Multi-scale Modelling of Allostery in Protein Homodimers

DAVID BURNELL

Multi-scale Modelling of Allostery in Protein Homodimers

David Burnell

A Thesis presented for the degree of
Doctor of Philosophy

Mark Wilson Research Group
Department Chemistry
University of Durham
England

2015

Abstract

Allostery is a form of signalling within biomolecules such that ligand binding to a protein affects its activity at a second site. Allostery was described by early models to be driven by structural changes in the protein. However, more recently there has been increasing evidence that dynamics can contribute to or even drive allostery. The protein studied in this thesis, the Catabolite Activator Protein (CAP), is an allosteric protein homodimer that has been shown to exhibit negatively cooperative binding of the ligand cyclic Adenosine Monophosphate (cAMP) to each of its monomers. Interestingly, CAP is a protein whose allostery is believed to be driven by dynamics rather than a conformational change.

In this thesis, a number of coarse grained models are employed to investigate this dynamic allostery in CAP. One family of models, termed Super Coarse Grained (SCG) models explore the global properties of the dynamics of the CAP dimer that cause it to exhibit negatively cooperative allostery. It is shown through these models that changes in protein interactions can provide a basis for changing cooperativity. A second family of coarse grained models called Elastic Network Models (ENM) are studied. These are used to show that adjusting the interactions between specific residues can affect cooperative binding of cAMP to CAP.

A number of atomistic approaches are also used to study the cAMP-CAP system, including Molecular Dynamics (MD) and Normal Mode Analysis (NMA). The efficacy of using such approaches for studying the thermodynamics of the allostery in CAP is investigated. The motion observed within the protein is also studied closely to identify potential allosteric pathways.

X-ray crystallography and Isothermal Titration Calorimetry (ITC) are finally used to investigate how accurately computational methods can describe the cooperative binding of cAMP to CAP. They are also used to try and determine whether the allostery in CAP can be manipulated experimentally without any observed changes to its structure.

Declaration

The work in this thesis is based on research carried out at the Biophysical Sciences Institute, the Department of Chemistry, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I would like to thank my supervisor Mark Wilson for his help and guidance throughout my studies. The knowledge that he has shared on the subject of atomistic and coarse grained simulations has been extremely beneficial.

I would also like to thank my second supervisor Ehmke Pohl for his guidance and support. His shared knowledge of X-ray crystallography and biochemistry has been immensely helpful.

I am very grateful to Dr. Tom Rodgers for his contribution to the project and his help on many of the models and simulations discussed in this thesis. I would also like to thank him for the useful conversations we have shared and his company during the first half of my studies. Likewise, I would also like to thank Dr. Fatima Chami for her help towards the end of the project.

Many thanks also go to Dr. Phil Townsend, who was a great help with my experimental work. I would also like to thank him for his contribution to the experimental part of this project.

I am also very grateful to the rest of the "Protein multitools" team, especially Tom McLeish and Martin Cann, the wisdom they have shared has been extremely beneficial to me. Likewise I am grateful to the other occupants of rooms CG200X and CG229.

I am also indebted to my family, my girlfriend and her family, who have supported me throughout my time in Durham. There are many more people who I would like to acknowledge either for their academic help throughout the project, or for generally improving my days. They should all know who they are.

I would finally like to acknowledge the financial support I have received from the EPSRC, for which I am very grateful.

Contents

List of Figures

List of Tables

List of Abbreviations

Coot	Crystallographic Object-Oriented Toolkit	119
Pfu	*Pyrococcus furiosus*	173
2D	Two Dimension(al)	107
3D	Three Dimension(al)	3
AA	All Atom	40
AFM	Atomic Force Microscopy	73
AMBER	Assisted Model Building with Energy Refinement	49
ANM	Anisotropic Network Model	54
ATP	Adenosine Triphosphate	26
BCA	Bovine Serum Albumin	183
cAMP	Cyclic Adenosine Monophosphate	12
CAP	Catabolite Activator Protein	3
CG	Coarse-Grained	59
CIAP	Calf Intestinal Alkaline Phosphatase	174
CP	Cyclic Peptide	164
CPPTRAJ	Command Prompt Process Trajectory	119
CPU	Central Processing Unit	51
DDPT	Durham Dynamic Protein Toolbox	100
DNA	Deoxyribonucleic Acid	8
DNABD	DNA Binding Domain	27
DTT	Dithiothreitol	183, 185

