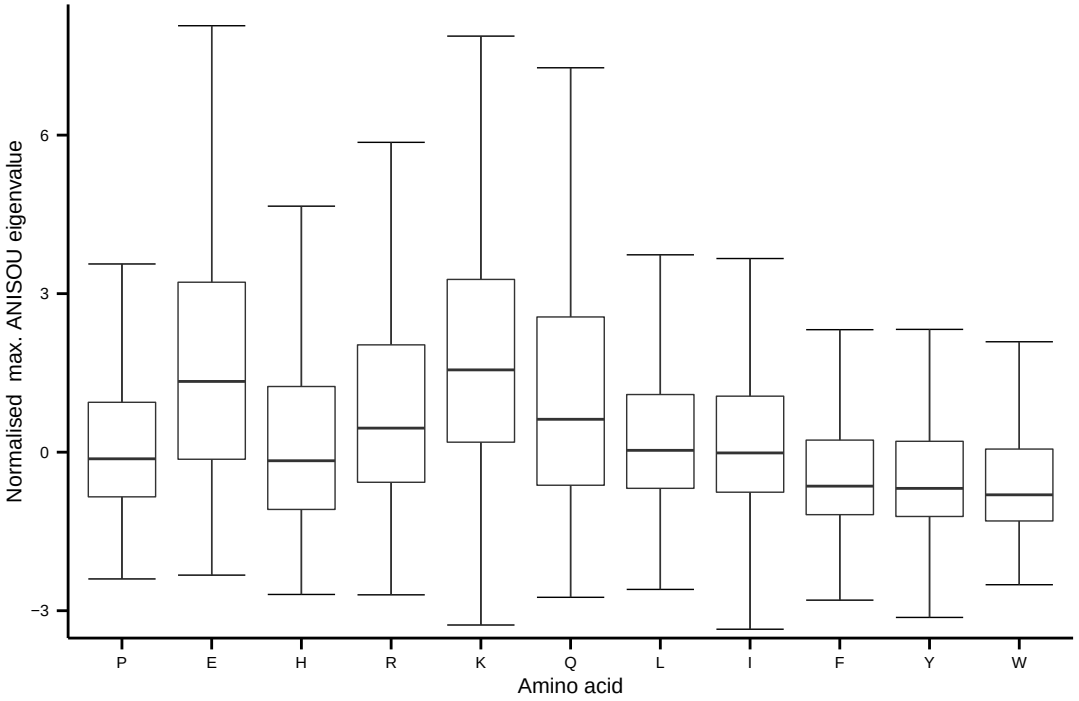


(b) Delta-carbon AADPs. The proportion of outliers was less than 8% for all groupings.

Figure A.2: Boxplots of AADPs grouped according to amino acid type



(a) Alpha-carbon AADPs. The proportion of outliers was less than 7% in all groupings except for methionine (9.5%) and arginine (7.5%).



(b) Delta-carbon AADPs. The proportion of outliers was less than 6% in all groupings.

Figure A.3: Boxplots of alpha-carbon B-factors grouped according to normalised SASA for the amino acid. The bin width is 0.05 units except for the final bin (0.9 to 1.0). The proportions of all outliers were less than 8% in each grouping except at 0.8 (11.5%).

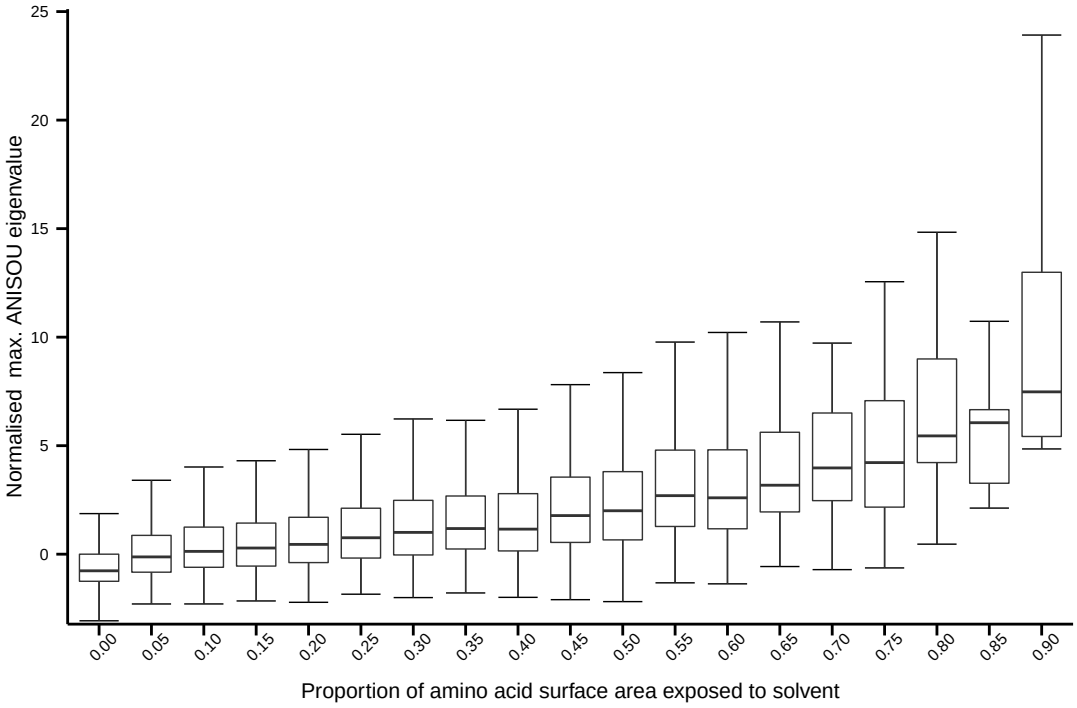


Figure A.4: Boxplots of alpha-carbon B-factors grouped according to the atom's distance from the surface. The bin width is 0.5Å. The proportions of all outliers were less than 8% in each grouping except < 0.5Å (8.4%) and $7.5 - 8.0$Å (12.9%)).

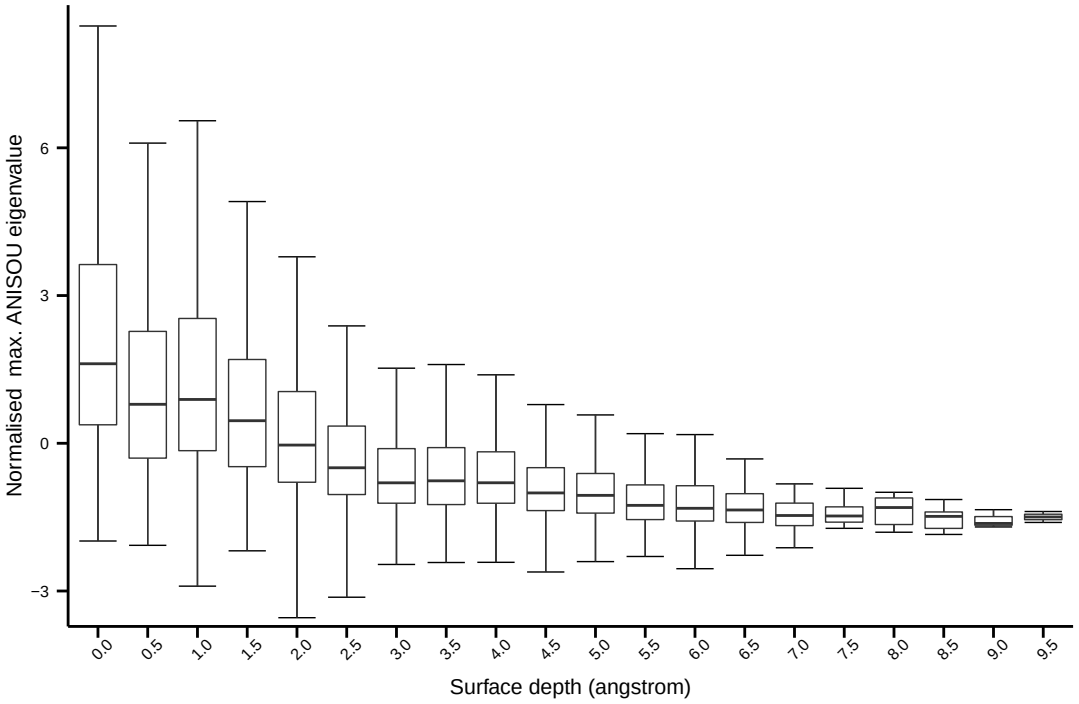


Figure A.5: Boxplots of alpha-carbon B-factors grouped according to the atoms's distance to the protein's COM. The bin width is 1Å except for the final bin ($\geq$ 45Å). The proportions of all outliers were less than 10% in each grouping in the range $2 - 35$Å and up to 18% otherwise

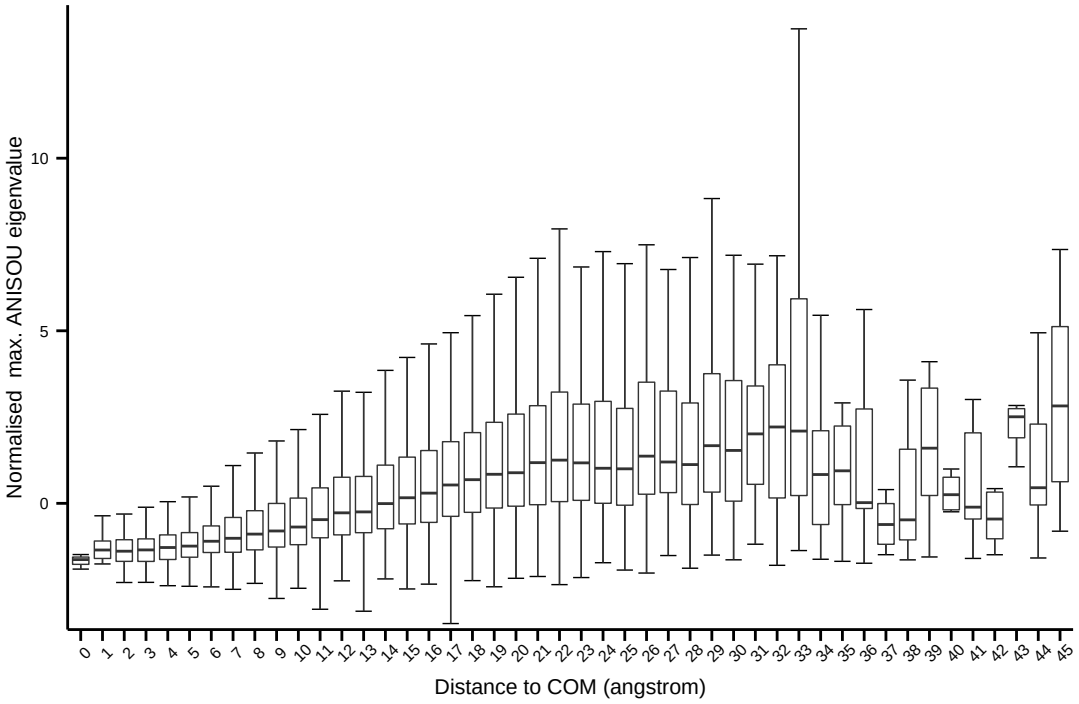


Figure A.6: Distribution of alpha-carbon to alpha-carbon distances for the maximum occupancy protein structures of the dataset. Lower and upper quartiles are represented with error bars to indicate the spread of proportions.

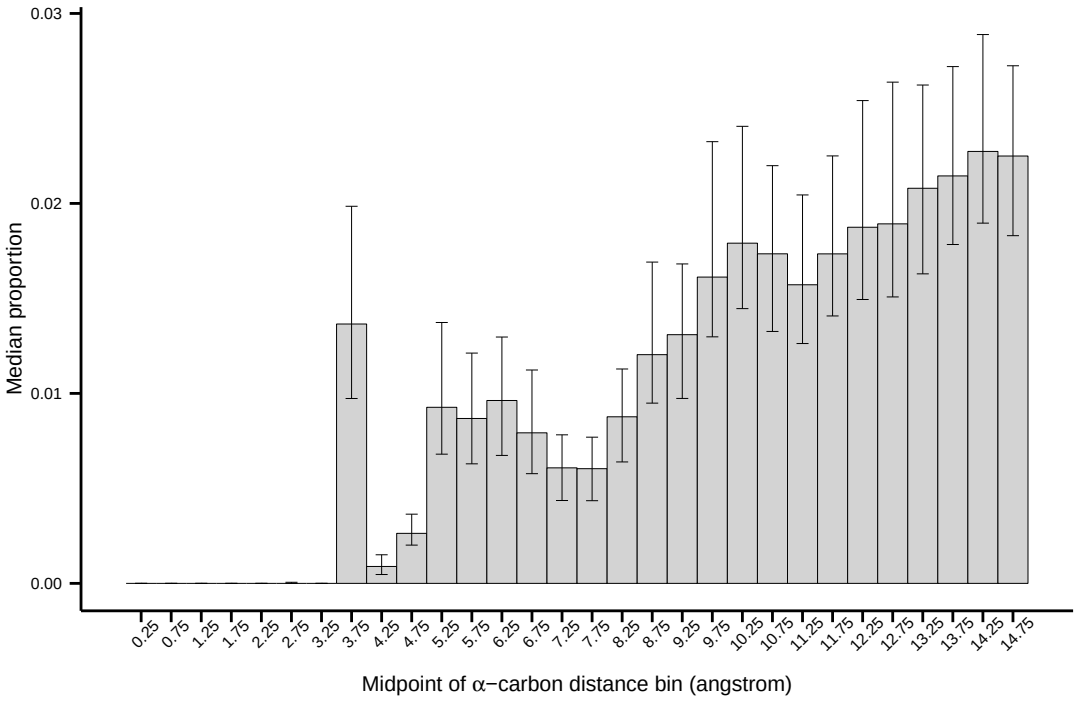
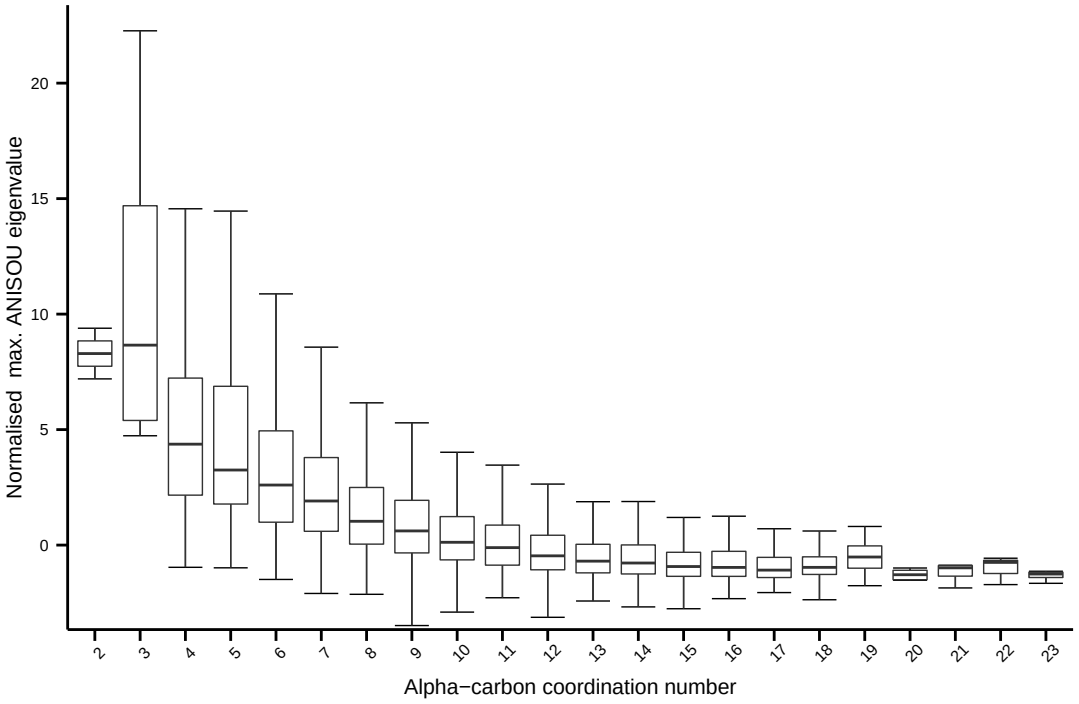


Figure A.7: Boxplots of alpha-carbon B-factors grouped according to the coordination number of the amino acid.The proportions of all outliers were less than 6.5% in each grouping except at 19 (9.5%).