ECL	Enhanced Chemiluminescence	185, 186
EDTA	Ethylenediaminetetraacetic Acid	180
EM	Electron Microscopy	6
ENM	Elastic Network Model	54
ESI	Electrospray Ionization	185
GAFF	General AMBER Force Field	117
GNM	Gaussian Network Model	54
GPU	Graphics Processing Unit	118
GROMACS	GROningen MAchine for Chemical Simulations	118
ID	Identification Data	125
ITC	Isothermal Titration Calorimetry	12
L-BFGS	Limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm	44
LBD	Ligand Binding Domain	27
LDS	Lithium Dodecyl Sulphate	185
MAD	Multiple wavelength Anomalous Dispersion	85
MD	Molecular Dynamics	7
MIR	Multiple Isomorphous Replacement	85
MM/PBSA	Molecular Mechanics/Poisson Boltzmann Surface Area method	68
MMM	Mobile Monomer Model	96
MS	Mass Spectrometry	175
NMA	Normal Mode Analysis	34
NMR	Nuclear Magnetic Resonance	6
NOESY	Nuclear Overhauser Effect Spectroscopy	29
NP	Nonpolar	50

PAGE	Polyacrylamide Gel Electrophoresis	185
PB	Poisson-Boltzmann	48
PBC	Periodic Boundary Conditions	159
PC	Principal Component	64
PCA	Principal Component Analysis	34
PCR	Polymerase Chain Reaction	173
PDB	Protein Data Bank	6
PEG	Polyethylene Glycol	76
pI	Isoelectric point	77
PTS	Phototransferase System	26
RMS	Root-Mean-Square	202
RMSD	Root-Mean-Square Deviation	125
RMSF	Root-Mean-Square Fluctuation	24
RNA	Ribonucleic Acid	27
RTB	Rotational-Translational Block	56
SAD	Single wavelength Anomalous Dispersion	85
SASA	Solvent Accessible Surface Area	50
SCG	Super Coarse Grained	59
SDS	Sodium Dodecyl Sulphate	185, 186
SMM	Static Monomer Model	96
SOC	Super Optimal broth with Catabolite repression	181
TAE	Tris base, Acetic acid and EDTA	180
TMM	Truncated Monomer Model	96
UA	United Atom	40
VMD	Visual Molecular Dynamics	99

WT	Wild-Type	29

Part I

General Introduction

Chapter 1

Project Overview and Background

Proteins are biological molecules essential for life. Knowing their three dimension(al) (3D) structure contributes to understanding how proteins function. Proteins consist of amino acids joined by peptide bonds into a polypeptide chain. For most proteins, these chains of amino acids are folded into complex 3D structures. It is understood that proteins have evolved over time to perform specific functions dependent on these structures [1], however it was not until the 1970s that the importance of thermal fluctuations of proteins on their function was recognised [2].

This thesis explores the importance of structure and dynamics in a form of intramolecular signalling common in proteins called allostery. An example protein, where allostery is believed to occur without a global conformational change, the catabolite activator protein (CAP) is investigated in detail.

1.1 Protein Structure

Amino acids, the building blocks of proteins follow a specific template, with a central carbon atom (C_α) bonded to an amino group, a carboxyl group, a hydrogen atom and a side chain. The α-carbons of adjacent amino acids are separated by three covalent bonds arranged as C_α-C-N-C_α (see figure 1.1). The C-N bond for each amino acid is unable to freely rotate, as electron sharing occurs between the carbonyl oxygen and the amide nitrogen, forcing the C_α-C-N-C_α atoms into a planar conformation. The rigid peptide bonds therefore limit to some extent the number of conformations that can be adopted by the folded protein. Typically the torsional angles in the amino acids