Appendix B

Consensus B-factor profiles

Figure B.1: Consensus B-factor profile for T4 lysozyme

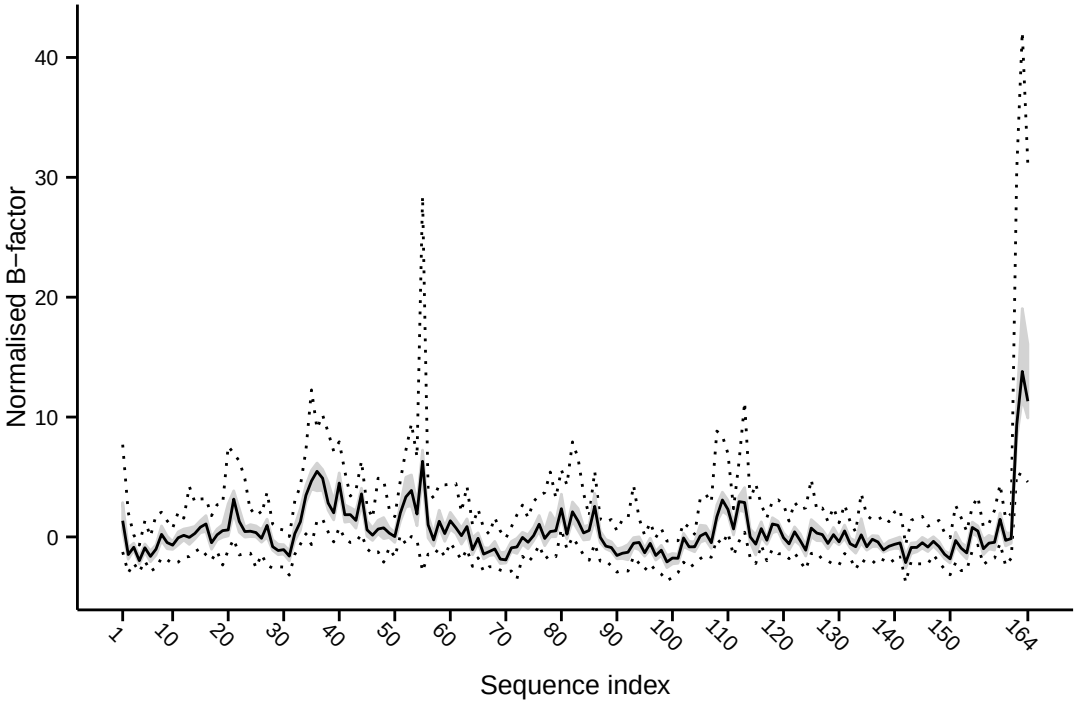


Figure B.2: Consensus B-factor profile for human lysozyme

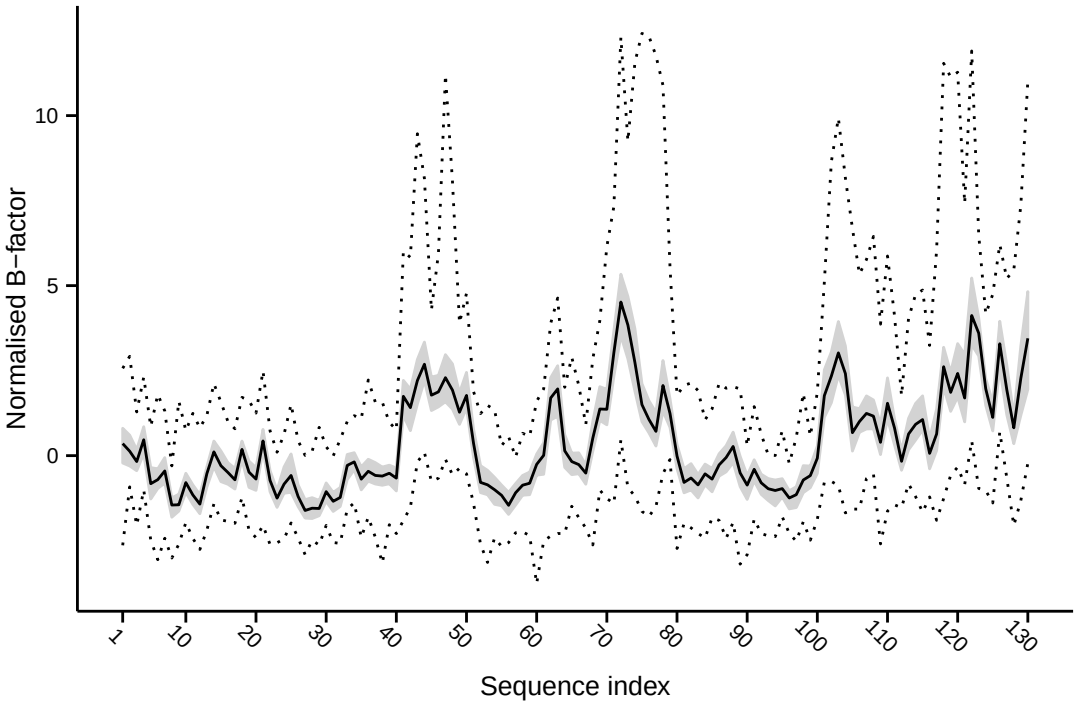


Figure B.3: Consensus B-factor profile for Staphylococcal nuclease

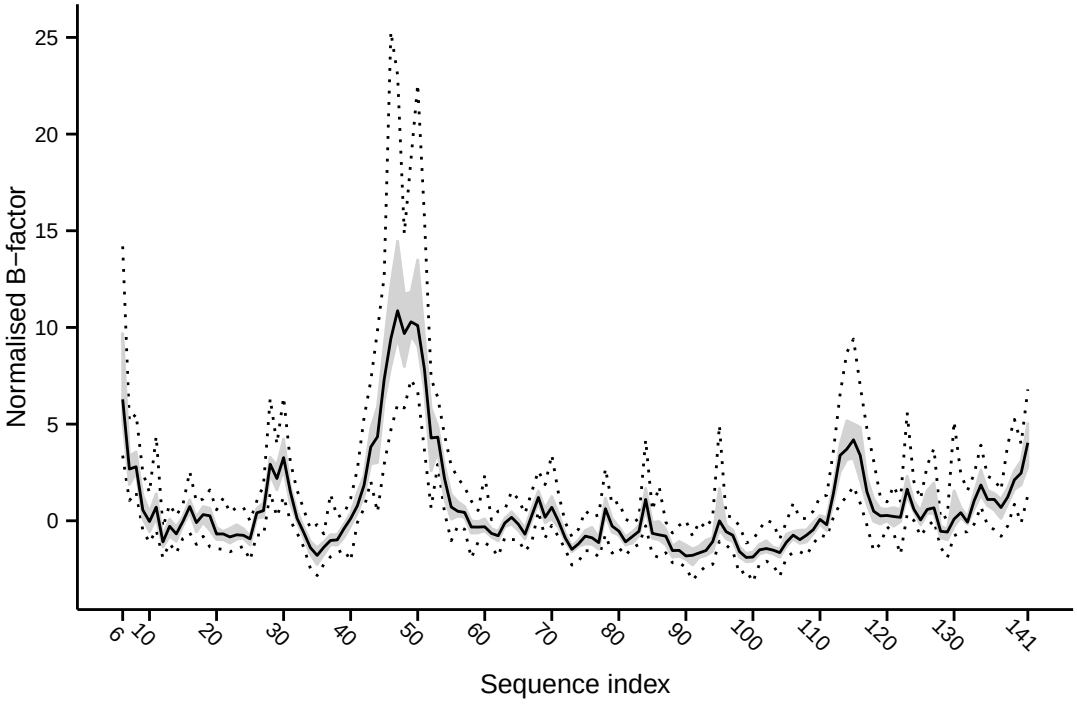
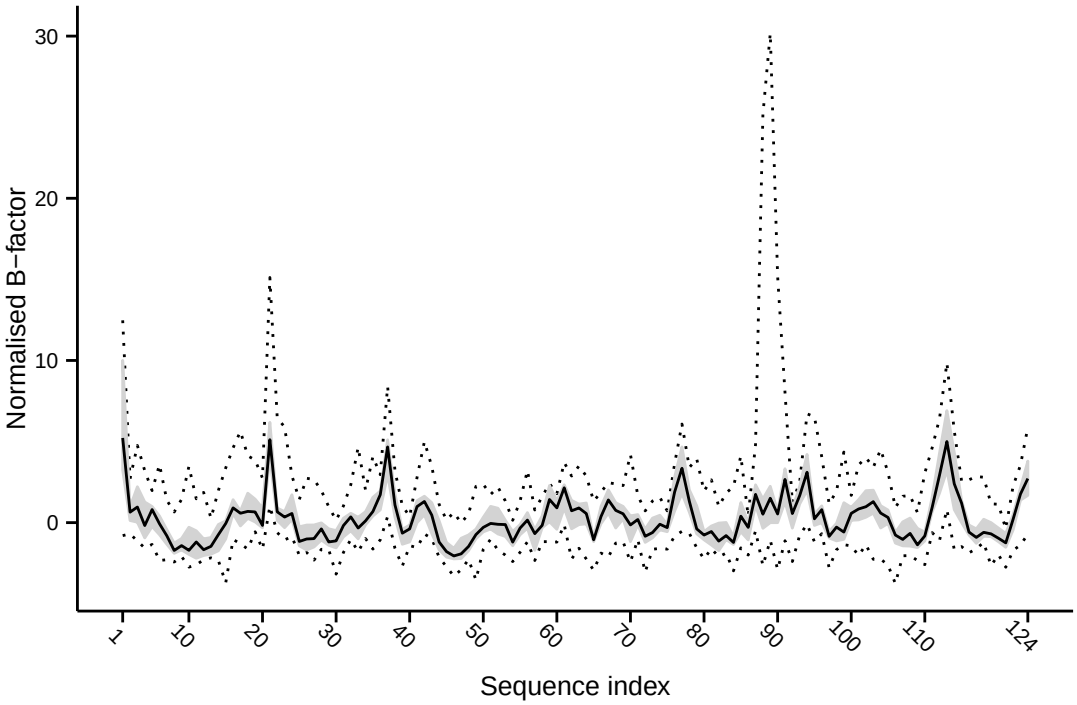
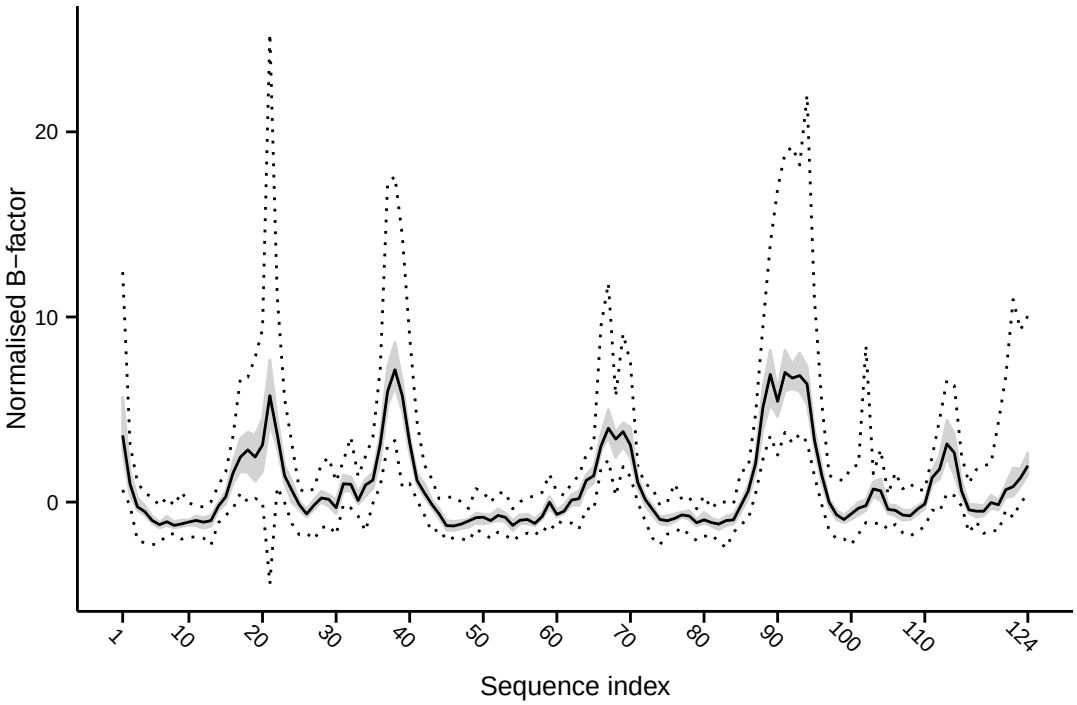


Figure B.4: Consensus B-factor profiles for pancreatic ribonuclease

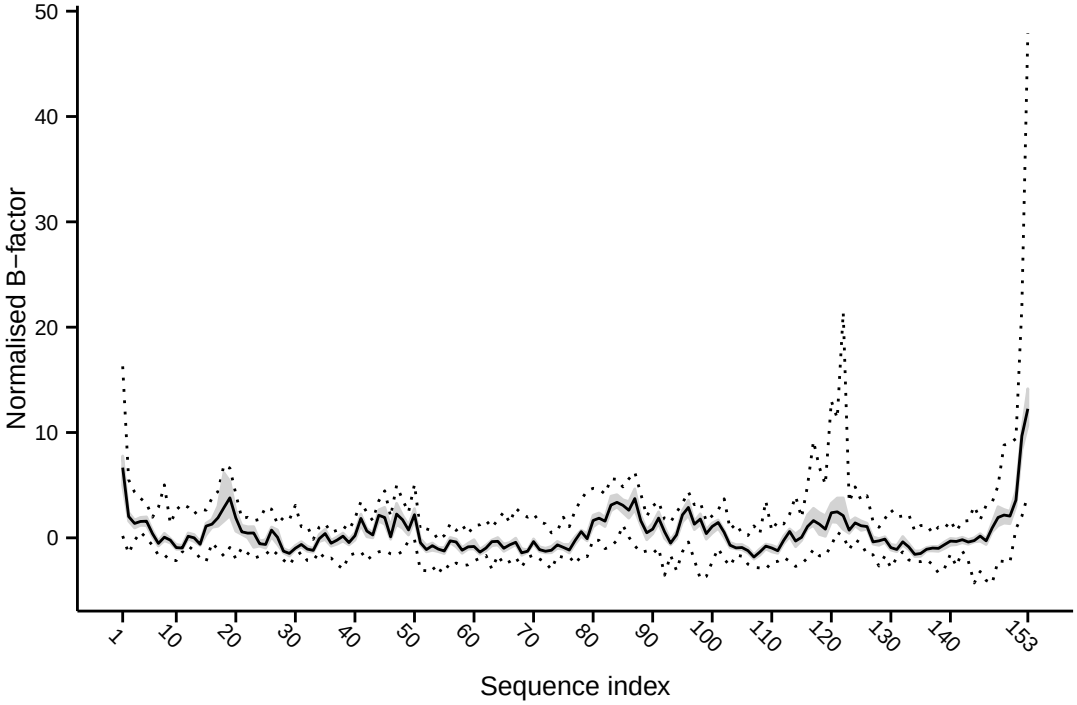


(a) Structures in space group P 1 2_1 1

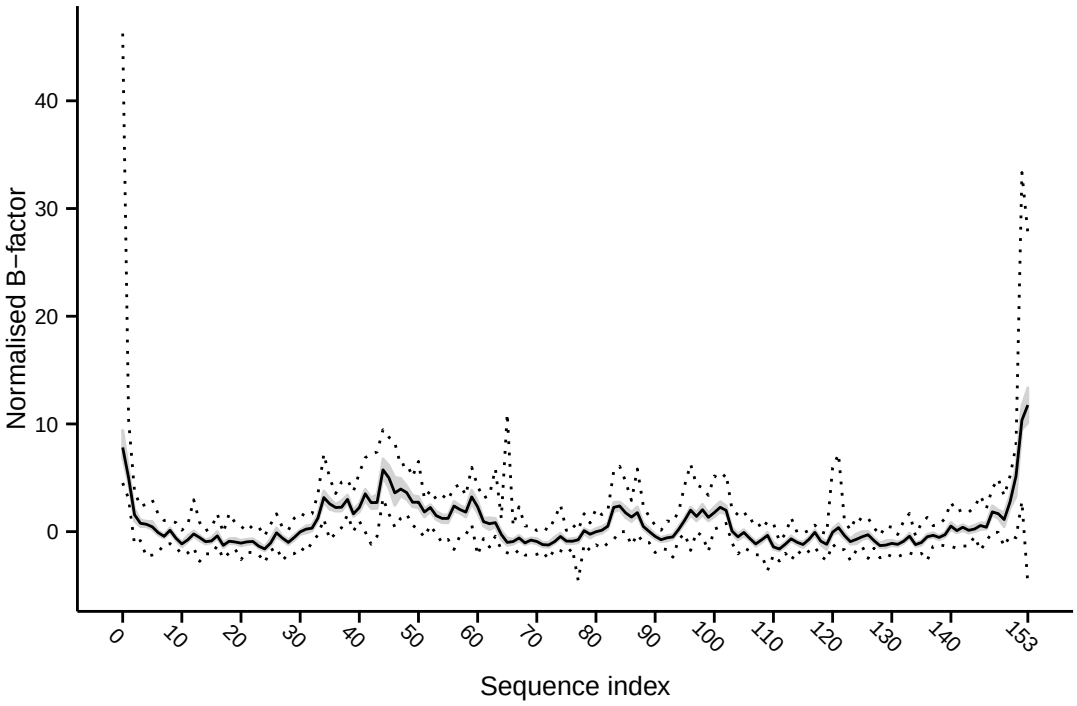


(b) Structures in space group P 3_2 2 1

Figure B.5: Consensus B-factor profiles for sperm whale myoglobin



(a) Structures in space group P 1 2_1 1



(b) Structures in space group P 6

Figure B.6: Consensus B-factor profile for yeast cytochrome c peroxidase

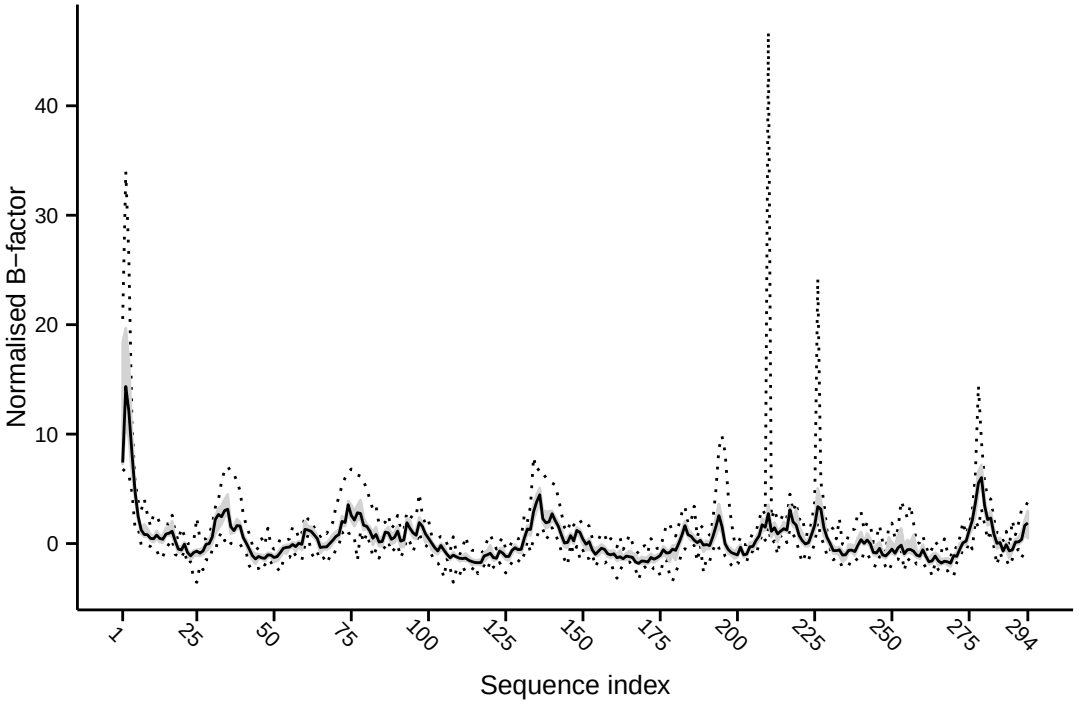


Figure B.7: Consensus B-factor profile for Pseudomonas cytochrome P450 with camphor

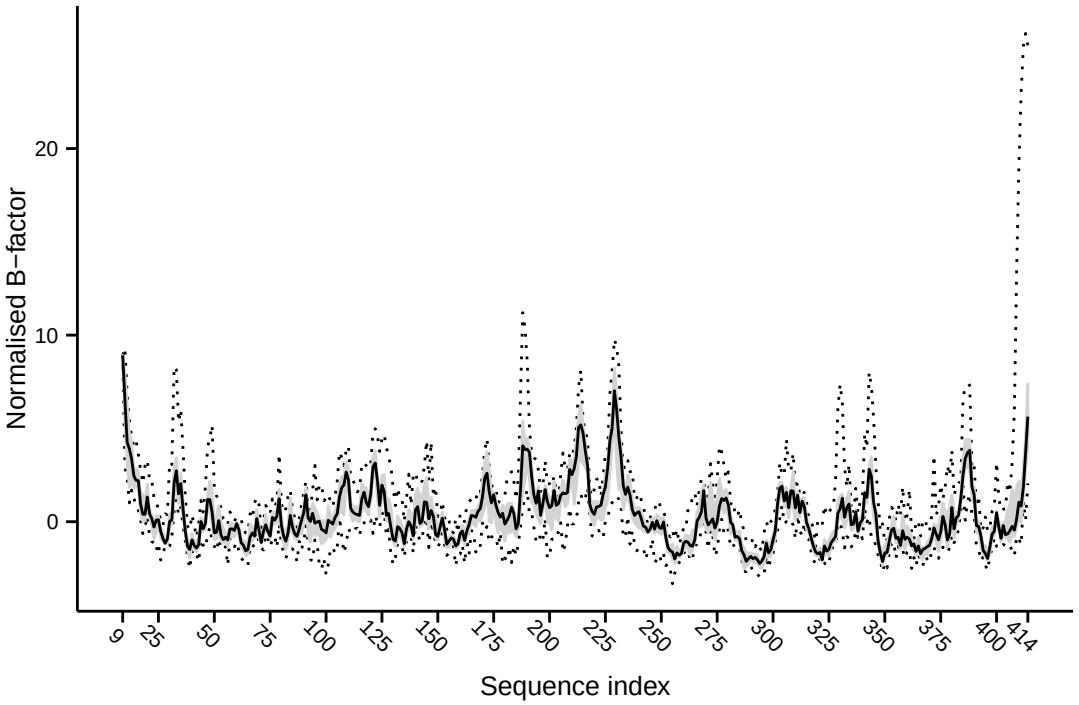


Figure B.8: Consensus B-factor profile for human heat shock protein 90 bound to various ligands

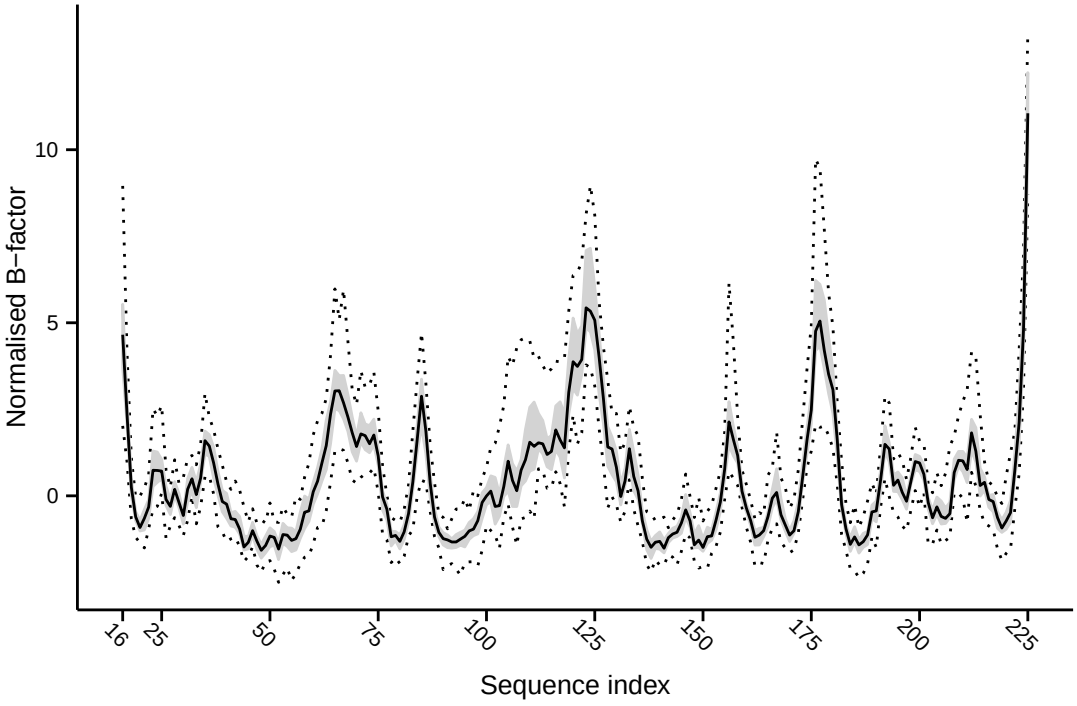


Figure B.9: Consensus B-factor profile for thermolysin bound to various ligands

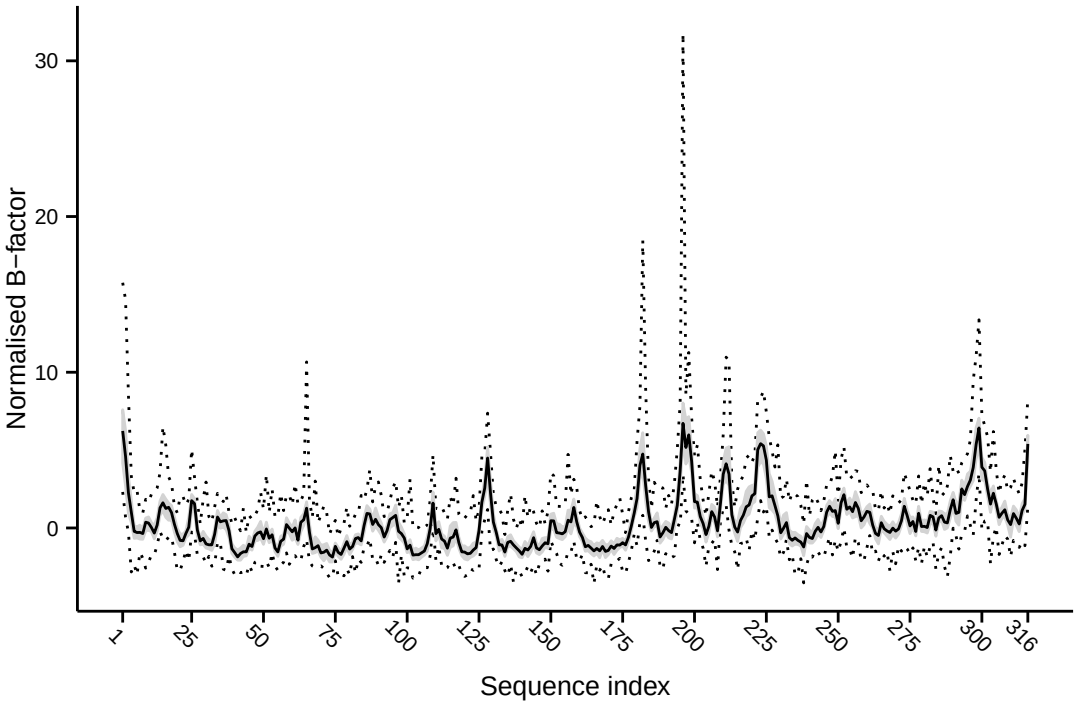


Figure B.10: Consensus B-factor profile for human HRas GTPase bound to various ligands