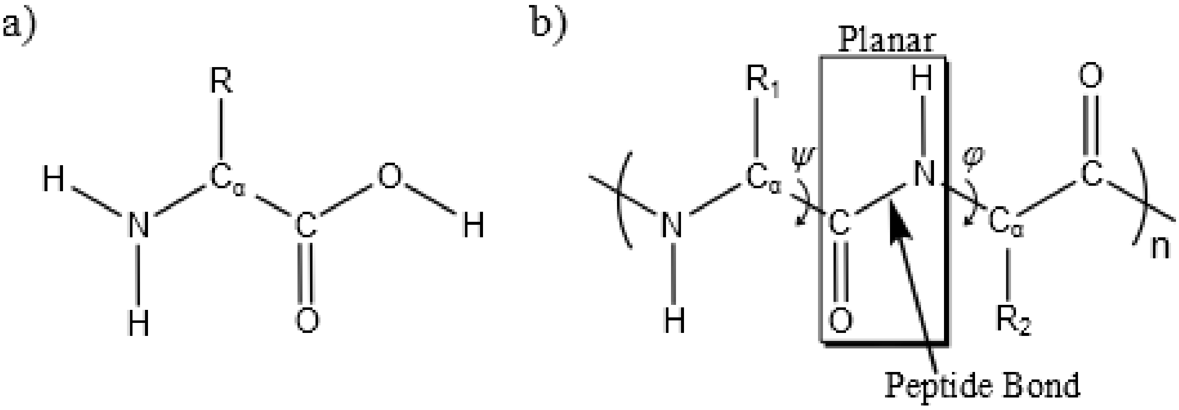


Figure 1.1: a) The basic structure of an amino acid showing the C_α atom and the amino acid side chain, R, and b) the peptide bond between adjacent amino acids in a polypeptide chain. The torsional angles ϕ and *psi* are displayed.

are labelled ϕ for C-N-C_α-C and ψ for N-C_α-C-N. The allowed values for ϕ and ψ are limited as many values are prohibited by steric interference. Ramachandran plots are a graphical representation that plots allowed values for ϕ and ψ and are useful for validating a protein's structure [3].

A protein's 3D structure is the result of folding of the polypeptide chain to create compact units with specific 3D structures, such as α-helices and β-sheets. Most proteins exist with more than one unique spacial arrangement of their atoms. These unique spacial arrangements, called conformations, exist at local minima on a complex multidimensional potential energy surface. Usually separate conformations will exist for the active and the inactive states of the protein [4].

The folded state of a protein is held together by disulphide bonds and many weaker intermolecular interactions, such as hydrophobic interactions, hydrogen bonding and ionic interactions. However, a protein's stability can not be calculated by simply summing these interactions; there are other factors to take into account. Firstly, all of the hydrogen bonds formed internally in the protein would include the breaking of a hydrogen bond between the protein and water prior to the protein folding. Secondly, the presence of the protein in water disrupts the hydrogen bonding network of pure water, causing the water to form a highly structured solvation shell around the protein. And thirdly, the entropic penalty associated with ordering the water around the protein, is worse for hydrophobic regions of the protein. Therefore, the protein folds in such a manner that the hydrophobic groups are shielded from the solvent as much as possible and the hydrophilic groups are found on the exterior. The gain in entropy from the

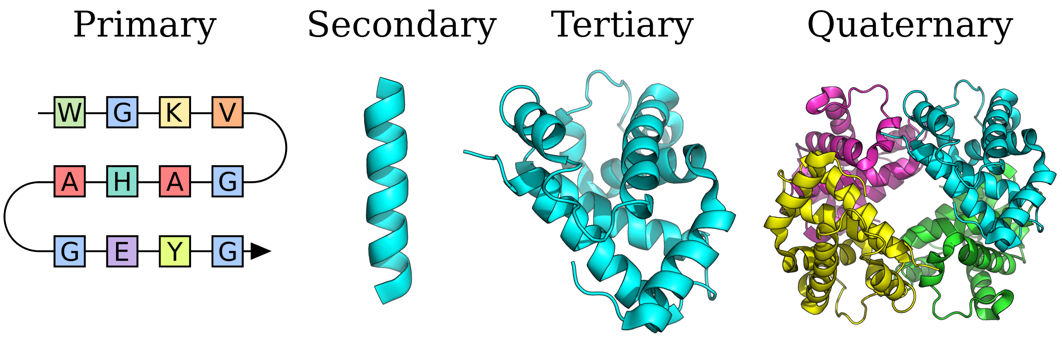


Figure 1.2: The primary, secondary, tertiary and quaternary structures of proteins. The primary structure is given by the ordering of the amino acids, the secondary is the structural elements such as an α-helix (shown) or a β-sheet. The tertiary structure gives the 3D arrangement of the protein monomer in space, while the quaternary structure is the 3D structure of the entire protein complex. Figure adapted from [1].

release of the structured water when the intramolecular interactions in the protein are formed compensates somewhat for the loss of conformational entropy associated with folding the protein into a single conformation.

A protein's structure is described over four levels. The primary structure of a protein is given by the sequence of amino acids or residues in the polypeptide chain. The secondary structure of a protein is the arrangement of local domains of the polypeptide chain into structural elements, where the structure of the amino acids in these local domains are spatially related to each other. The most common secondary structural elements are α helices, β sheets and β turns. An α helix is a helical structure in which the protein backbone is tightly wound around an axis drawn along the length of the helix with the side chains pointing away from axis of the helix. A single turn of the alpha helix extends approximately 5.4 Å along the length of the helical axis. The ϕ and ψ values of the residues in an α helix are -60° and -45° to -50° respectively. When the structure of the protein chain extend into a zigzag rather than a helical structure, it forms a β strand. These β strands stack in either a parallel or antiparallel manner to form β sheets. β turns connect two adjacent β strands in an antiparallel stacked β sheet. The tertiary structure is given by the 3D arrangement of these structural elements in space for the polypeptide chain. For a protein complex of two or more polypeptide chains, the arrangement of these chains in 3D space is the quaternary structure. A depiction of the four structure levels of a protein can be seen in figure 1.2.

The tertiary and quaternary structures of a protein can bring residues that seem distant in the primary structure, close together in 3D space to form a functional region of the protein such as a catalytic or binding site. Proteins are able to exist with more than one conformation, often having an active form and an inactive form. The conversion of a protein from its inactive form to its active form can be triggered by an event such as the binding of a small molecule.

Knowing a protein's 3D structure is a very powerful tool for understanding its biological function. However, predicting a protein's tertiary or quaternary structure from its primary structure is not yet possible. Currently the most powerful method is homology modelling, which makes use of known structures of protein domains that have similar sequences [5]. Typically homology modelling can work well with proteins that have greater than 30% sequence identity, however, as structure is conserved more than sequence, proteins that share no sequence similarity can share structural similarities. This makes predicting the protein's 3D structure such a difficult task.

Rather than calculating a protein's structure computationally, since 1958 when Kendrew solved the structure of Myoglobin [6], it has been possible to solve a protein's structure experimentally using X-ray crystallography. X-ray crystallography is able to resolve the average positions of the heavy atoms in the protein, generating a 3D static structure. Nuclear magnetic resonance (NMR) or electron microscopy (EM) can now also be used to solve a protein's crystal structure. The protein data bank (PDB) [7, 8] is a database of the 3D structures of biomolecules derived predominantly from crystallographic studies. At the time of writing, this contains over 100,000 structures. Around 90,000 are derived from X-ray crystallography, 10,000 from NMR and 1000 by EM and other methods. The protein data bank is a priceless tool for studying protein structure and dynamics.

1.2 Protein Dynamics

A protein's structure, however, is not static and the structure seen by X-ray crystallography is a high precision average structure located at a minimum on the complicated potential energy landscape. In reality thermal fluctuations of the protein allows it to explore a multitude of other conformations in the surrounding conformational space. It wasn't until the 1970s that Weber [2] indicated that the thermal fluctuations can be essential for protein function and Austin et al. [9] demonstrated this with Myoglobin, showing that an ensemble of conformational substates can affect ligand binding.

Over the last few decades a number of techniques have been used to study protein dynamics. With X-ray crystallography, for example, it is possible to identify flexible regions of proteins. However, it is often hard to gain useful dynamic information such as timescales from this. NMR is also very sensitive to protein motions [10, 11] and can resolve the dynamics of independent residues because of their distinct chemical shifts. Isotopic labelling can also be used as a further set of tools to study dynamics in specific residues or regions of the protein. NMR can also be performed with the protein in solution, and at room temperature, closer to native conditions than is possible with X-ray crystallography. The use of molecular dynamics (MD) to study protein dynamics also quickly gained popularity since the first simulations of a simplified protein folding process by Levitt and Warshel in 1975 [12].