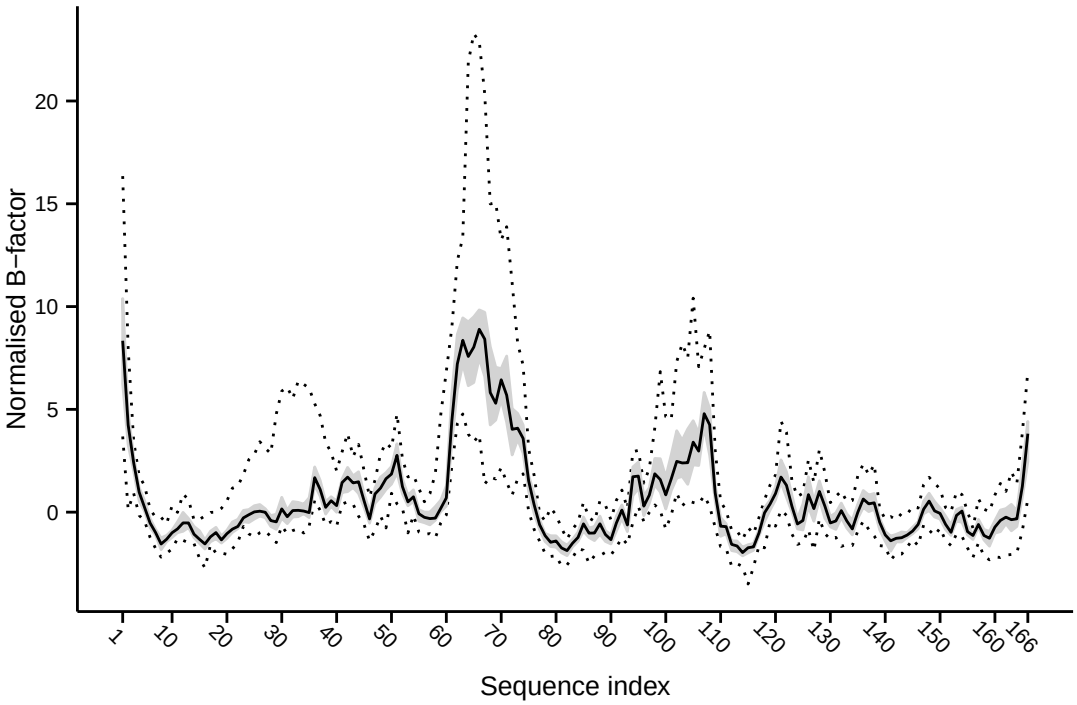
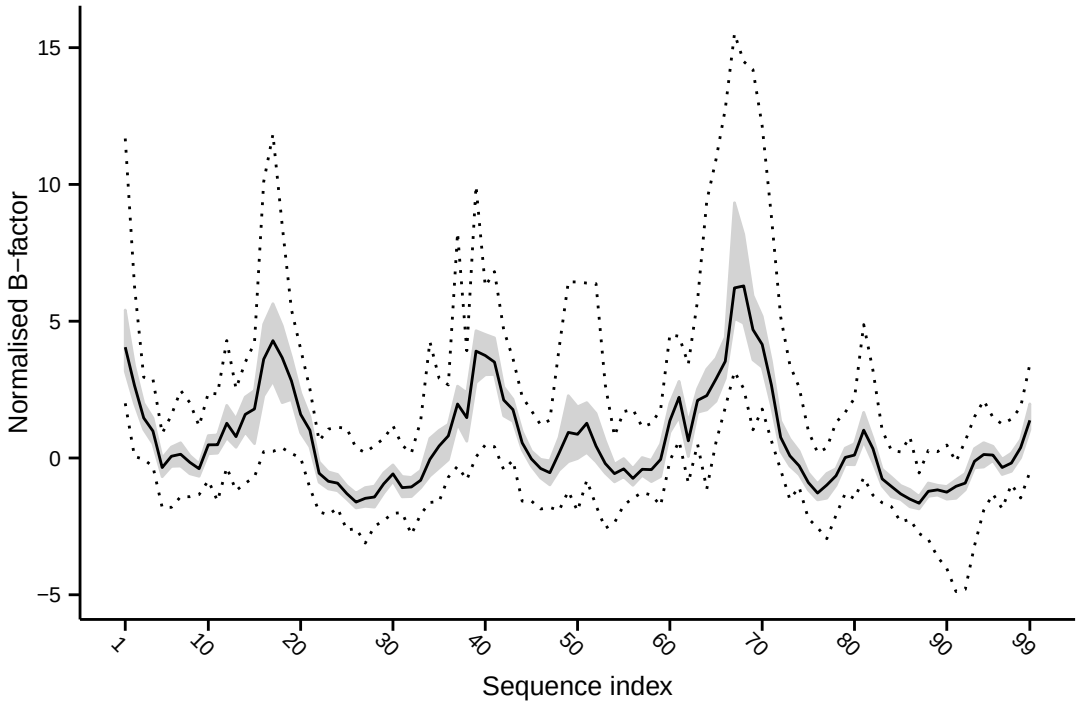
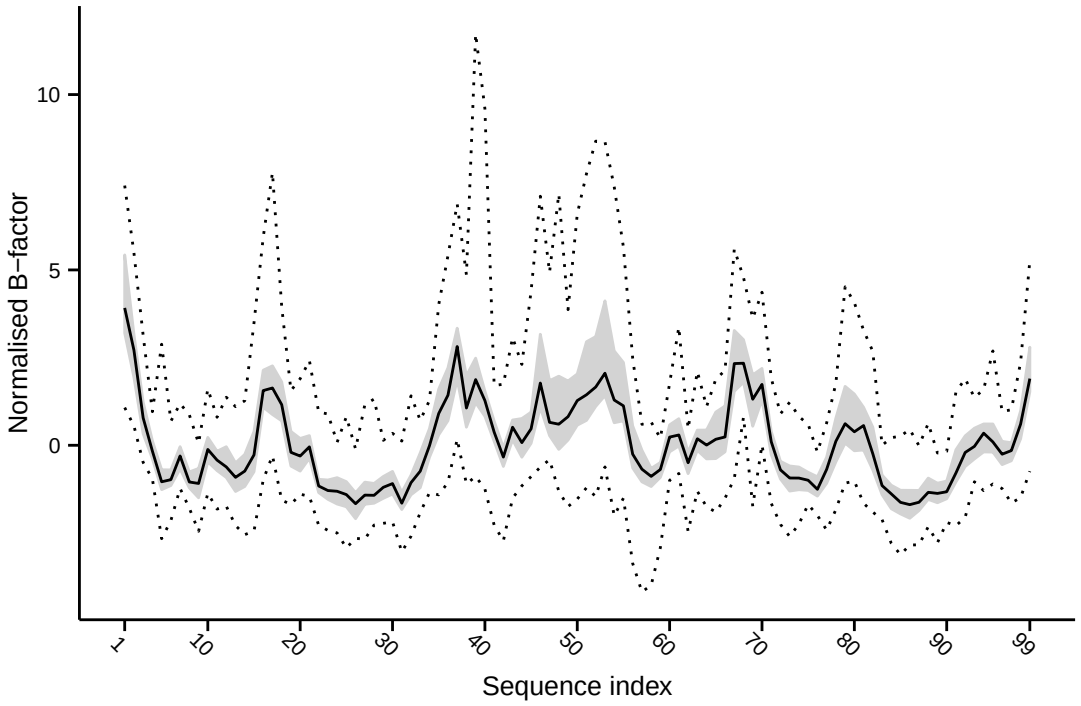


Figure B.11: Consensus B-factor profiles for HIV-1 protease homodimers bound to various ligands

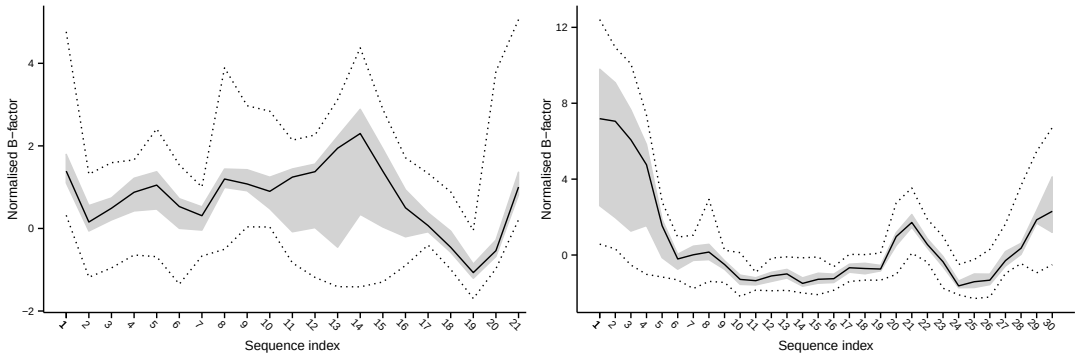


(a) "A" chain of the dimer



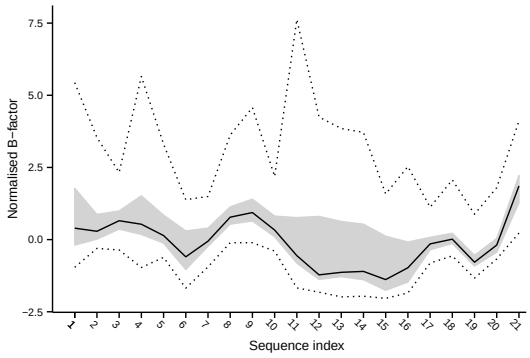
(b) "B" chain of the dimer

Figure B.12: Consensus B-factor profiles for human insulin where the unit cell is a homodimer

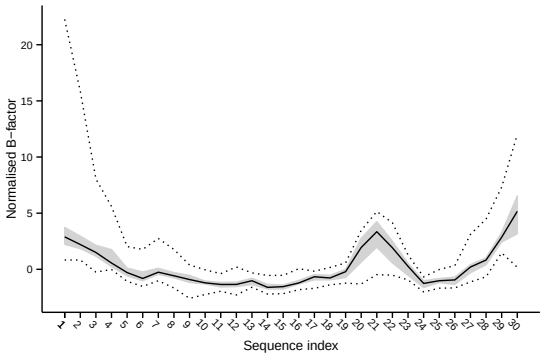


(a) "A" chain

(b) "B" chain

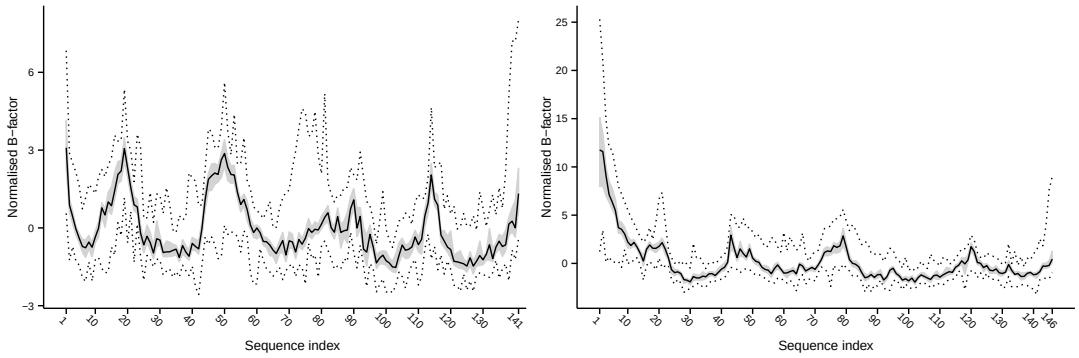


(c) "A" chain (second copy)



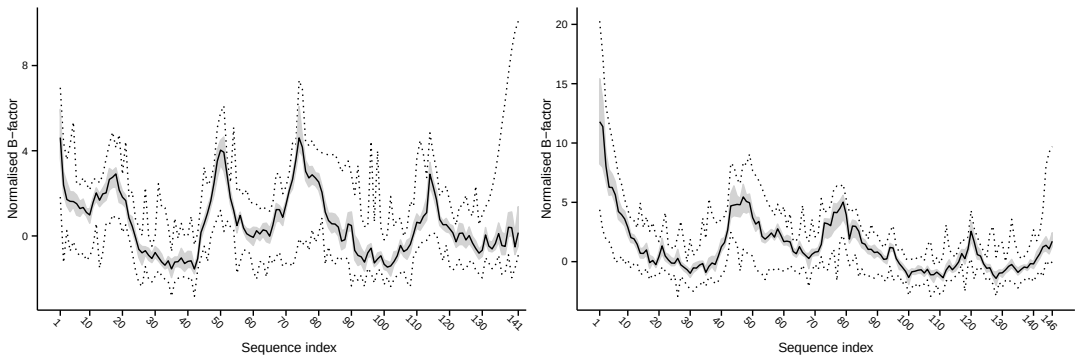
(d) "B" chain (second copy)

Figure B.13: Consensus B-factor profiles for human haemoglobin where the unit cell is a tetramer



(a) Alpha chain

(b) Beta chain



(c) Alpha chain (second copy)

(d) Beta chain (second copy)

Appendix C

MD simulations

Figure C.1: Profile of the square deviations in the positions of the alpha-carbon for hen egg white lysozyme with respect to the centroid of the cluster (structure `3A92`). The solid line plots the median deviation and the grey ribbon plots the interquartile range. The dotted line plots the mean deviations which are equivalent to the MSF measurements of a MD simulation.

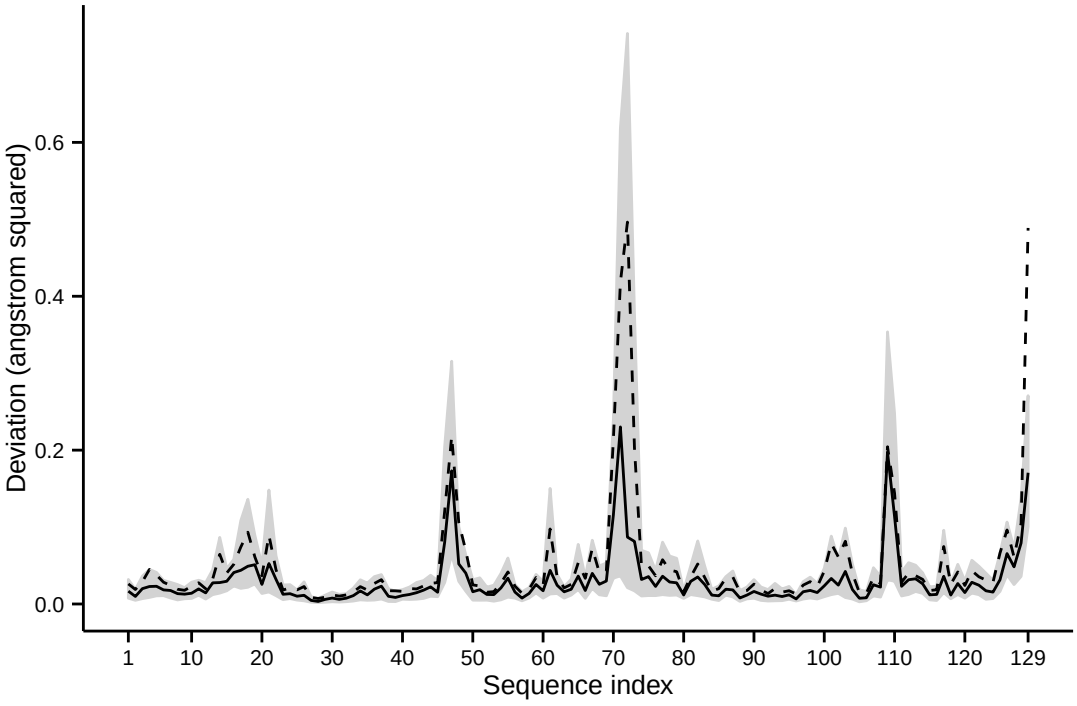
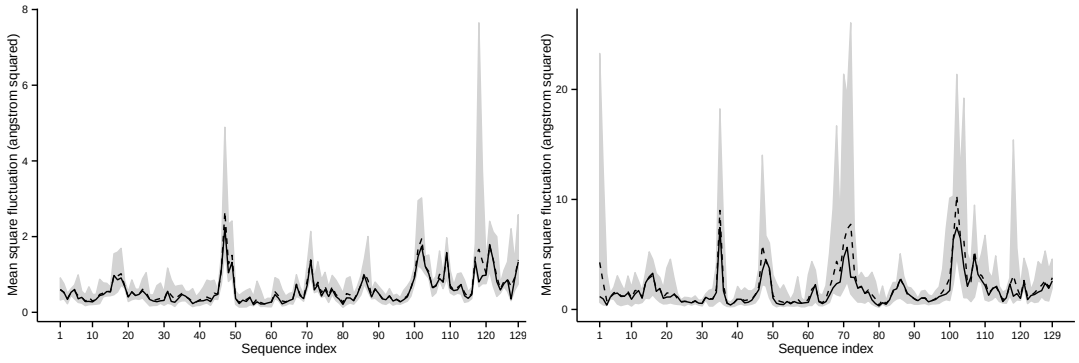
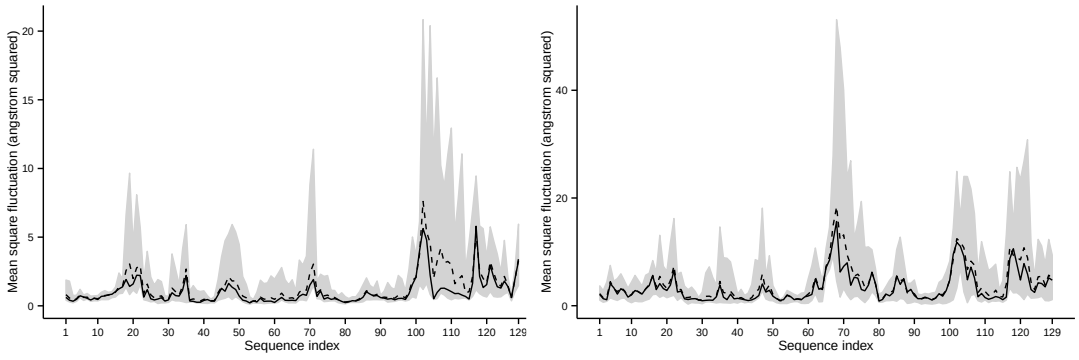


Figure C.2: Alpha-carbon MD MSF profiles for hen egg white lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.



(a) Amber99SB-ILDN

(b) OPLS-AA



(c) CHARMM27

(d) GROMOS54a7

Figure C.3: Profile of the backbone phi and psi torsion angle dispersions for hen egg white lysozyme structures.

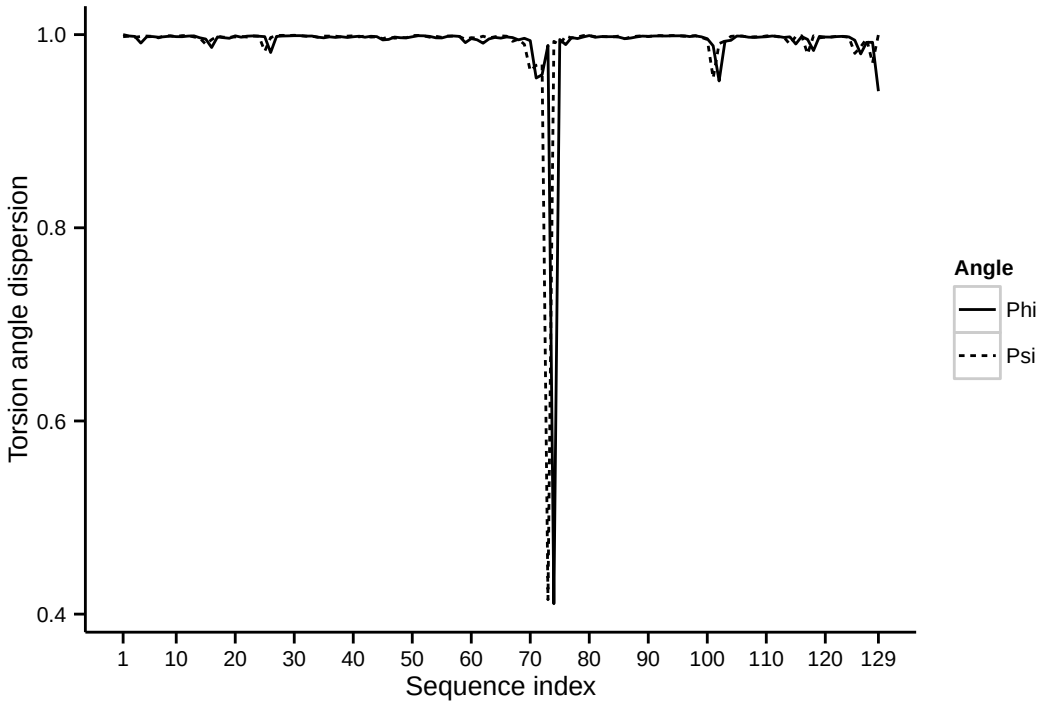
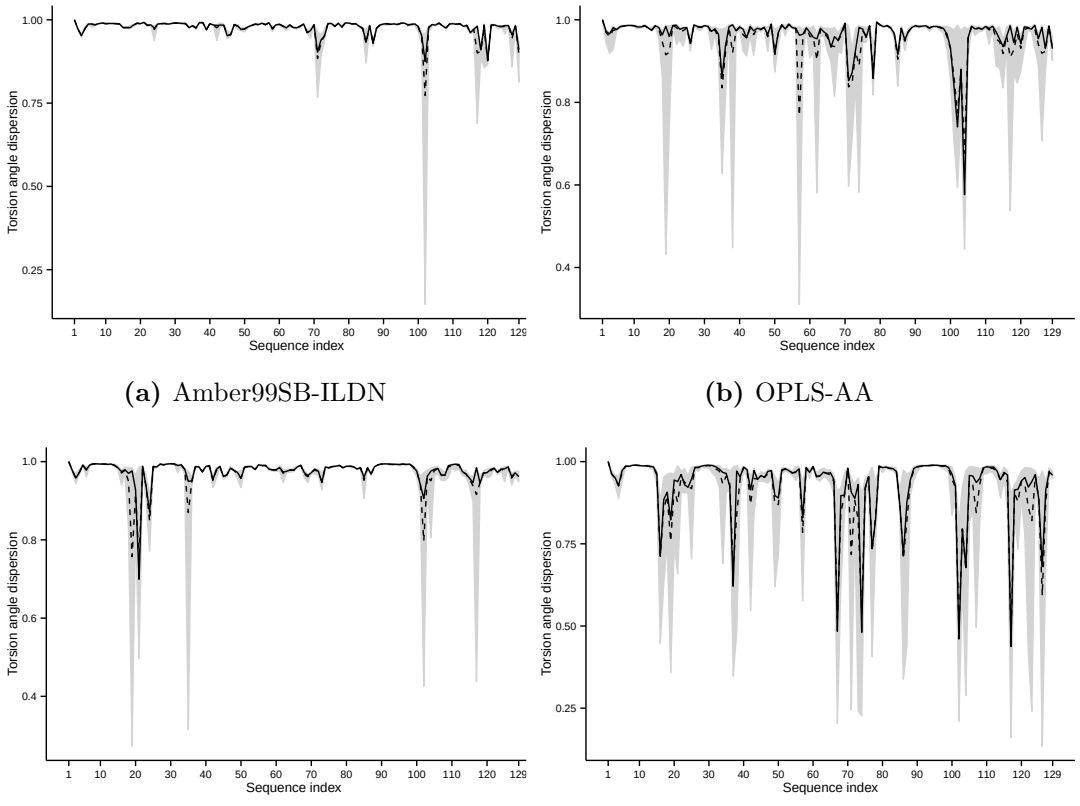


Figure C.4: MD phi torsion angle profiles for hen egg white lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.