The recent emphasis put on the importance dynamics on protein functionality is evident in the increased number of studies of protein dynamics over the past 10 years. A number of recent studies show the importance of dynamics in certain aspects in protein function such as ligand recognition [13], protein folding [14], enzyme catalysis [15], ion channel selectivity [16], and allosteric regulation [17, 18]. This new emphasis has led to a model of protein evolution predicting that they have evolved on a dynamic basis, alongside a structural basis, to fulfil their role within the cell [19].

Henzler-Wildman and Kern defined protein dynamics as any time-dependent change in atomic coordinates [20]. This can of course refer to both equilibrium fluctuations, or other non-equilibrium phenomena. The majority of biological processes, including allostery (section 1.4), are regulated by equilibrium fluctuations. These range from fast motions on the fs to ns timescale to slow motions on the microsecond to second timescale. The fast motions indicate motions caused by local flexibility including bond vibrations and side chain motions. The slow motions are the biologically important collective motions of much larger amplitude such as large domain motions and motions on the timescale of protein folding. Figure 1.3 displays a break-down of these motions.

This thesis primarily investigates slow collective protein motions. It can be argued that these motions are biologically more important than the fast modes as they involve large rearrangements felt over large distances, whereas the fast modes are localised so are unlikely to cause long distance signalling such as that seen in allostery. Protein motions on timescales less than 100 ps were argued in a review article by Daniel *et*

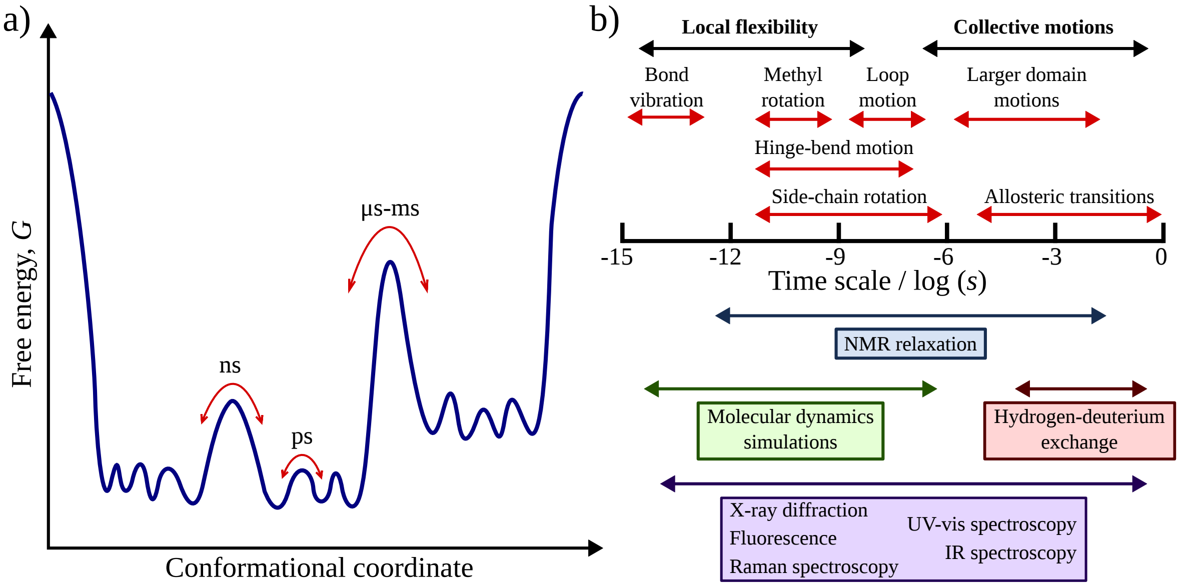


Figure 1.3: The large spectrum of timescales observed in the thermal fluctuations of proteins. (a) A schematic protein potential energy landscape along one conformational coordinate showing barrier transitions between local minima corresponding to motions on three different time scales. (b) Examples of motions observed across the spectrum of protein timescales and the techniques used to study protein motion at corresponding timescales. Adapted from [20], with extra information from [21]

al. [22] to be unnecessary for enzyme function. There is some disagreement between the timescale of motions active in allostery [17, 23–25]. However, the most likely answer is that motion at most timescales contribute in some way to the signalling. Furthermore, it has also been shown that a change in a protein's backbone shape is not always necessary for allosteric signalling [17]. This report will investigate further, allostery that is driven by protein dynamics rather than a structural change.