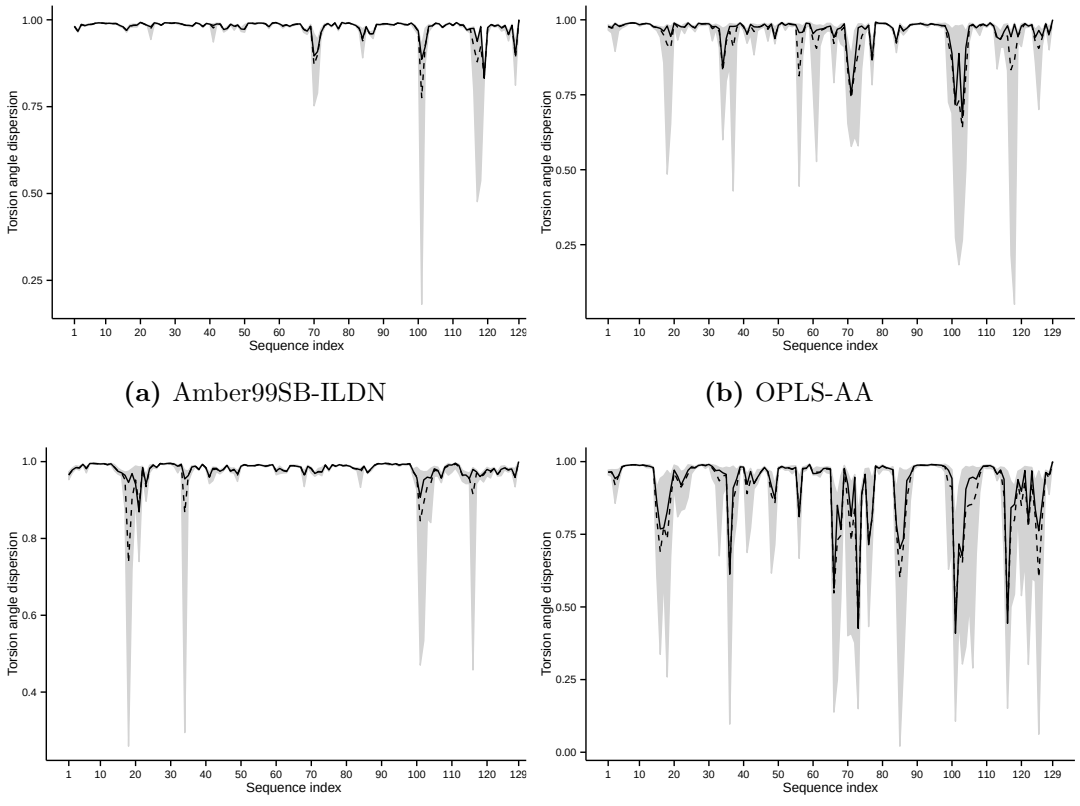
(a) Amber99SB-ILDN

(b) OPLS-AA

(c) CHARMM27

(d) GROMOS54a7

Figure C.5: MD psi torsion angle profiles for hen egg white lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.



(a) Amber99SB-ILDN

(b) OPLS-AA

(c) CHARMM27

(d) GROMOS54a7

Figure C.6: Profile of the square deviations in the positions of the alpha-carbon for human lysozyme with respect to the centroid of the cluster (structure `1C43`). The solid line plots the median deviation and the grey ribbon plots the interquartile range. The dotted line plots the mean deviations which are equivalent to the MSF measurements of a MD simulation.

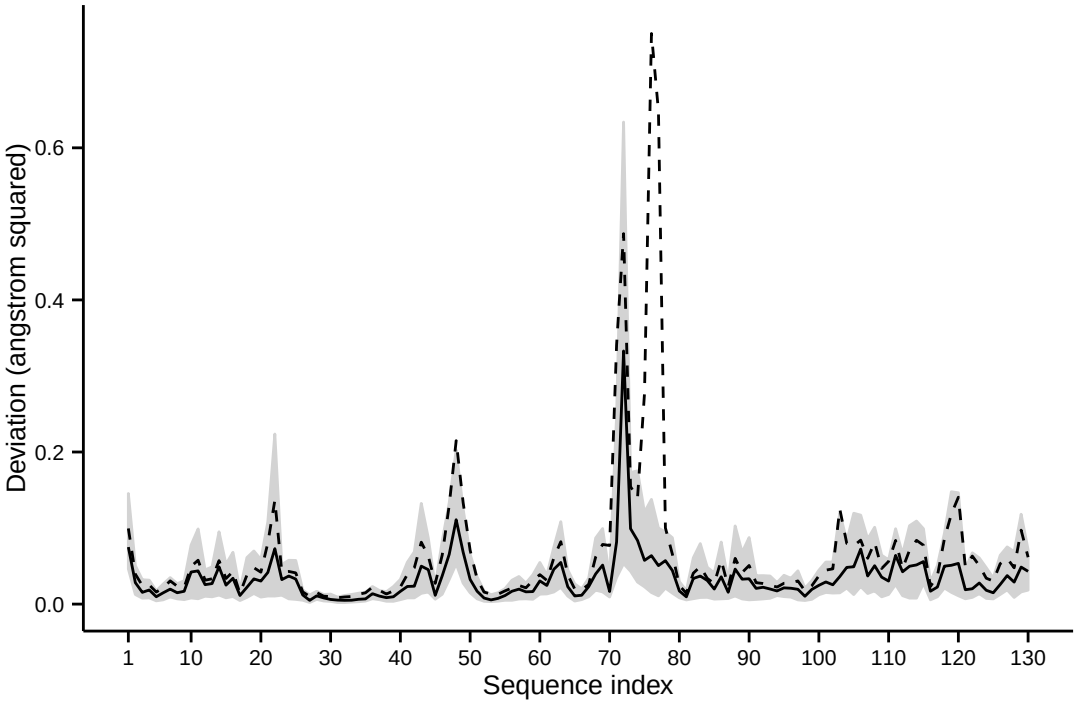
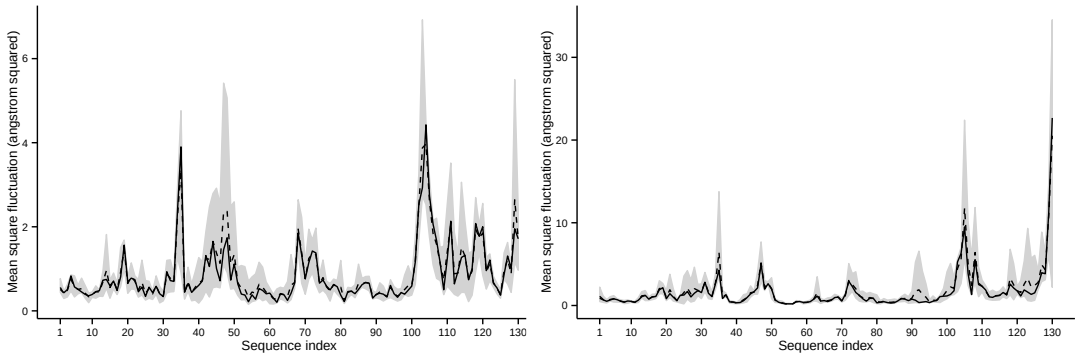
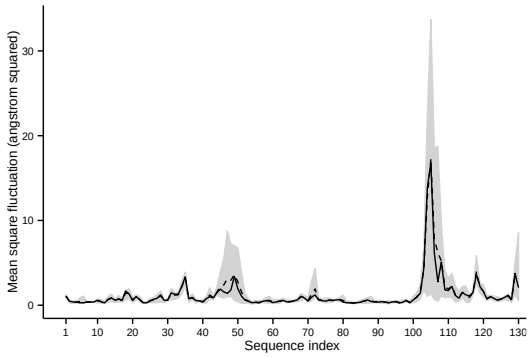


Figure C.7: Alpha-carbon MD MSF profiles for human lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.

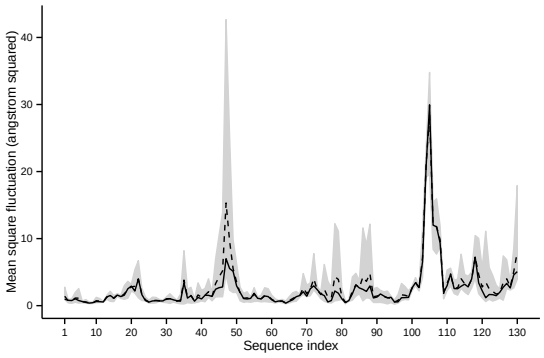


(a) Amber99SB-ILDN

(b) OPLS-AA



(c) CHARMM27



(d) GROMOS54a7

Figure C.8: Profile of the backbone psi and phi torsion angle dispersions for human lysozyme structures.

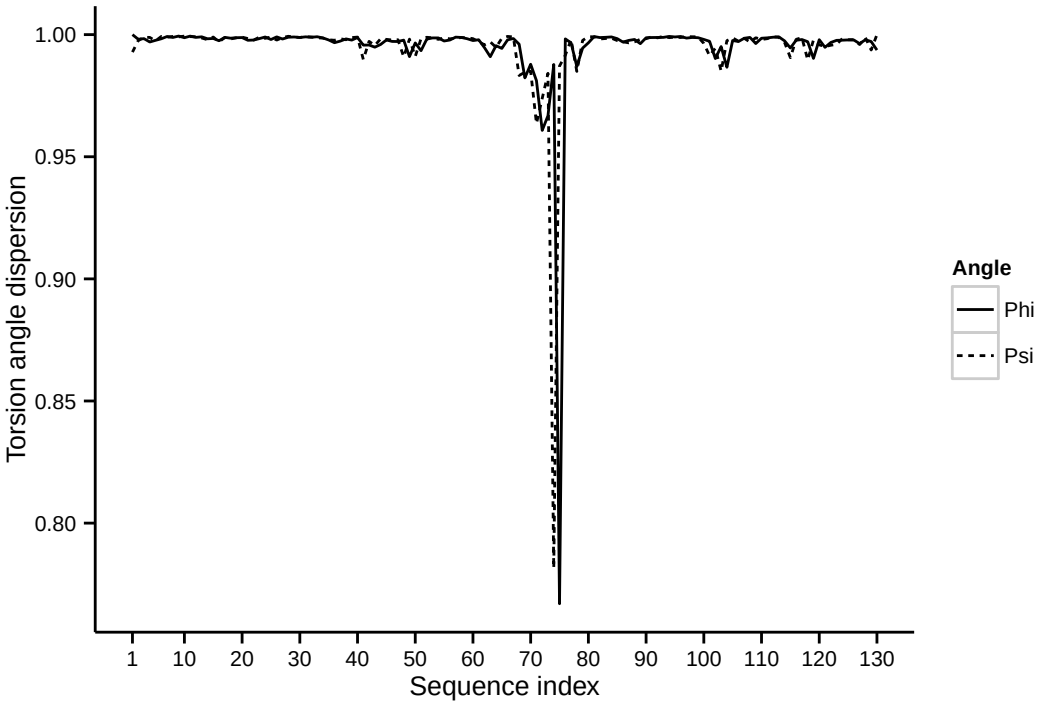
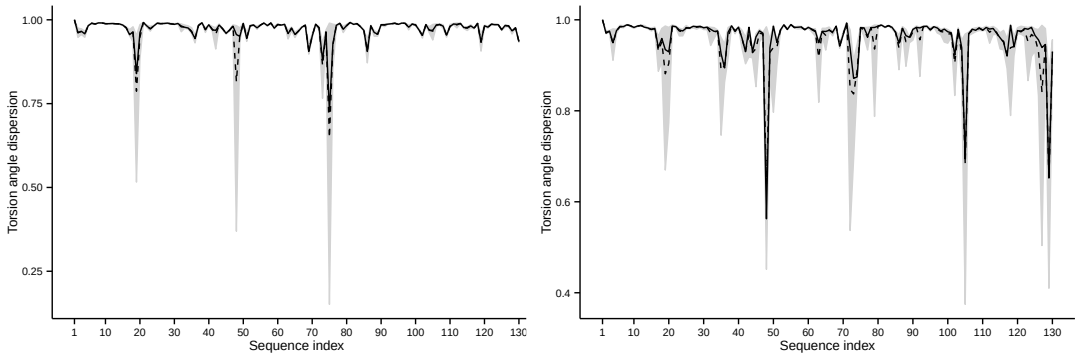
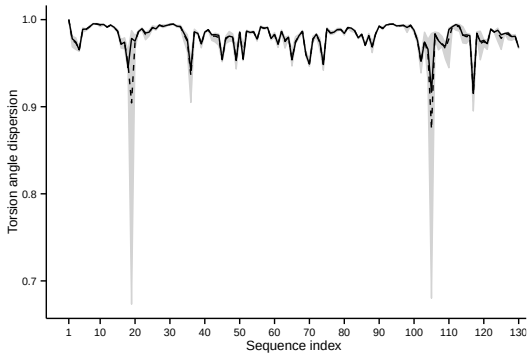


Figure C.9: MD phi torsion angle profiles for human lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.

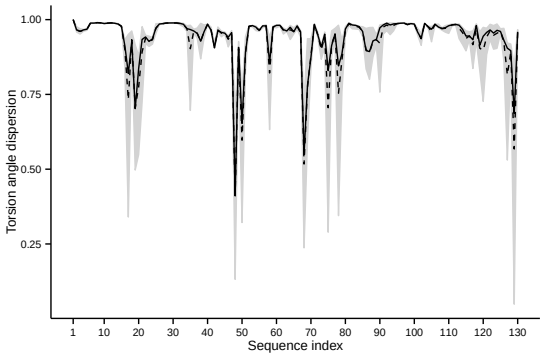


(a) Amber99SB-ILDN

(b) OPLS-AA

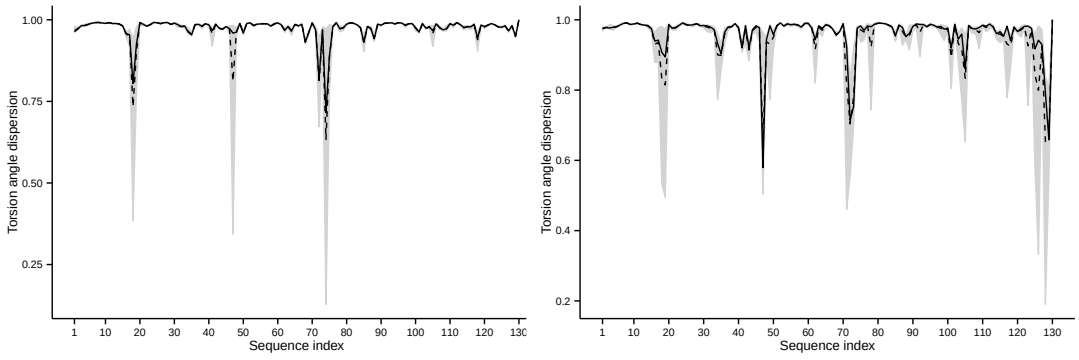


(c) CHARMM27



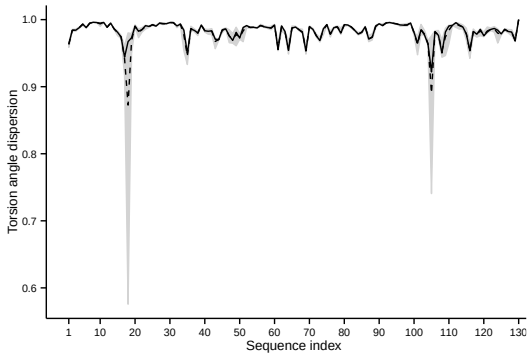
(d) GROMOS54a7

Figure C.10: MD psi torsion angle profiles for human lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.

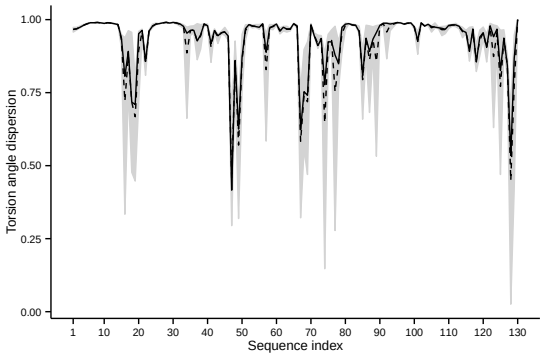


(a) Amber99SB-ILDN

(b) OPLS-AA



(c) CHARMM27



(d) GROMOS54a7

Figure C.11: Profile of the square deviations in the positions of the alpha-carbon for T4 lysozyme with respect to the centroid of the cluster (structure `1L19`). The solid line plots the median deviation and the grey ribbon plots the interquartile range. The dotted line plots the mean deviations which are equivalent to the MSF measurements of a MD simulation.

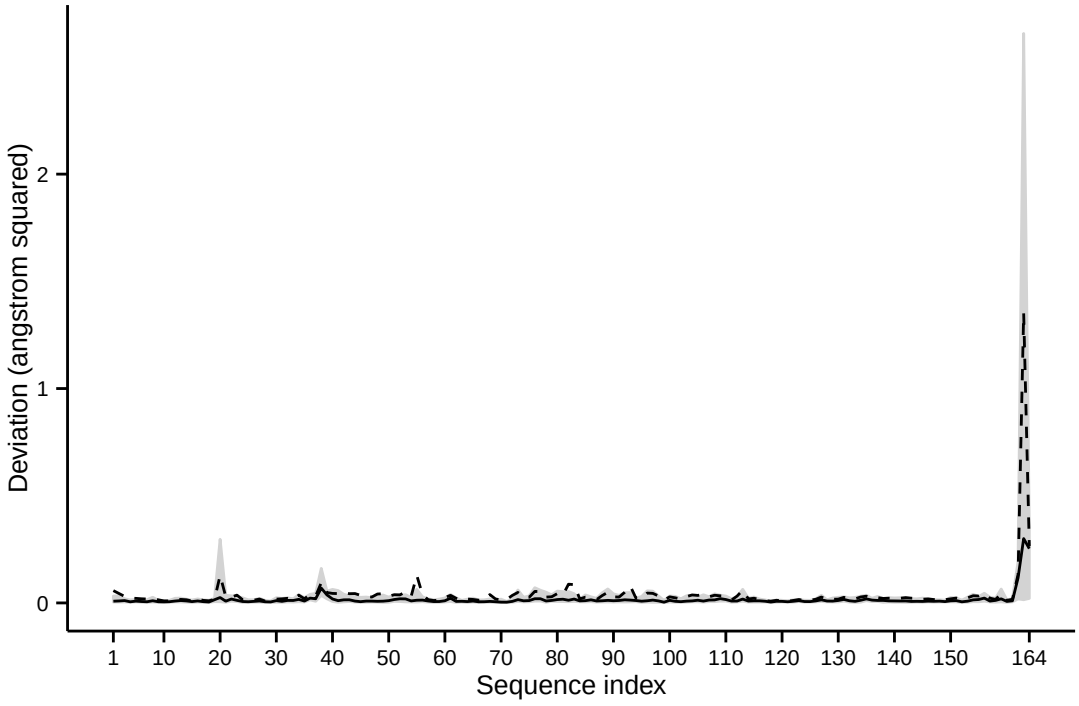
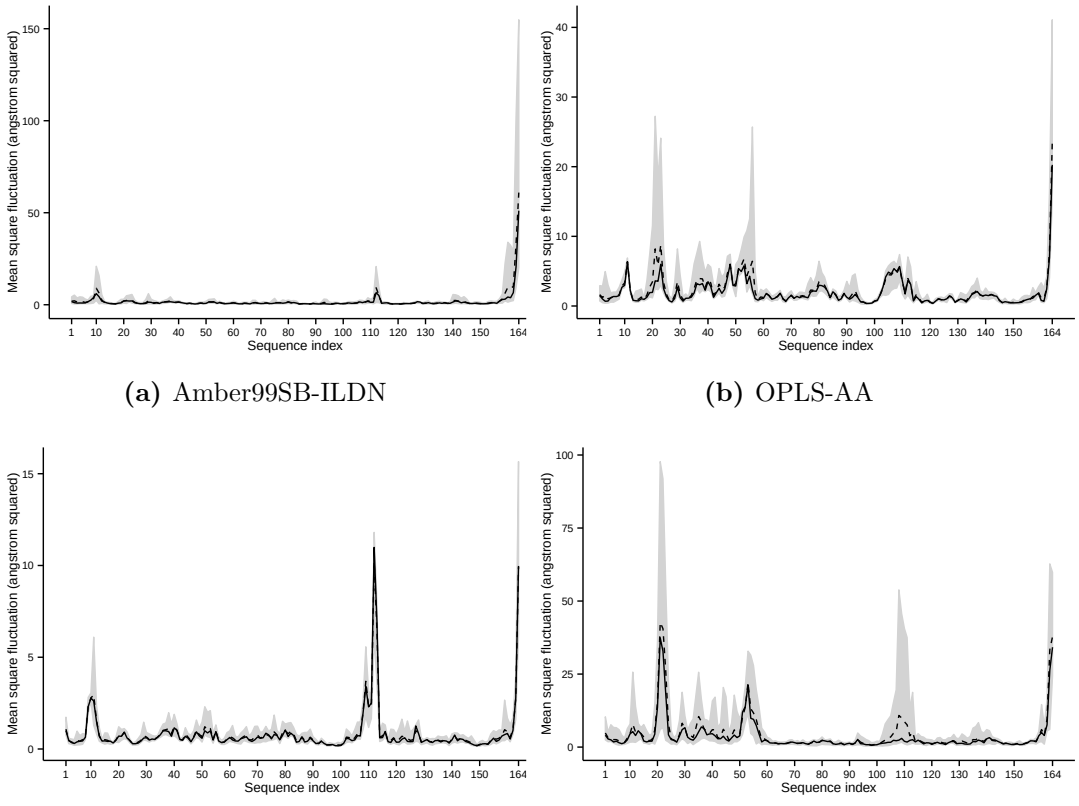


Figure C.12: Alpha-carbon MD MSF profiles for T4 lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.



(a) Amber99SB-ILDN

(b) OPLS-AA

(c) CHARMM27

(d) GROMOS54a7

Figure C.13: Profile of the backbone phi and psi torsion angle dispersions for T4 lysozyme structures.

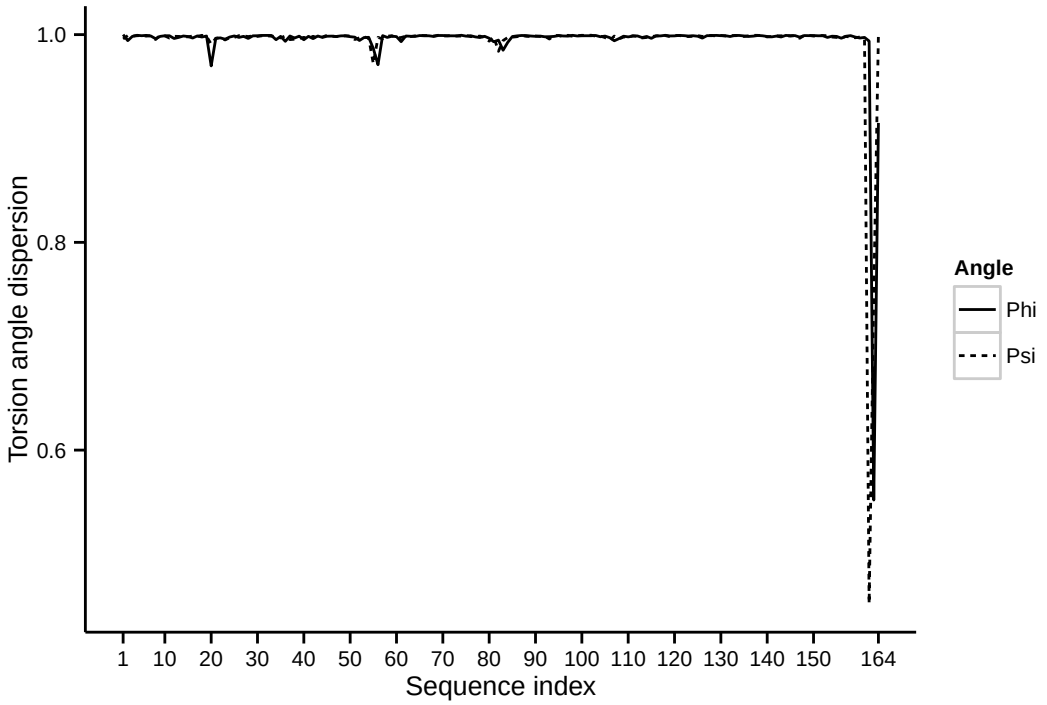
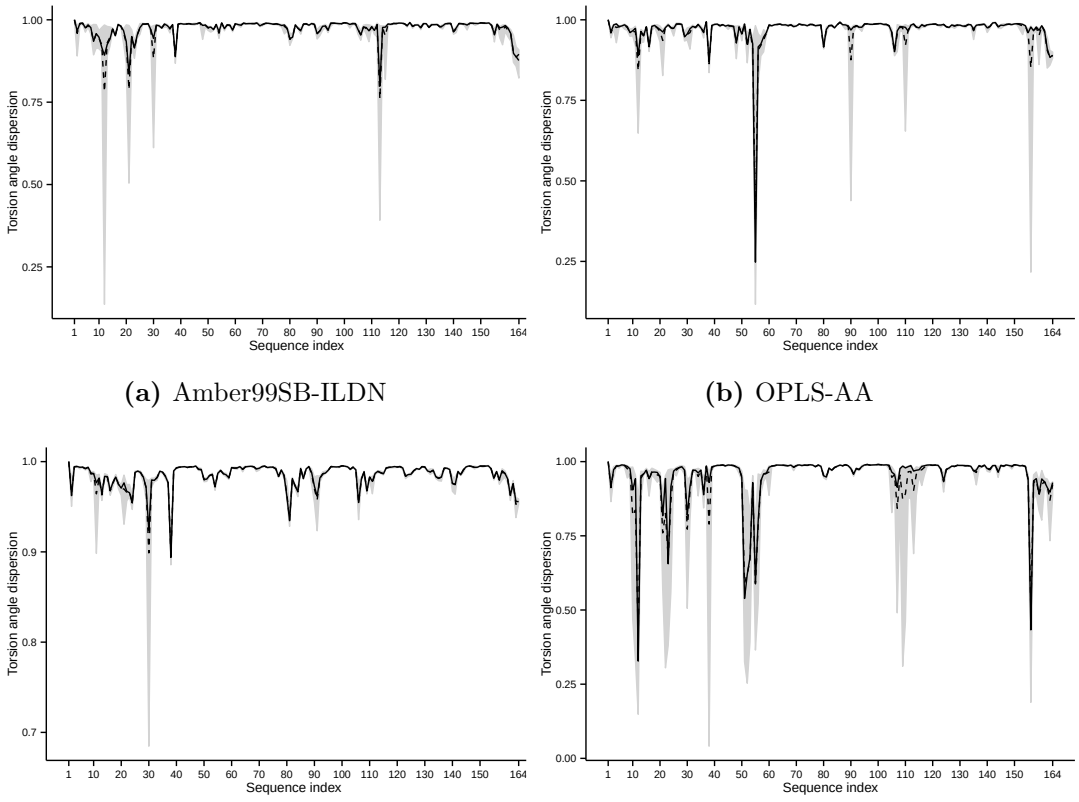


Figure C.14: MD phi torsion angle profiles for T4 lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.



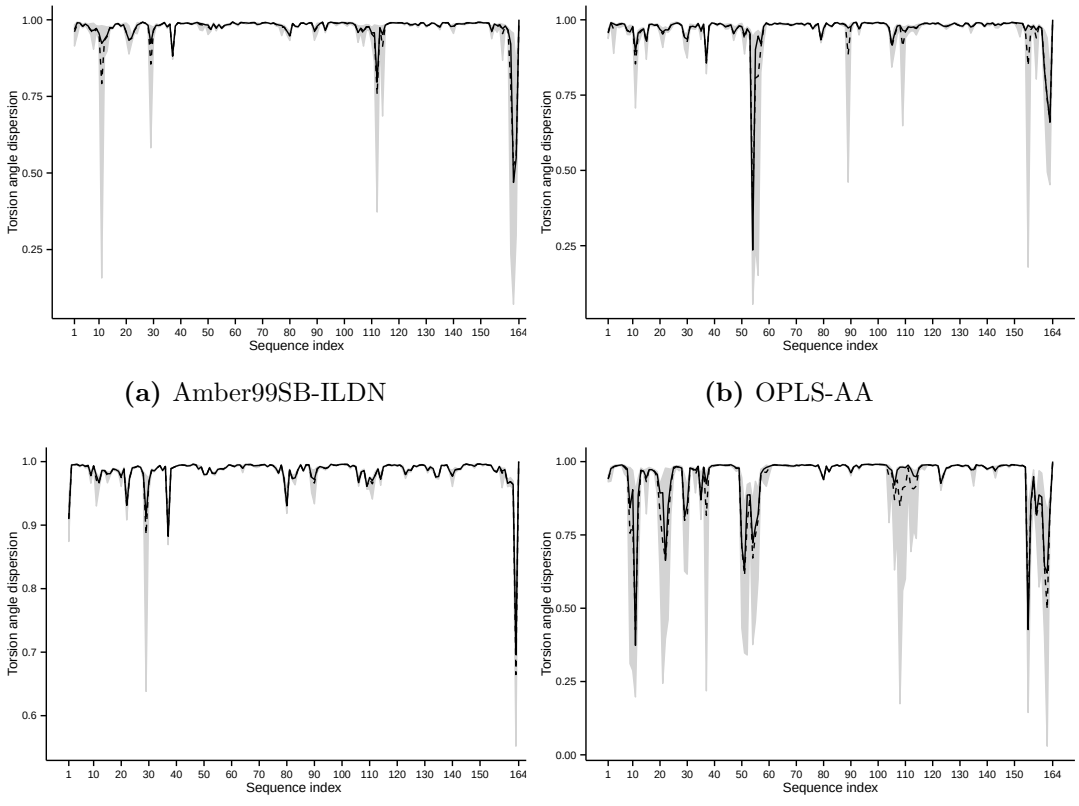
(a) Amber99SB-ILDN

(b) OPLS-AA

(c) CHARMM27

(d) GROMOS54a7

Figure C.15: MD psi torsion angle profiles for T4 lysozyme. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.



(a) Amber99SB-ILDN

(b) OPLS-AA

(c) CHARMM27

(d) GROMOS54a7

Figure C.16: Profile of the square deviations in the positions of the alpha-carbon for pancreatic ribonuclease with respect to the centroid of the cluster (structure `1KF4`). The solid line plots the median deviation and the grey ribbon plots the interquartile range. The dotted line plots the mean deviations which are equivalent to the MSF measurements of a MD simulation.

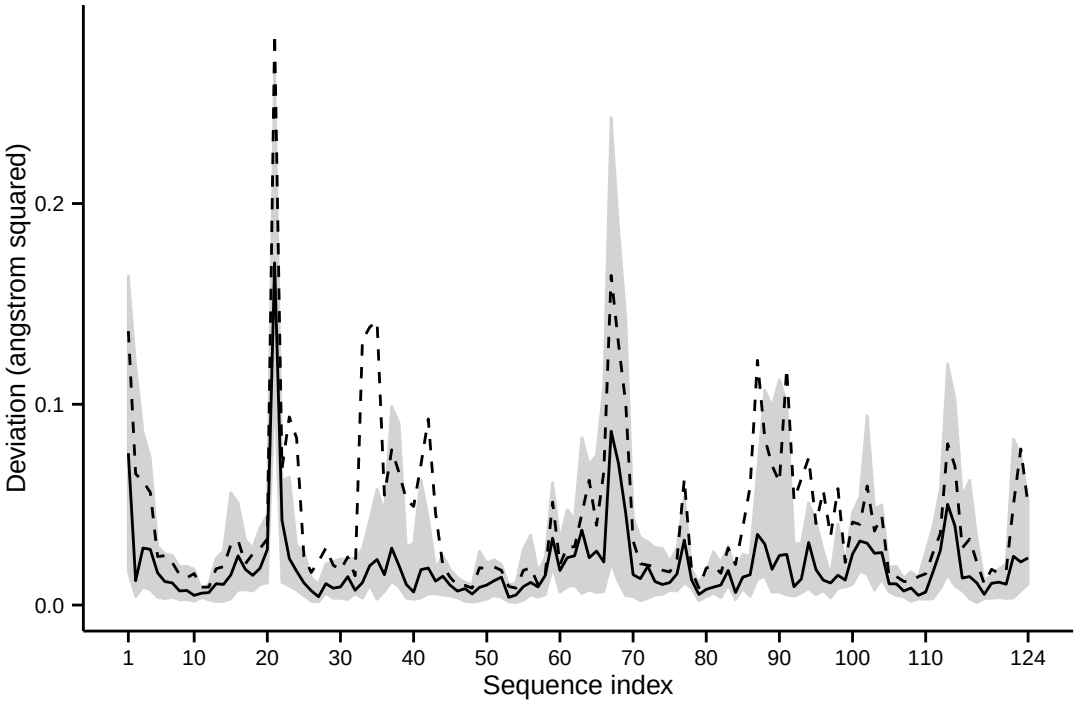
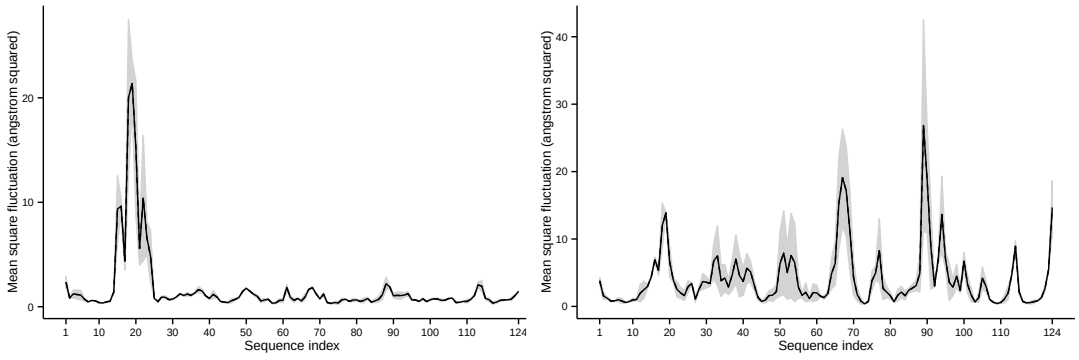
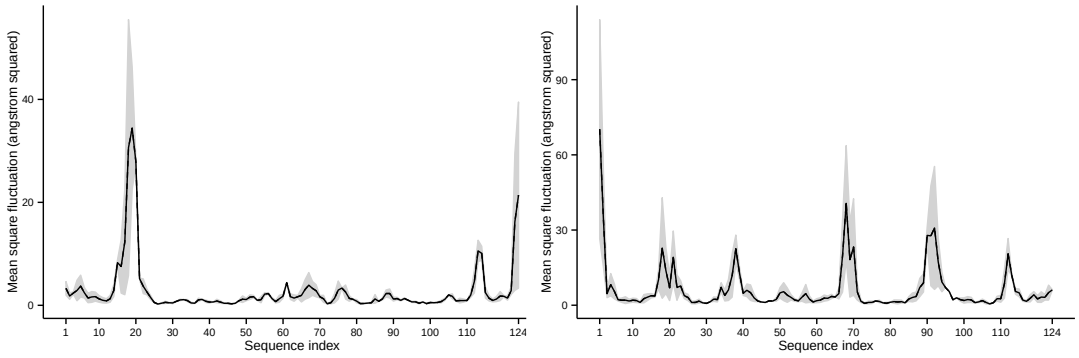


Figure C.17: Alpha-carbon MD MSF profiles for pancreatic ribonuclease.The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.



(a) Amber99SB-ILDN

(b) OPLS-AA



(c) CHARMM27

(d) GROMOS54a7

Figure C.18: Profile of the backbone torsion angle dispersions for pancreatic ribonuclease structures.

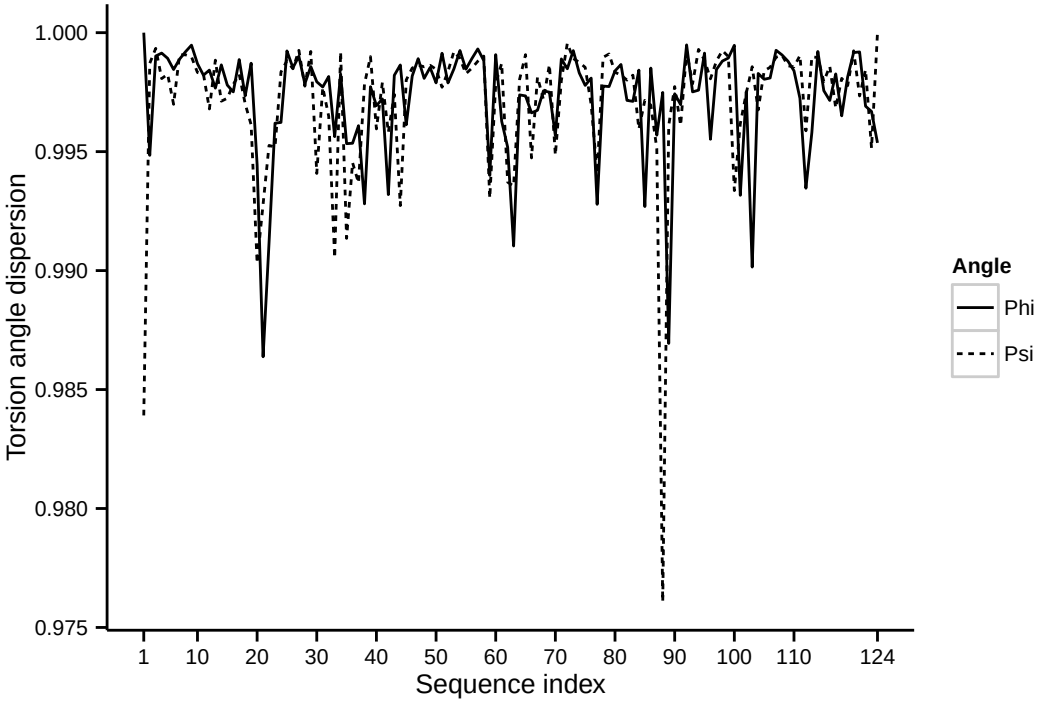
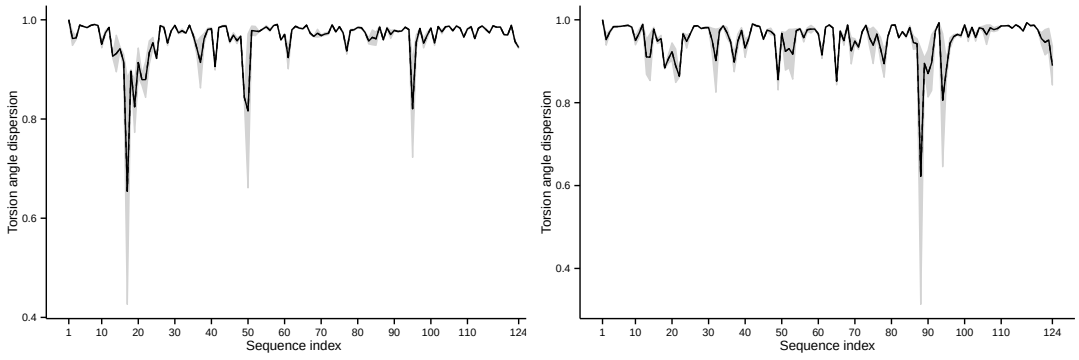
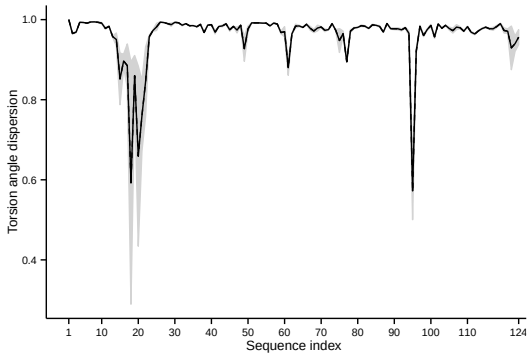


Figure C.19: MD phi torsion angle profiles for pancreatic ribonuclease. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.

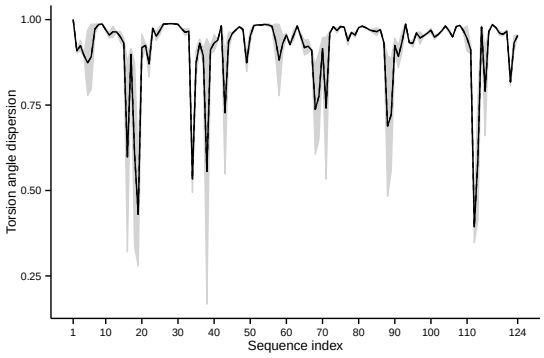


(a) Amber99SB-ILDN

(b) OPLS-AA

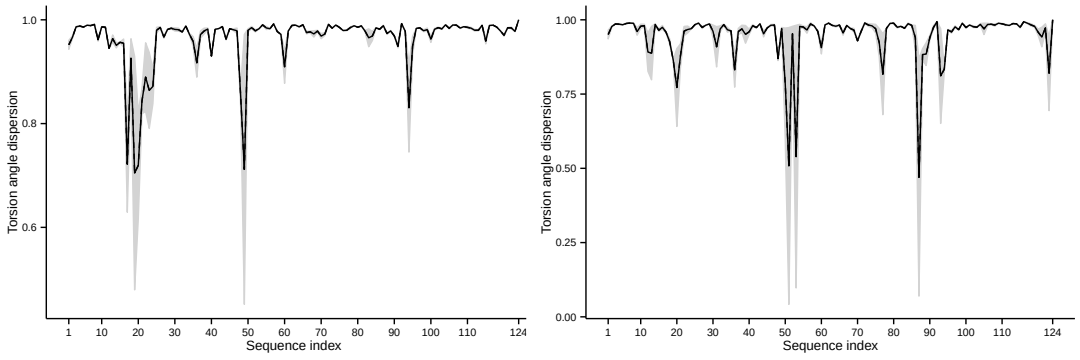


(c) CHARMM27



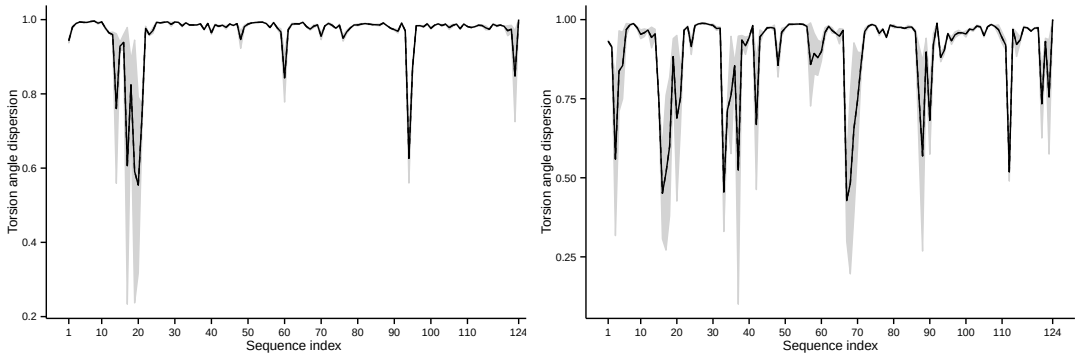
(d) GROMOS54a7

Figure C.20: MD psi torsion angle profiles for pancreatic ribonuclease. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.



(a) Amber99SB-ILDN

(b) OPLS-AA



(c) CHARMM27

(d) GROMOS54a7

Figure C.21: Profile of the square deviations in the positions of the alpha-carbon for staphylococcal nuclease with respect to the centroid of the cluster (structure `2F0I`). The solid line plots the median deviation and the grey ribbon plots the interquartile range. The dotted line plots the mean deviations which are equivalent to the MSF measurements of a MD simulation.

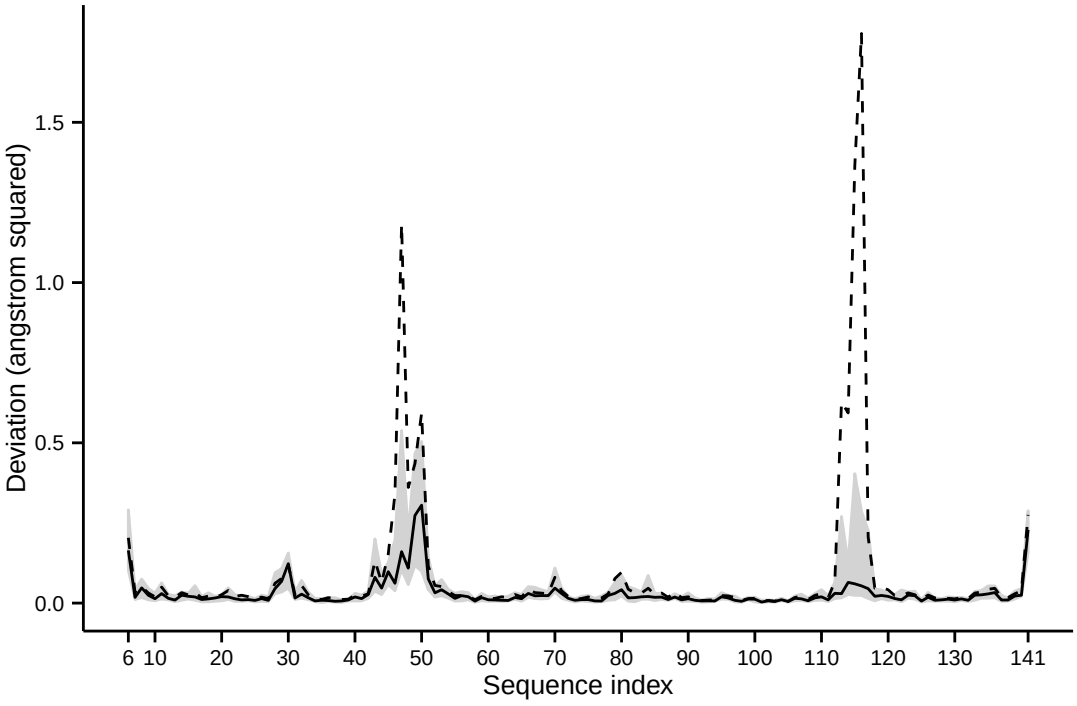
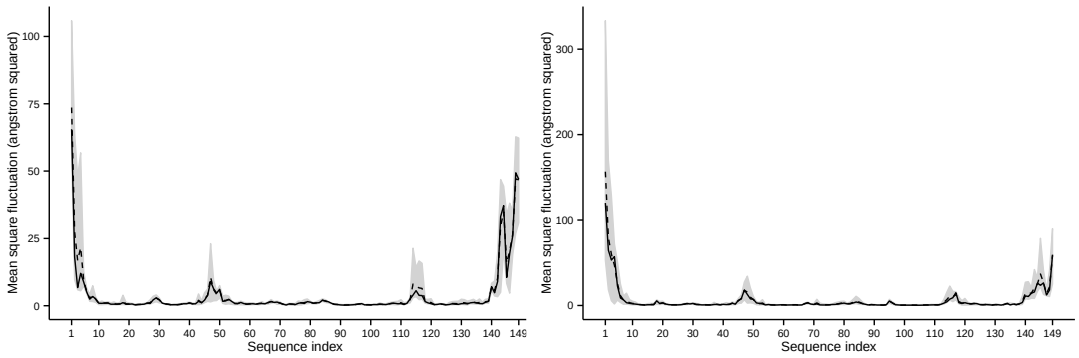
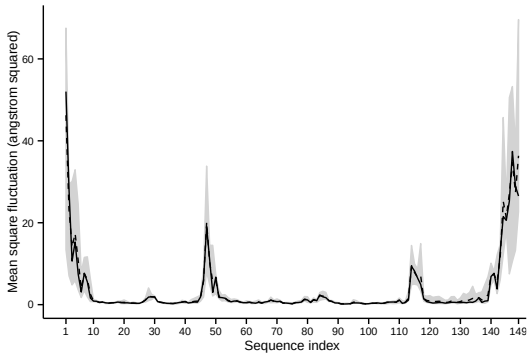


Figure C.22: Alpha-carbon MD MSF profiles for staphylococcal nuclease. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.

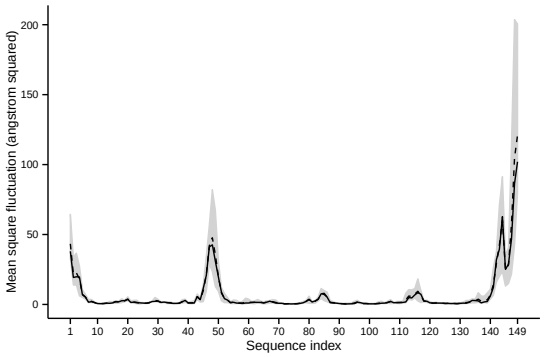


(a) Amber99SB-ILDN

(b) OPLS-AA



(c) CHARMM27



(d) GROMOS54a7

Figure C.23: Profile of the backbone phi and psi torsion angle dispersions for staphylococcal nuclease structures.

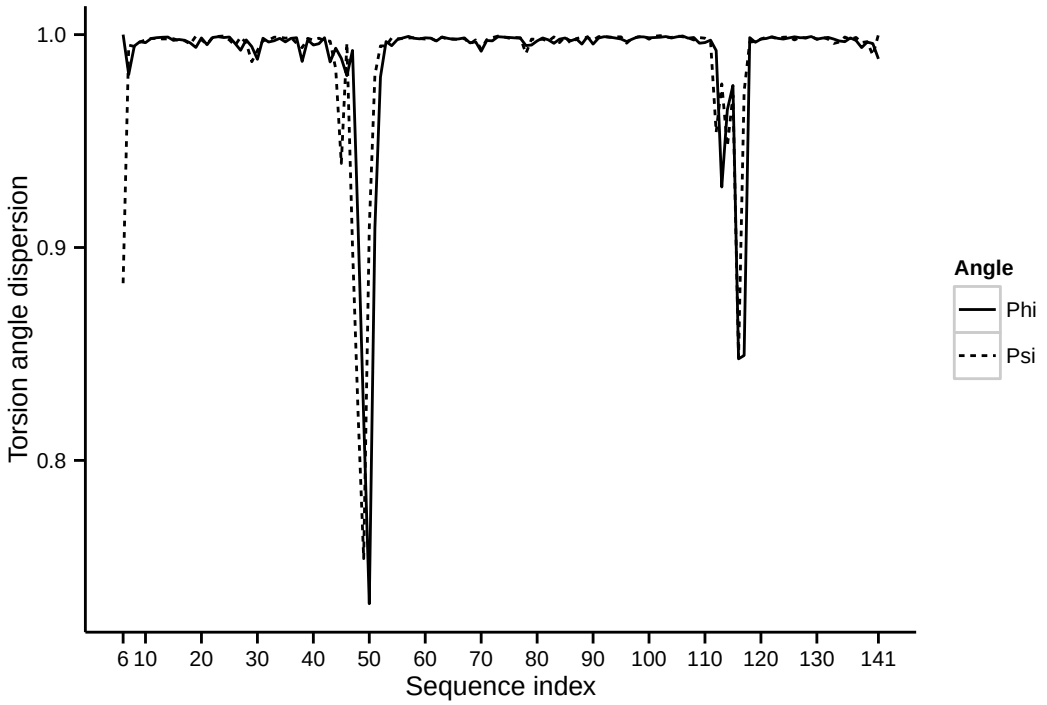
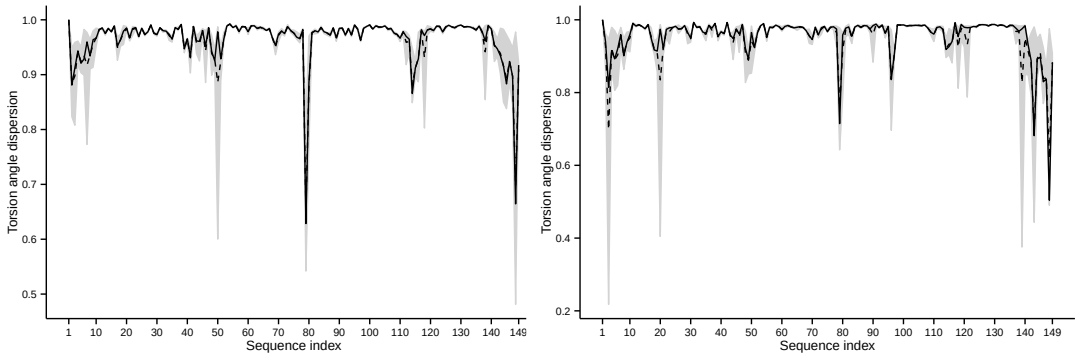
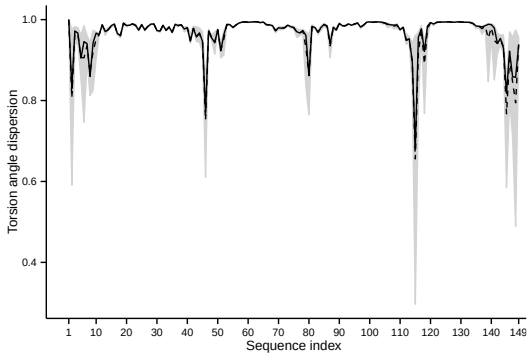


Figure C.24: MD phi torsion angle profiles for staphylococcal nuclease. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.

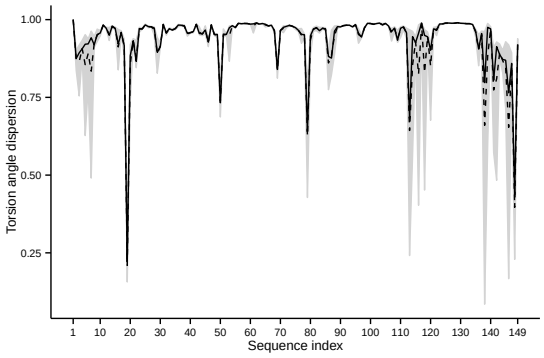


(a) Amber99SB-ILDN

(b) OPLS-AA

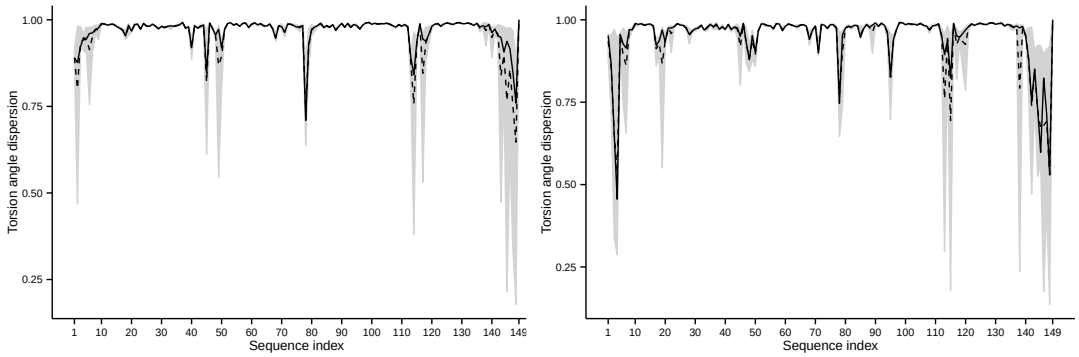


(c) CHARMM27



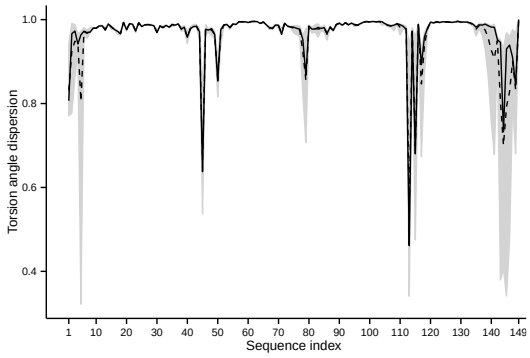
(d) GROMOS54a7

Figure C.25: MD psi torsion angle profiles for staphylococcal nuclease. The solid and dotted lines plots the median and mean dispersions respectively. The grey ribbon plots the interquartile range.

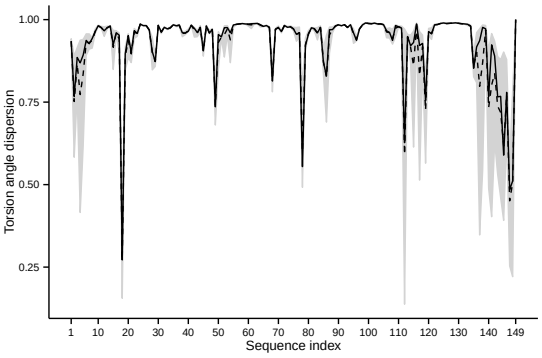


(a) Amber99SB-ILDN

(b) OPLS-AA



(c) CHARMM27



(d) GROMOS54a7

Appendix D

Qualitative analysis

Key to table headings:

Seq. index Amino acid sequence index.

PDB Conformation variability measures derived from PDB files.

B Consensus normalised B-factor profile.

SqD Square deviations derived from aligned PDB files.

r Torsion angle dispersion derived from PDB files.

MD MSF Mean square fluctuations derived from MD simulations.

MD torsion Torsion angle dispersion derived from MD simulations.

A MD simulation using the Amber99SB-ILDN force field.

O MD simulation using the OPLS-AA force field.

C MD simulation using the CHARMM27 force field.

G MD simulation using the GROMOS54a7 force field.

The five most dynamic residues as measured by a metric derived from PDB data or MD simulation are marked in the table. Each column in the table must, therefore, contain exactly five entries. The key to the table entries is:

$*$ A dynamic residue as measured using data derived from PDB files.

$+$ A dynamic residues as measured by MD simulation in agreement with the PDB data.

$-$ A residue classified as one of the most dynamic by MD simulation but not by the PDB data.

Table D.1: Locating the most flexible regions of hen egg white lysozyme

	Fluctuations						Torsion angles									
							Psi					Phi				
Seq.	PDB		MD MSF				PDB	MD torsion				PDB	MD torsion			
index	B	SqD	A	O	C	G	r	A	O	C	G	r	A	O	C	G
18										−						
20										−					−	
21										−					−	
23										−					−	
24															−	
34									−							
35				−												
36											−					
37																−
47	∗	∗	+													
66											−					
67						−										−
68						−										
70		∗					∗	+	+							
71	∗	∗		+			∗	+	+			∗	+	+		
72	∗						∗					∗				
73							∗				+					
74												∗				+
78														−		
101			−	−	−		∗	+	+	+	+			−		
102			−	−	−	−						∗	+	+	+	+
103				−	−	−			−							
104														−		
109		∗	+													
116											−					
117					−											−
118						−							−			
119								−								
120													−			
121			−													
128	∗							−								
129	∗	∗			+							∗	+			

Table D.2: Locating the most flexible regions of human lysozyme

Seq. index	Fluctuations						Torsion angles									
							Psi					Phi				
	PDB		MD MSF				PDB	MD torsion				PDB	MD torsion			
	B	SqD	A	O	C	G	r	A	O	C	G	r	A	O	C	G
17										−						
18								−							−	
19								−			−		−			−
35			−							−						
36															−	
47									−		−					
48		*												−		−
49											−					
50																−
67											−					
68							*									−
69												*	+			
71		*					*					*				
72	*	*					*	+	+			*				
73	*	*					*		+			*	+	+		
74		*					*	+						−	−	
75								−				*	+			
86													−			
102			−													
103			−		−											
104			−	−	−	−				−						
105			−	−	−	−								−	−	
106					−	−										
107						−										
108				−	−					−						
116										−						
117														−	−	
122	*															
123	*															
128									−		−					
129				−				−								−
130	*			+												

Table D.3: Locating the most flexible regions of T4 lysozyme

Seq. index	Fluctuations						Torsion angles									
							Psi					Phi				
	PDB		MD MSF				PDB	MD torsion				PDB	MD torsion			
	B	SqD	A	O	C	G	r	A	O	C	G	r	A	O	C	G
1										−						
10			−													
11				−							−					
12													−	−		−
20							*					*				
21						−							−			
22						−										
23				−												
24															−	
29										−						
30															−	
36	*															
37								−	−	−						
38		*											−	−	−	
39		*														
48				−												
51											−					−
52																−
53						−										
54								−								
55	*						*							−		−
56												*				
80										−						
81															−	
82							*									
83												*				
106															−	
109					−											
112			−		−			−								
113					−								−			
155											−					
156																−
161								−	−		−					
162	*	*	+				*	+	+		+					
163	*	*	+	+	+	+	*	+	+	+		*	+	+		
164	*	*	+	+	+	+						*		+		

Table D.4: Locating the most flexible regions of pancreatic ribonuclease

	Fluctuations						Torsion angles									
							Psi					Phi				
Seq.	PDB		MD MSF				PDB	MD torsion				PDB	MD torsion			
index	B	SqD	A	O	C	G	r	A	O	C	G	r	A	O	C	G
1	*	*				+	*									
2						−										
16			−								−					
17								−		−	−		−			
18			−		−										−	
19			−		−			−		−			−			−
20			−		−		*	+	+	+					−	
21	*	*								−		*			+	
22			−									*		+	+	
33							*				+					
34																−
35							*									
37	*															
38																−
49								−					−	−		
50									−				−			
51									−							
53									−							
63												*				
65														−		
66				−												
67		*		+							−					
68		*		+		+					−			−		
77	*															
87									−							
88							*							−		
89				−								*				
90				−		−										
92						−										
94								−		−				−		
95													−		−	
103												*				
112																−
113	*	*														−
123					−											
124					−											

Table D.5: Locating the most flexible regions of staphylococcal ribonuclease

	Fluctuations						**Torsion angles**									
							Psi					**Phi**				
Seq.	**PDB**		**MD MSF**				**PDB**	**MD torsion**				**PDB**	**MD torsion**			
index	**B**	**SqD**	**A**	**O**	**C**	**G**	**r**	**A**	**O**	**C**	**G**	**r**	**A**	**O**	**C**	**G**
6		*					*									
7					−											
8														−	−	
18											−					
19																−
45								−		−						
46	*					+									−	
47	*	*	+	+	+	+	*									
48	*		+	+	+	+	*		+							
49	*	*				+	*				+	*	+	+		
50	*	*	+			+		−		−		*				+
51												*				
78								−	−		−					
79										−			−	−		−
80													−		−	
95									−							
96														−		
97														−		
112											−					
113								−		−						−
114					−			−					−		−	
115			−		−			−		−			−		−	
116							*					*				
117				−								*				
119											−					
138																−
140			−	−												
141		*		+				−								

Bibliography

Adams M. J., Buehner M., Chandrasekhar K., Ford G. C., Hackert M. L., Liljas A., Rossmann M. G., Smiley I. E., Allison W. S., Everse J., Kaplan N. O. and Taylor S. S. (1973). Structure-function relationships in lactate dehydrogenase. *Proceedings of the National Academy of Sciences of the United States of America*, **70**: pp. 1968–1972.

Aliev A. E. and Courtier-Murias D. (2010). Experimental verification of force fields for molecular dynamics simulations using Gly-Pro-Gly-Gly. *Journal of Physical Chemistry B*, **114**: pp. 12358–12375.

Allen M. and Tildesley D. (1987). Computer Simulation of Liquids. Oxford University Press, Oxford, UK.

Anfinsen C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**: pp. 223–230.

Artymiuk P. J. and Blake C. C. (1981). Refinement of human lysozyme at 1.5 A resolution analysis of non-bonded and hydrogen-bond interactions. *Journal of Molecular Biology*, **152**: pp. 737–762.

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

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

Bai X.-c., McMullan G. and Scheres S. H. W. (2015). How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, **40**: pp. 49–57.

Balakrishnan G., Weeks C. L., Ibrahim M., Soldatova A. V. and Spiro T. G. (2008). Protein dynamics from time resolved UV Raman spectroscopy. *Current Opinion in Structural Biology*, **18**: pp. 623–629.

Baldwin R. L. (2007). Energetics of protein folding. *Journal of Molecular Biology*, **371**: pp. 283–301.

Bañuelos S., Saraste M. and Djinović Carugo K. (1998). Structural comparisons of calponin homology domains: implications for actin binding. *Structure*, **6**: pp. 1419–1431.

Bartlett G. J., Porter C. T., Borkakoti N. and Thornton J. M. (2002). Analysis of Catalytic Residues in Enzyme Active Sites. *Journal of Molecular Biology*, **324**: pp. 105–121.

Beadle B. M. and Shoichet B. K. (2002). Structural bases of stability-function tradeoffs in enzymes. *Journal of Molecular Biology*, **321**: pp. 285–296.

Beauchamp K. a., Lin Y.-S., Das R. and Pande V. S. (2012). Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of Chemical Theory and Computation*, **8**: pp. 1409–1414.

Benaglia T., Chauveau D., Hunter D. R. and Young D. (2009). mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, **32**: pp. 1–29.

Berendsen H. J. and Hayward S. (2000). Collective protein dynamics in relation to function. *Current Opinion in Structural Biology*, **10**: pp. 165–169.

Berendsen H., Postma J., Gunsteren W. van and Hermans J. (1981). Interaction Models for Water in Relation to Protein Hydration. In: *Intermolecular Forces* ed. by B. Pullman. Springer Netherlands,

Berendsen H., Spoel D. van der and Drunen R. van (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, **91**: pp. 43–56.

Berisio R., Sica F., Lamzin V. S., Wilson K. S., Zagari A. and Mazzarella L. (2002). Atomic resolution structures of ribonuclease A at six pH values. *Acta Crystallographica. Section D, Biological Crystallography*, **58**: pp. 441–450.

Bernstein F. C., Koetzle T. F., Williams G. J., Meyer E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanouchi T. and Tasumi M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, **112**: pp. 535–542.

Best R. B., Buchete N.-V. and Hummer G. (2008). Are current molecular dynamics force fields too helical? *Biophysical Journal*, **95**: pp. L07–L09.

Blow D. (2002). Outline of Crystallography for Biologists. Oxford University Press, Oxford, UK.

Bourgeois D., Vallone B., Schotte F., Arcovito A., Miele A. E., Sciara G., Wulff M., Anfinrud P. and Brunori M. (2003). Complex landscape of protein structural dynamics unveiled by nanosecond Laue crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, **100**: pp. 8704–8709.

Braxenthaler M., Unger R., Auerbach D., Given J. A. and Moult J. (1997). Chaos in protein dynamics. *Proteins*, **29**: pp. 417–425.

Brooks B. and Karplus M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, **80**: pp. 6571–6575.

Bryngelson J. D. and Wolynes P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **84**: pp. 7524–7528.

Brzozowski A. M., Pike A. C., Dauter Z., Hubbard R. E., Bonn T., Engström O., Ohman L., Greene G. L., Gustafsson J. A. and Carlquist M. (1997). Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, **389**: pp. 753–758.

Carugo O. and Argos P. (1997). Correlation between side chain mobility and conformation in protein structures. *Protein Engineering*, **10**: pp. 777–787.

Cerutti D. S., Freddolino P. L., Duke R. E. and Case D. a. (2010). Simulations of a protein crystal with a high resolution X-ray structure: evaluation of force fields and water models. *Journal of Physical Chemistry B*, **114**: pp. 12811–12824.

Chandler D. (1986). Roles of classical dynamics and quantum dynamics on activated processes occurring in liquids. *Journal of Statistical Physics*, **42**: pp. 49–67.

Chang C.-C. and Lin C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**: 27:1–27:27.

Chapman H. N., Fromme P., Barty A., White T. A., Kirian R. A., Aquila A., Hunter M. S., Schulz J., DePonte D. P., Weierstall U., Doak R. B., Maia F. R. N. C., Martin A. V., Schlichting I., Lomb L., Coppola N., Shoeman R. L., Epp S. W., Hartmann R., Rolles D., Rudenko A., Foucar L., Kimmel N., Weidenspointner G., Holl P., Liang M., Barthelmess M., Caleman C., Boutet S., Bogan M. J., Krzywinski J., Bostedt C., Bajt S., Gumprecht L., Rudek B., Erk B., Schmidt C., Hömke A., Reich C., Pietschner D., Strüder L., Hauser

G., Gorke H., Ullrich J., Herrmann S., Schaller G., Schopper F., Soltau H., Kühnel K.-U., Messerschmidt M., Bozek J. D., Hau-Riege S. P., Frank M., Hampton C. Y., Sierra R. G., Starodub D., Williams G. J., Hajdu J., Timneanu N., Seibert M. M., Andreasson J., Rocker A., Jönsson O., Svenda M., Stern S., Nass K., Andritschke R., Schröter C.-D., Krasniqi F., Bott M., Schmidt K. E., Wang X., Grotjohann I., Holton J. M., Barends T. R. M., Neutze R., Marchesini S., Fromme R., Schorb S., Rupp D., Adolph M., Gorkhover T., Andersson I., Hirsemann H., Potdevin G., Graafsma H., Nilsson B. and Spence J. C. H. (2011). Femtosecond X-ray protein nanocrystallography. *Nature*, **470**: pp. 73–77.

Chaudhri A., Zarraga I. E., Kamerzell T. J., Brandt J. P., Patapoff T. W., Shire S. J. and Voth G. a. (2012). Coarse-Grained Modeling of the Self-Association of Therapeutic Monoclonal Antibodies. *Journal of Physical Chemistry B*,

Chen D.-H., Song J.-L., Chuang D. T., Chiu W. and Ludtke S. J. (2006). An expanded conformation of single-ring GroEL-GroES complex encapsulates an 86 kDa substrate. *Structure*, **14**: pp. 1711–1722.

Chiti F. and Dobson C. M. (2006). Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry*, **75**: pp. 333–366.

Cino E. a., Choy W.-Y. and Karttunen M. (2012). Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, **8**: pp. 2725–2740.

Clore G. M. and Schwieters C. D. (2006). Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small α/β protein: a unified picture of high probability, fast atomic motions in proteins. *Journal of molecular biology*, **355**: pp. 879–886.

Cock P. J. A., Antao T., Chang J. T., Chapman B. A., Cox C. J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B. and de Hoon M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**: pp. 1422–1423.

Cohen A. E., Soltis S. M., González A., Aguila L., Alonso-Mori R., Barnes C. O., Baxter E. L., Brehmer W., Brewster A. S., Brunger A. T., Calero G., Chang J. F., Chollet M., Ehrensberger P., Eriksson T. L., Feng Y., Hattne J., Hedman B., Hollenbeck M., Holton J. M., Keable S., Kobilka B. K., Kovaleva E. G., Kruse A. C., Lemke H. T., Lin G., Lyubimov A. Y., Manglik A., Mathews I. I., McPhillips S. E., Nelson S., Peters J. W., Sauter N. K., Smith C. A., Song J., Stevenson H. P., Tsai Y., Uervirojnangkoorn M., Vinetsky V., Wakatsuki S., Weis W. I., Zadvornyy O. A., Zeldin O. B., Zhu D. and Hodgson K. O. (2014). Goniometer-based femtosecond crystallography with X-ray free electron lasers. *Proceedings of the National Academy of Sciences of the United States of America*, **111**: pp. 17122–17127.

Commons Math Developers (2013). Apache Commons Math. Version 3.2: The Apache Software Foundation.

Cuff M., Bigelow L., Abdullah J., Collart F. and Joachimiak A. (2005). 'Structure of a Protein of Unknown Function from Bacteroides thetaiotaomicron'. Unpublished PDB structure from the Midwest Center for Structural Genomics (MCSG).

Deniz A. A., Laurence T. A., Beligere G. S., Dahan M., Martin A. B., Chemla D. S., Dawson P. E., Schultz P. G. and Weiss S. (2000). Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proceedings of the National Academy of Sciences of the United States of America*, **97**: pp. 5179–5184.

Dill K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**: pp. 7133–7155.

Dobson C. M. (2004). Principles of protein folding, misfolding and aggregation. *Seminars in Cell and Developmental Biology*, **15**: pp. 3–16.

Donne D. G., Viles J. H., Groth D., Mehlhorn I., James T. L., Cohen F. E., Prusiner S. B., Wright P. E. and Dyson H. J. (1997). Structure of the recombinant full-length hamster prion protein PrP(29-231): the N terminus is highly flexible. *Proceedings of the National Academy of Sciences of the United States of America*, **94**: pp. 13452–13457.

Doruker P., Atilgan a. R. and Bahar I. (2000). Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins*, **40**: pp. 512–524.

Drenth J. (2007). Principles of Protein X-Ray Crystallography. Springer-Verlag, New York, USA.

Dunn M. F. (2005). Zinc-ligand interactions modulate assembly and stability of the insulin hexamer – a review. *Biometals*, **18**: pp. 295–303.

Durbin S. D. and Feher G. (1996). Protein crystallization. *Annual Review of Physical Chemistry*, **47**: pp. 171–204.

Eastman P., Pellegrini M. and Doniach S. (1999). Protein flexibility in solution and in crystals. *Journal of Chemical Physics*, **110**: pp. 10141–10152.

EclipseLink Project (2013). EclipseLink persistence libraries. Version 2.4.2: The Eclipse Foundation.

Eftink M. R. and Ghiron C. A. (1975). Dynamics of a protein matrix revealed by fluorescence quenching. *Proceedings of the National Academy of Sciences of the United States of America*, **72**: pp. 3290–3294.

Eisenmesser E. Z., Millet O., Labeikovsky W., Korzhnev D. M., Wolf-Watz M., Bosco D. A., Skalicky J. J., Kay L. E. and Kern D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, **438**: pp. 117–121.

Eklund H. and Brändén C. I. (1979). Structural differences between apo- and holoenzyme of horse liver alcohol dehydrogenase. *Journal of Biological Chemistry*, **254**: pp. 3458–3461.

Englander S. W., Mayne L., Bai Y. and Sosnick T. R. (1997). Hydrogen exchange: the modern legacy of Linderstrøm-Lang. *Protein Science*, **6**: pp. 1101–1109.

Eyal E., Chennubhotla C., Yang L.-W. and Bahar I. (2007). Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. *Bioinformatics*, **23**: pp. i175–84.

Eyal E., Gerzon S., Potapov V., Edelman M. and Sobolev V. (2005). The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *Journal of Molecular Biology*, **351**: pp. 431–442.

Feng W., Tejero R., Zimmerman D. E., Inouye M. and Montelione G. T. (1998). Solution NMR structure and backbone dynamics of the major cold-shock protein (CspA) from Escherichia coli: evidence for conformational dynamics in the single-stranded RNA-binding site. *Biochemistry*, **37**: pp. 10881–10896.

Fenwick R. B., van den Bedem H., Fraser J. S. and Wright P. E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proceedings of the National Academy of Sciences of the United States of America*, **111**: E445–E454.

Fields P. A. (2001). Review: Protein function at thermal extremes: balancing stability and flexibility. *Comparative Biochemistry and Physiology. Part A, Molecular and Integrative Physiology*, **129**: pp. 417–431.

Fischer M. W., Zeng L., Majumdar A. and Zuiderweg E. R. (1998). Characterizing semilocal motions in proteins by NMR relaxation studies. *Proceedings of the National Academy of Sciences of the United States of America*, **95**: pp. 8016–8019.

Fraser J. S., Clarkson M. W., Degnan S. C., Erion R., Kern D. and Alber T. (2009). Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, **462**: pp. 669–673.

Fraser J. S., van den Bedem H., Samelson A. J., Lang P. T., Holton J. M., Echols N. and Alber T. (2011). Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, **108**: pp. 16247–16252.

Frauenfelder H. and Petsko G. A. (1980). Structural dynamics of liganded myoglobin. *Biophysical Journal*, **32**: pp. 465–483.

Gamma E., Helm R., Johnson R. and Vlissides J. (1995). Design Patterns: Elements of Reusable Object-oriented Software. Addison-Wesley, Boston, MA, USA.

Ganim Z., Chung H. S., Smith A. W., Deflores L. P., Jones K. C. and Tokmakoff A. (2008). Amide I two-dimensional infrared spectroscopy of proteins. *Accounts of Chemical Research*, **41**: pp. 432–441.

Georlette D., Blaise V., Collins T., D'Amico S., Gratia E., Hoyoux A., Marx J.-C., Sonan G., Feller G. and Gerday C. (2004). Some like it cold: biocatalysis at low temperatures. *FEMS Microbiology Reviews*, **28**: pp. 25–42.

Go N., Noguti T. and Nishikawa T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences of the United States of America*, **80**: pp. 3696–3700.

Guha R., Howard M. T., Hutchison G. R., Murray-Rust P., Rzepa H., Steinbeck C., Wegner J. and Willighagen E. L. (2006). The Blue Obelisk-interoperability in chemical informatics. *Journal of Chemical Information and Modeling*, **46**: pp. 991–998.

Ha T., Enderle T., Ogletree D. F., Chemla D. S., Selvin P. R. and Weiss S. (1996). Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proceedings of the National Academy of Sciences of the United States of America*, **93**: pp. 6264–6268.

Hafner J. and Zheng W. (2011). All-atom modeling of anisotropic atomic fluctuations in protein crystal structures. *Journal of Chemical Physics*, **135**: p. 144114.

Haile J. (1992). Molecular dynamics simulation. Wiley, New York, USA.

Haliloglu T., Bahar I. and Erman B. (1997). Gaussian Dynamics of Folded Proteins. *Physical Review Letters*, **79**: pp. 3090–3093.

Halle B. (2002). Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **99**: pp. 1274–1279.

Hamelryck T. (2005). An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins*, **59**: pp. 38–48.

Hannay J. H. and Nye J. F. (2004). Fibonacci numerical integration on a sphere. *Journal of Physics A: Mathematical and General*, **37**: pp. 11591–11601.

Hansen D. F., Vallurupalli P. and Kay L. E. (2008). Using relaxation dispersion NMR spectroscopy to determine structures of excited, invisible protein states. *Journal of Biomolecular NMR*, **41**: pp. 113–120.

Hayre N. R., Singh R. R. P. and Cox D. L. (2011). Evaluating force field accuracy with long-time simulations of a β-hairpin tryptophan zipper peptide. *Journal of Chemical Physics*, **134**: p. 035103.

Hendrickson W. A. (1985). Stereochemically restrained refinement of macromolecular structures. *Methods in Enzymology*, **115**: pp. 252–270.

Henzler-Wildman K. A., Thai V., Lei M., Ott M., Wolf-Watz M., Fenn T., Pozharski E., Wilson M. A., Petsko G. A., Karplus M., Hübner C. G. and Kern D. (2007). Intrinsic motions along an enzymatic reaction trajectory. *Nature*, **450**: pp. 838–844.

Henzler-Wildman K. and Kern D. (2007). Dynamic personalities of proteins. *Nature*, **450**: pp. 964–972.

Herczenik E. and Gebbink M. F. B. G. (2008). Molecular and cellular aspects of protein misfolding and disease. *FASEB Journal*, **22**: pp. 2115–2133.

Hess B., Kutzner C., Spoel D. van der and Lindahl E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, **4**: pp. 435–447.

Hicklin J., Moler C., Webb P., Boisvert R. F., Miller B., Pozo R. and Remington K. (2012). JAMA : A Java Matrix Package. Version 1.0.3:

Hinsen K. (2008). Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics*, **24**: pp. 521–528.

Hollenstein K., Dawson R. J. P. and Locher K. P. (2007). Structure and mechanism of ABC transporter proteins. *Current Opinion in Structural Biology*, **17**: pp. 412–418.

Hollenstein K., Frei D. C. and Locher K. P. (2007). Structure of an ABC transporter in complex with its binding protein. *Nature*, **446**: pp. 213–216.

Holliday G. L., Almonacid D. E., Mitchell J. B. and Thornton J. M. (2007). The Chemistry of Protein Catalysis. *Journal of Molecular Biology*, **372**: pp. 1261–1277.

Hornak V., Abel R., Okur A., Strockbine B., Roitberg A. and Simmerling C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, **65**: pp. 712–725.

HSQL Development Group (2012). HyperSQL Database. Version 2.2.9:

Hu Z. and Jiang J. (2010). Assessment of biomolecular force fields for molecular dynamics simulations in a protein crystal. *Journal of Computational Chemistry*, **31**: pp. 371–380.

Humphrey W., Dalke A. and Schulten K. (1996). VMD: visual molecular dynamics. *Journal of Molecular Graphics*, **14**: pp. 33–8, 27–8.

Hunter M. S. and Fromme P. (2011). Toward structure determination using membrane-protein nanocrystals and microcrystals. *Methods*, **55**: pp. 387–404.

Hvidt A. and Linderstrøm-Lang K. (1955). Exchange of deuterium and 18O between water and other substances. III. Deuterium exchange of short peptides, Sanger's A-chain and insulin. *C R Trav Lab Carlsberg Chim*, **29**: pp. 385–402.

Hynes T. R. and Fox R. O. (1991). The crystal structure of staphylococcal nuclease refined at 1.7 A resolution. *Proteins*, **10**: pp. 92–105.

Jaenicke R. (1991). Protein stability and molecular adaptation to extreme conditions. *European Journal of Biochemistry*, **202**: pp. 715–728.

Jia Y., Talaga D. S., Lau W. L., Lu H. S., DeGrado W. F. and Hochstrasser R. M. (1999). Folding dynamics of single GCN-4 peptides by fluorescence resonant energy transfer confocal microscopy. *Chemical physics*, **247**: pp. 69–83.

Jorgensen W. L., Chandrasekhar J., Madura J. D., Impey R. W. and Klein M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, **79**: pp. 926–935.

Jorgensen W. L., Maxwell D. S. and Tirado-Rives J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, **118**: pp. 11225–11236.

Kabsch W. and Sander C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**: pp. 2577–2637.

Kabsch W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, **32**: pp. 922–923.

Kabsch W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, **34**: pp. 827–828.

Kamerzell T. J. and Middaugh C. R. (2008). The complex inter-relationships between protein flexibility and stability. *Journal of Pharmaceutical Sciences*, **97**: pp. 3494–3517.

Kamerzell T. J., Ramsey J. D. and Middaugh C. R. (2008). Immunoglobulin dynamics, conformational fluctuations, and nonlinear elasticity and their effects on stability. *Journal of Physical Chemistry B*, **112**: pp. 3240–3250.

Kaminski G. A., Friesner R. A., Tirado-Rives J. and Jorgensen W. L. (2001). Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *Journal of Physical Chemistry B*, **105**: pp. 6474–6487.

Kang Y., Zhou X. E., Gao X., He Y., Liu W., Ishchenko A., Barty A., White T. A., Yefanov O., Han G. W., Xu Q., de Waal P. W., Ke J., Tan M. H. E., Zhang C., Moeller A., West G. M., Pascal B. D., Van Eps N., Caro L. N., Vishnivetskiy S. A., Lee R. J., Suino-Powell K. M., Gu X., Pal K., Ma J., Zhi X., Boutet S., Williams G. J., Messerschmidt M., Gati C., Zatsepin N. A., Wang D., James D., Basu S., Roy-Chowdhury S., Conrad C. E., Coe J., Liu H., Lisova S., Kupitz C., Grotjohann I., Fromme R., Jiang Y., Tan M., Yang H., Li J., Wang M., Zheng Z., Li D., Howe N., Zhao Y., Standfuss J., Diederichs K., Dong Y., Potter C. S., Carragher B., Caffrey M., Jiang H., Chapman H. N., Spence J. C. H., Fromme P., Weierstall U., Ernst O. P., Katritch V., Gurevich V. V., Griffin P. R., Hubbell W. L., Stevens R. C., Cherezov V., Melcher K. and Xu H. E. (2015). Crystal structure of rhodopsin bound to arrestin by femtosecond X-ray laser. *Nature*, **523**: pp. 561–567.

Karplus M. and McCammon J. A. (1983). Dynamics of proteins: elements and function. *Annual Review of Biochemistry*, **52**: pp. 263–300.

Karplus P. A. and Schulz G. E. (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften*, **72**: pp. 212–213.

Kay L. E., Torchia D. A. and Bax A. (1989). Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry*, **28**: pp. 8972–8979.

Keedy D. A., Kenner L. R., Warkentin M., Woldeyes R. A., Hopkins J. B., Thompson M. C., Brewster A. S., Van Benschoten A. H., Baxter E. L., Uervirojnangkoorn M., McPhillips S. E., Song J., Alonso-Mori R., Holton J. M., Weis W. I., Brunger A. T., Soltis S. M., Lemke H., Gonzalez A., Sauter N. K., Cohen A. E., van den Bedem H., Thorne R. E. and Fraser J. S. (2015). Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. *eLife*, **4**:

Kendrew J. C., Bodo G., Dintzis H. M., Parrish R., Wyckoff H. and Phillips D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, **181**: pp. 662–666.

Kikhney A. G. and Svergun D. I. (2015). A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters*, **589**: pp. 2570–2577.

Kleckner I. R. and Foster M. P. (2011). An introduction to NMR-based approaches for measuring protein dynamics. *Biochimica et Biophysica Acta*, **1814**: pp. 942–968.

Koeniger S. L., Merenbloom S. I. and Clemmer D. E. (2006). Evidence for many resolvable structures within conformation types of electrosprayed ubiquitin ions. *Journal of Physical Chemistry B*, **110**: pp. 7017–7021.

Komsta L. and Novomestky F. (2012). moments: Moments, cumulants, skewness, kurtosis and related tests:

Kondrashov D. a., Van Wynsberghe A. W., Bannen R. M., Cui Q. and Phillips G. N. (2007). Protein structural variation in computational models and crystallographic data. *Structure*, **15**: pp. 169–177.

Kondrashov D. A., Zhang W., Aranda R., Stec B. and Phillips G. N. (2008). Sampling of the native conformational ensemble of myoglobin via structures in different crystalline environments. *Proteins*, **70**: pp. 353–362.

Konermann L., Pan J. and Liu Y.-H. (2011). Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chemical Society Reviews*, **40**: pp. 1224–1234.

Koshland D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **44**: pp. 98–104.

Krieger E., Darden T., Nabuurs S. B., Finkelstein A. and Vriend G. (2004). Making optimal use of empirical energy functions: force-field parameterization in crystal space. *Proteins*, **57**: pp. 678–683.

Kundu S., Melton J. S., Sorensen D. C. and Phillips G. N. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophysical Journal*, **83**: pp. 723–732.

Kupitz C., Basu S., Grotjohann I., Fromme R., Zatsepin N. A., Rendek K. N., Hunter M. S., Shoeman R. L., White T. A., Wang D., James D., Yang J.-H., Cobb D. E., Reeder B., Sierra R. G., Liu H., Barty A., Aquila A. L., Deponte D., Kirian R. A., Bari S., Bergkamp J. J., Beyerlein K. R., Bogan M. J., Caleman C., Chao T.-C., Conrad C. E., Davis K. M., Fleckenstein H., Galli L., Hau-Riege S. P., Kassemeyer S., Laksmono H., Liang M., Lomb L., Marchesini S., Martin A. V., Messerschmidt M., Milathianaki D., Nass K., Ros A., Roy-Chowdhury S., Schmidt K., Seibert M., Steinbrener J., Stellato F., Yan L., Yoon C., Moore T. A., Moore A. L., Pushkar Y., Williams G. J., Boutet S., Doak R. B., Weierstall U., Frank M., Chapman H. N., Spence J. C. H. and Fromme P. (2014). Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature*, **513**: pp. 261–265.

Kuriyan J. and Weis W. I. (1991). Rigid protein motion as a model for crystallographic temperature factors. *Proceedings of the National Academy of Sciences of the United States of America*, **88**: pp. 2773–2777.

Kuzmanic A., Pannu N. S. and Zagrovic B. (2014). X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature Communications*, **5**:

Lang P. T., Holton J. M., Fraser J. S. and Alber T. (2014). Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proceedings of the National Academy of Sciences of the United States of America*, **111**: pp. 237–242.

Lang P. T., Ng H.-L., Fraser J. S., Corn J. E., Echols N., Sales M., Holton J. M. and Alber T. (2010). Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Science*, **19**: pp. 1420–1431.

Lee B. and Richards F. (1971). The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, **55**: 379–IN4.

Lei H. and Duan Y. (2007). Improved sampling methods for molecular simulation. *Current Opinion in Structural Biology*, **17**: pp. 187–191.

Leopold P. E., Montal M. and Onuchic J. N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America*, **89**: pp. 8721–8725.

Levitt M., Sander C. and Stern P. S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of Molecular Biology*, **181**: pp. 423–447.

Lindahl E., Hess B. and Spoel D. V. D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling Journal of Molecular Modeling*, pp. 306–317.

Lindorff-Larsen K., Maragakis P., Piana S., Eastwood M. P., Dror R. O. and Shaw D. E. (2012). Systematic validation of protein force fields against experimental data. *PLoS One*, **7**: e32131.

Lindorff-Larsen K., Piana S., Palmo K., Maragakis P., Klepeis J. L., Dror R. O. and Shaw D. E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, **78**: pp. 1950–1958.

Lipscomb W. N. (1973). Enzymatic activities of carobxypeptidase A's in solution and in crystals. *Proceedings of the National Academy of Sciences of the United States of America*, **70**: pp. 3797–3801.

Liu R., Jiang W. and Zhou Y. (2010). Identifying protein-protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area. *Amino Acids*, **38**: pp. 263–270.

Liu W., Wacker D., Gati C., Han G. W., James D., Wang D., Nelson G., Weierstall U., Katritch V., Barty A., Zatsepin N. A., Li D., Messerschmidt M., Boutet S., Williams G. J., Koglin J. E., Seibert M. M., Wang C., Shah S. T. A., Basu S., Fromme R., Kupitz C., Rendek K. N., Grotjohann I., Fromme P., Kirian R. A., Beyerlein K. R., White T. A., Chapman H. N., Caffrey M., Spence J. C. H., Stevens R. C. and Cherezov V. (2013). Serial femtosecond crystallography of G protein-coupled receptors. *Science*, **342**: pp. 1521–1524.

Lomb L., Barends T. R. M., Kassemeyer S., Aquila A., Epp S. W., Erk B., Foucar L., Hartmann R., Rudek B., Rolles D., Rudenko A., Shoeman R. L., Andreasson J., Bajt S., Barthelmess M., Barty A., Bogan M. J., Bostedt C., Bozek J. D., Caleman C., Coffee R., Coppola N., Deponte D. P., Doak R. B., Ekeberg T., Fleckenstein H., Fromme P., Gebhardt M., Graafsma H., Gumprecht L., Hampton C. Y., Hartmann A., Hauser G., Hirsemann H., Holl P., Holton J. M., Hunter M. S., Kabsch W., Kimmel N., Kirian R. A., Liang M., Maia F. R. N. C., Meinhart A., Marchesini S., Martin A. V., Nass K., Reich C., Schulz J., Seibert M. M., Sierra R., Soltau H., Spence J. C. H., Steinbrener J., Stellato F., Stern S., Timneanu N., Wang X., Weidenspointner G., Weierstall U., White T. A., Wunderer C., Chapman H. N., Ullrich J., Strüder L. and Schlichting I. (2011). Radiation damage in protein serial femtosecond crystallography using an X-ray free-electron laser. *Physical Review B*, **84**: p. 214111.

Ludtke S. J., Chen D.-H., Song J.-L., Chuang D. T. and Chiu W. (2004). Seeing GroEL at 6 A resolution by single particle electron cryomicroscopy. *Structure*, **12**: pp. 1129–1136.

Ma J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, **13**: pp. 373–380.

MacKerell A. D., Bashford D., Dunbrack R. L., Evanseck J. D., Field M. J., Fischer S., Gao J., Guo H., Ha S., Joseph-McCarthy D., Kuchnir L., Kuczera K., Lau F. T. K., Mattos C., Michnick S., Ngo T., Nguyen D. T., Prodhom B., Reiher W. E., Roux B., Schlenkrich M., Smith J. C., Stote R., Straub J., Watanabe M., Wiórkiewicz-Kuczera J., Yin D. and Karplus M. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *Journal of Physical Chemistry B*, **102**: pp. 3586–3616.

Mackerell A. D., Feig M. and Brooks C. L. (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reprodu-

cing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, **25**: pp. 1400–1415.

Makinen M. W. and Fink A. L. (1977). Reactivity and cryoenzymology of enzymes in the crystalline state. *Annual Review of Biophysics and Bioengineering*, **6**: pp. 301–343.

Marion D. (2013). An introduction to biological NMR spectroscopy. *Molecular and Cellular Proteomics*, **12**: pp. 3006–3025.

Marrink S. J., Risselada H. J., Yefimov S., Tieleman D. P. and Vries A. H. de (2007). The MARTINI force field: coarse grained model for biomolecular simulations. *Journal of Physical Chemistry B*, **111**: pp. 7812–7824.

Matsumura M., Wozniak J. A., Sun D. P. and Matthews B. W. (1989). Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. *Journal of Biological Chemistry*, **264**: pp. 16059–16066.

McNaught A. and Wilkinson A. (1997). Compendium of Chemical Terminology: IUPAC Recommendations. Blackwell Scientific Publications, Oxford.

Meinhold L. and Smith J. C. (2005). Fluctuations and correlations in crystalline protein dynamics: a simulation analysis of staphylococcal nuclease. *Biophysical Journal*, **88**: pp. 2554–2563.

Merritt E. A. (2012). To B or not to B: a question of resolution? *Acta Crystallographica Section D: Biological Crystallography*, **68**: pp. 468–477.

Mertens H. D. T. and Svergun D. I. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *Journal of Structural Biology*, **172**: pp. 128–141.

Meyer D., Dimitriadou E., Hornik K., Weingessel A. and Leisch F. (2014). e1071: Misc. Functions of the Department of Statistics (e1071), The Vienna University of Technology - TU Wien:

Michaud-Agrawal N., Denning E. J., Woolf T. B. and Beckstein O. (2011). MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, **32**: pp. 2319–2327.

Michie A. D., Orengo C. A. and Thornton J. M. (1996). Analysis of domain structural class using an automated class assignment protocol. *Journal of Molecular Biology*, **262**: pp. 168–185.

Mittermaier A. K. and Kay L. E. (2009). Observing biological dynamics at atomic resolution using NMR. *Trends in Biochemical Sciences*, **34**: pp. 601–611.

Moss G. (1996). Basic terminology of stereochemistry (IUPAC Recommendations 1996). *Pure and Applied Chemistry*, **68**: pp. 2193–2222.

Mozzarelli A. and Rossi G. L. (1996). Protein function in the crystal. *Annual Review of Biophysics and Biomolecular Structure*, **25**: pp. 343–365.

Mukherjee P., Kass I., Arkin I. T., Arkin I. and Zanni M. T. (2006). Picosecond dynamics of a membrane protein revealed by 2D IR. *Proceedings of the National Academy of Sciences of the United States of America*, **103**: pp. 3528–3533.

Nar H., Schmid A., Puder C. and Potterat O. (2010). High-resolution crystal structure of a lasso Peptide. *Chemmedchem*, **5**: pp. 1689–1692.

Neutze R., Wouts R., van der Spoel D., Weckert E. and Hajdu J. (2000). Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, **406**: pp. 752–757.

Nic M., Jirat J., Kosat B. and Jenkins A. (2014). IUPAC Gold Book. Version 2.3.3: IUPAC.

Nishikawa K. and Ooi T. (1980). Prediction of the surface-interior diagram of globular proteins by an empirical method. *International Journal of Peptide and Protein Research*, **16**: pp. 19–32.

Norvell J. C., Nunes A. C. and Schoenborn B. P. (1975). Neutron diffraction analysis of myoglobin: structure of the carbon monoxide derivative. *Science*, **190**: pp. 568–570.

Opron K., Xia K. and Wei G.-W. (2014). Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *The Journal of Chemical Physics*, **140**: p. 234105.

Pace C. N. (1975). The stability of globular proteins. *CRC Critical Reviews in Biochemistry*, **3**: pp. 1–43.

Parthasarathy S. and Murthy M. R. (1997). Analysis of temperature factor distribution in high-resolution protein structures. *Protein Science*, **6**: pp. 2561–2567.

Petsko G. A. and Ringe D. (1984). Fluctuations in protein structure from X-ray diffraction. *Annual Review of Biophysics and Bioengineering*, **13**: pp. 331–371.

Phillips S. E. (1980). Structure and refinement of oxymyoglobin at 1· 6 Å resolution. *Journal of Molecular Biology*, **142**: pp. 531–554.

Pollastri G., Baldi P., Fariselli P. and Casadio R. (2002). Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**: pp. 142–153.

Ponder J. W. and Case D. A. (2003). Force fields for protein simulations. *Advances in Protein Chemistry*, **66**: pp. 27–85.

Prade L., Engh R. A., Girod A., Kinzel V., Huber R. and Bossemeyer D. (1997). Staurosporine-induced conformational changes of cAMP-dependent protein kinase catalytic subunit explain inhibitory potential. *Structure*, **5**: pp. 1627–1637.

Privalov P. L. and Khechinashvili N. N. (1974). A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *Journal of Molecular Biology*, **86**: pp. 665–684.

Prlić A., Yates A., Bliven S. E., Rose P. W., Jacobsen J., Troshin P. V., Chapman M., Gao J., Koh C. H., Foisy S., Holland R., Rimsa G., Heuer M. L., Brandstätter-Müller H., Bourne P. E. and Willis S. (2012). BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**: pp. 2693–2695.

Pronk S., Páll S., Schulz R., Larsson P., Bjelkmar P., Apostolov R., Shirts M. R., Smith J. C., Kasson P. M., Spoel D. van der, Hess B. and Lindahl E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**: pp. 845–854.

Purdy M. D., Bennett B. C., McIntire W. E., Khan A. K., Kasson P. M. and Yeager M. (2014). Function and dynamics of macromolecular complexes explored by integrative structural and computational biology. *Current Opinion in Structural Biology*, **27**: pp. 138–148.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Radivojac P., Obradovic Z., Smith D. K., Zhu G., Vucetic S., Brown C. J., Lawson J. D. and Dunker A. K. (2004). Protein flexibility and intrinsic disorder. *Protein Science*, **13**: pp. 71–80.

Razvi A. and Scholtz J. M. (2006). Lessons in stability from thermophilic proteins. *Protein Science*, **15**: pp. 1569–1578.

Reichert D., Zinkevich T., Saalwächter K. and Krushelnitsky A. (2012). The relation of the X-ray B-factor to protein dynamics: insights from recent dynamic solid-state NMR data. *Journal of Biomolecular Structure and Dynamics*, **30**: pp. 617–627.

Riccardi D., Cui Q. and Phillips Jr G. N. (2009). Application of elastic network models to proteins in the crystalline state. *Biophysical Journal*, **96**: pp. 464–475.

Rosenberg A. and Chakravarti K. (1968). Studies of hydrogen exchange in proteins. I. The exchange kinetics of bovine carbonic anhydrase. *Journal of Biological Chemistry*, **243**: pp. 5193–5201.

Rosenberg A. and Enberg J. (1969). Studies of hydrogen exchange in proteins. II. The reversible thermal unfolding of chymotrypsinogen A as studied by exchange kinetics. *Journal of Biological Chemistry*, **244**: pp. 6153–6159.

Rueda M., Chacón P. and Orozco M. (2007). Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, **15**: pp. 565–575.

Saff E. B. and Kuijlaars A. B. J. (1997). Distributing many points on a sphere. *The Mathematical Intelligencer*, **19**: pp. 5–11.

Sali A. and Blundell T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**: pp. 779–815.

Sali A., Shakhnovich E. and Karplus M. (1994). How does a protein fold? *Nature*, **369**: pp. 248–251.

Schlessinger A., Yachdav G. and Rost B. (2006). PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22**: pp. 891–893.

Schlichting I. and Miao J. (2012). Emerging opportunities in structural biology with X-ray free-electron lasers. *Curr Opin Struct Biol*, **22**: pp. 613–626.

Schmid N., Eichenberger A. P., Choutko A., Riniker S., Winger M., Mark A. E. and van Gunsteren W. F. (2011). Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *European Biophysics Journal*, **40**: pp. 843–856.

Schneider B., Gelly J. C., de Brevern A. G. and Černý J. (2014). Local dynamics of proteins and DNA evaluated from crystallographic B factors. *Acta Crystallographica Section D: Biological Crystallography*, **70**: pp. 2413–2419.

Schomaker V. and Trueblood K. N. (1968). On the rigid-body motion of molecules in crystals. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, **24**: pp. 63–76.

Schotte F., Lim M., Jackson T. A., Smirnov A. V., Soman J., Olson J. S., Phillips Jr G. N., Wulff M. and Anfinrud P. A. (2003). Watching a protein as it functions with 150-ps time-resolved X-ray crystallography. *Science*, **300**: pp. 1944–1947.

Schuler B. and Eaton W. A. (2008). Protein folding studied by single-molecule FRET. *Current Opinion in Structural Biology*, **18**: pp. 16–26.

Schwarzenbach D. (2011). The success story of crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, **68**: pp. 57–67.

Scott W. R. P., Hünenberger P. H., Tironi I. G., Mark A. E., Billeter S. R., Fennen J., Torda A. E., Huber T., Krüger P. and Gunsteren W. F. van (1999). The GROMOS Biomolecular Simulation Program Package. *Journal of Physical Chemistry A*, **103**: pp. 3596–3607.

Shaw D., Dror R., Salmon J., Grossman J., Mackenzie K., Bank J., Young C., Deneroff M., Batson B., Bowers K., Chow E., Eastwood M., Ierardi D., Klepeis J., Kuskin J., Larson R., Lindorff-Larsen K., Maragakis P., Moraes M., Piana S., Shan Y. and Towles B. (2009). Millisecond-scale molecular dynamics simulations on Anton. In: *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*, pp. 1–11.

Sheldrick G. M. and Schneider T. R. (1997). SHELXL: high-resolution refinement. *Methods in Enzymology*, **277**: pp. 319–343.

Sheriff S., Hendrickson W. a., Stenkamp R. E., Sieker L. C. and Jensen L. H. (1985). Influence of solvent accessibility and intermolecular contacts on atomic mobilities in hemerythrins. *Proceedings of the National Academy of Sciences of the United States of America*, **82**: pp. 1104–1107.

Shoichet B. K., Baase W. A., Kuroki R. and Matthews B. W. (1995). A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences of the United States of America*, **92**: pp. 452–456.

Shrake A. and Rupley J. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, **79**: pp. 351–371.

Skjaerven L., Hollup S. M. and Reuter N. (2009). Normal mode analysis for proteins. *Journal of Molecular Structure: THEOCHEM*, **898**: pp. 42–48.

Smith D. K., Radivojac P., Obradovic Z., Dunker A. K. and Zhu G. (2003). Improved amino acid flexibility parameters. *Protein Science*, **12**: pp. 1060–1072.

Smith G. D., Blessing R. H., Ealick S. E., Fontecilla-Camps J. C., Hauptman H. A., Housset D., Langs D. A. and Miller R. (1997). Ab initio structure determination and refinement of a scorpion protein toxin. *Acta Crystallographica Section D: Biological Crystallography*, **53**: pp. 551–557.

Smith G. D., Swenson D. C., Dodson E. J., Dodson G. G. and Reynolds C. D. (1984). Structural stability in the 4-zinc human insulin hexamer. *Proceedings of the National Academy of Sciences of the United States of America*, **81**: pp. 7093–7097.

Smith G. D., Pangborn W. A. and Blessing R. H. (2003). The structure of T6 human insulin at 1.0 A resolution. *Acta Crystallographica Section D: Biological Crystallography*, **59**: pp. 474–482.

Soares T. A., Daura X., Oostenbrink C., Smith L. J. and van Gunsteren W. F. (2004). Validation of the GROMOS force-field parameter set 45Alpha3 against nuclear magnetic resonance data of hen egg lysozyme. *Journal of Biomolecular NMR*, **30**: pp. 407–422.

Soheilifard R., Makarov D. E. and Rodin G. J. (2008). Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. *Physical Biology*, **5**: p. 026008.

Sonavane S., Jaybhaye A. A. and Jadhav A. G. (2013). Prediction of temperature factors from protein sequence. *Bioinformation*, **9**: pp. 134–140.

Spiro T. G., Smulevich G. and Su C. (1990). Probing protein structure and dynamics with resonance Raman spectroscopy: cytochrome c peroxidase and hemoglobin. *Biochemistry*, **29**: pp. 4497–4508.

Stryer L. (1978). Fluorescence energy transfer as a spectroscopic ruler. *Annual Review of Biochemistry*, **47**: pp. 819–846.

Swinbank R. and Purser R. J. (2006). Fibonacci grids: A novel approach to global modelling. *Quarterly Journal of the Royal Meteorological Society*, **132**: pp. 1769–1793.

Tai K. (2004). Conformational sampling for the impatient. *Biophysical Chemistry*, **107**: pp. 213–220.

Taverna D. M. and Goldstein R. A. (2002). Why are proteins marginally stable? *Proteins*, **46**: pp. 105–109.

Teilum K., Olsen J. G. and Kragelund B. B. (2009). Functional aspects of protein flexibility. *Cellular and Molecular Life Sciences*, **66**: pp. 2231–2247.

Tenboer J., Basu S., Zatsepin N., Pande K., Milathianaki D., Frank M., Hunter M., Boutet S., Williams G. J., Koglin J. E., Oberthuer D., Heymann M., Kupitz C., Conrad C., Coe J., Roy-Chowdhury S., Weierstall U., James D., Wang D., Grant T., Barty A., Yefanov O., Scales J., Gati C., Seuring C., Srajer V., Henning R., Schwander P., Fromme R., Ourmazd A., Moffat K., Van Thor J. J., Spence J. C. H., Fromme P., Chapman H. N. and Schmidt M. (2014). Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science*, **346**: pp. 1242–1246.

Tertoolen L. G., Blanchetot C., Jiang G., Overvoorde J., Gadella Jr T., Hunter T. and den Hertog J. (2001). Dimerization of receptor protein-tyrosine phosphatase alpha in living cells. *BMC Cell Biology*, **2**: p. 8.

The Blue Obelisk Group (2013). Blue Obelisk Data Repository (BODR). Version 10:

Tirion M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, **77**: pp. 1905–1908.

Tomatis P. E., Fabiane S. M., Simona F., Carloni P., Sutton B. J. and Vila A. J. (2008). Adaptive protein evolution grants organismal fitness by improving catalysis and flexibility. *Proceedings of the National Academy of Sciences of the United States of America*, **105**: pp. 20605–20610.

Touw W. G. and Vriend G. (2014). BDB: databank of PDB files with consistent B-factors. *Protein Engineering, Design and Selection*, **27**: pp. 457–462.

Trueblood K. N., Bürgi H. B., Burzlaff H., Dunitz J. D., Gramaccioli C. M., Schulz H. H., Shmueli U. and Abrahams S. C. (1996). Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. *Acta Crystallographica Section A: Foundations of Crystallography*, **52**: pp. 770–781.

van den Bedem H. and Fraser J. S. (2015). Integrative, dynamic structural biology at atomic resolution–it's about time. *Nature Methods*, **12**: pp. 307–318.

Van Der Spoel D., Lindahl E., Hess B., Groenhof G., Mark A. E. and Berendsen H. J. C. (2005). GROMACS: fast, flexible, and free. *Journal of Computational Chemistry*, **26**: pp. 1701–1718.

Wales T. E. and Engen J. R. (2006). Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrometry Reviews*, **25**: pp. 158–170.

Wall M. E., Van Benschoten A. H., Sauter N. K., Adams P. D., Fraser J. S. and Terwilliger T. C. (2014). Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse X-ray scattering. *Proceedings of the National Academy of Sciences of the United States of America*, **111**: pp. 17887–17892.

Whittaker S. B., Boetzel R., MacDonald C., Lian L. Y., Pommer A. J., Reilly A., James R., Kleanthous C. and Moore G. R. (1998). NMR detection of slow conformational dynamics in an endonuclease toxin. *Journal of Biomolecular NMR*, **12**: pp. 145–159.

Wickham H. (2009). ggplot2: elegant graphics for data analysis. Springer New York, New York.

Wickham H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, **40**: pp. 1–29.

Wilcox R. (2010). Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy. Springer-Verlag, New York, USA.

Wilson K. P., Malcolm B. A. and Matthews B. W. (1992). Structural and thermodynamic analysis of compensating mutations within the core of chicken egg white lysozyme. *Journal of Biological Chemistry*, **267**: pp. 10842–10849.

Woldeyes R. A., Sivak D. A. and Fraser J. S. (2014). E pluribus unum, no more: from one crystal, many conformations. *Current Opinion in Structural Biology*, **28**: pp. 56–62.

Wolynes P. G. (2005). Energy landscapes and solved protein-folding problems. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, **363**: pp. 453–64, 453–64.

Xia K. and Wei G.-W. (2013). Stochastic model for protein flexibility analysis. *Physical Review E*, **88**: p. 062709.

Xue Y. and Skrynnikov N. R. (2014). Ensemble MD simulations restrained via crystallographic data: Accurate structure leads to accurate dynamics. *Protein Science*, **23**: pp. 488–507.

Yang L., Song G. and Jernigan R. L. (2007). How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophysical Journal*, **93**: pp. 920–929.

Yuan Z., Zhao J. and Wang Z.-X. (2003). Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Engineering, Design and Selection*, **16**: pp. 109–114.

Yuan Z., Bailey T. L. and Teasdale R. D. (2005). Prediction of protein B-factor profiles. *Proteins*, **58**: pp. 905–912.

Zagrovic B. and van Gunsteren W. F. (2006). Comparing atomistic simulation data with the NMR experiment: how much can NOEs actually tell us? *Proteins*, **63**: pp. 210–218.

Zar J. (2010). Biostatistical Analysis. Pearson Prentice Hall, Upper Saddle River, New Jersey.

Zhang H., Unal H., Gati C., Han G. W., Liu W., Zatsepin N. A., James D., Wang D., Nelson G., Weierstall U., Sawaya M. R., Xu Q., Messerschmidt M., Williams G. J., Boutet S., Yefanov O. M., White T. A., Wang C., Ishchenko A., Tirupula K. C., Desnoyer R., Coe J., Conrad C. E., Fromme P., Stevens R. C., Katritch V., Karnik S. S. and Cherezov V. (2015). Structure of the Angiotensin receptor revealed by serial femtosecond crystallography. *Cell*, **161**: pp. 833–844.

Zhang H., Zhang T., Chen K., Shen S., Ruan J. and Kurgan L. (2009). On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins*, **76**: pp. 617–636.

Zhou Z. H., Liao W., Cheng R. H., Lawson J. E., McCarthy D. B., Reed L. J. and Stoops J. K. (2001). Direct evidence for the size and conformational variability of the pyruvate dehydrogenase complex revealed by three-dimensional electron microscopy. The "breathing" core and its functional relationship to protein dynamics. *Journal of Biological Chemistry*, **276**: pp. 21704–21713.