1.3 Thermodynamics of Protein-Ligand Binding Events

The function of many proteins is dependent on the reversible binding of regulatory molecules, termed ligands. Ligands can be anything as small as an ion, or can be as large as deoxyribonucleic acid (DNA) or another protein. Proteins are able to interact selectively with ligands, only binding one or a select few at specific binding sites. This process is of utmost importance for a protein's function and is critical to life itself.

The original view of protein-ligand binding was that the shape of the ligand was complementary to the shape of the binding site of the protein. This theory introduced in 1894 by Fischer [26] described the protein and the ligand as rigid structures and the

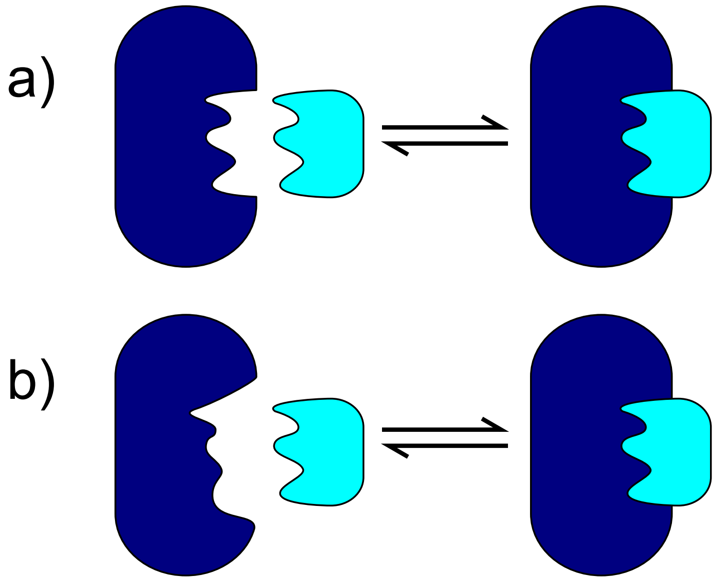


Figure 1.4: Different models for protein-ligand binding. (a) The lock and key model, where the ligand is the complementary geometric shape of the binding site. (b) The induced fit model, where the shape of the binding site adapts upon binding to better fit the ligand.

ligand slotted into the protein like a key in a lock. This model was able to explain the specificity of protein-ligand binding, but was unable to explain the stabilisation that occurs upon binding.

However, proteins are flexible and in 1958 Koshland introduced the induced fit model to accommodate this [27]. This model accepts that the binding of a ligand to a protein is often coupled with a conformational change of the protein that induces tighter binding of the ligand. In proteins with multiple subunits, the conformational change upon binding of one subunit can then affect the conformation of the other subunits. Figure 1.4 shows a graphical representation of these models.

As with any reversible chemical reaction, it is possible to describe the binding of a ligand (L) to a protein (P) with a simple equilibrium expression:

$$\mathrm{P} + \mathrm{L} \underset{k_\mathrm{d}}{\overset{k_\mathrm{a}}{\rightleftharpoons}} \mathrm{PL}, \tag{1.3.1}$$

where the binding event has an association constant, K_a. K_a can be calculated either from the ratio of product and reactant concentrations,

$$K_\mathrm{a} = \frac{[\mathrm{PL}]}{[\mathrm{P}][\mathrm{L}]}, \tag{1.3.2}$$

or the ratio of the rates of binding (k_a) and unbinding (k_d),

$$K_\mathrm{a} = \frac{k_\mathrm{a}}{k_\mathrm{d}}. \tag{1.3.3}$$

Here K_a is a measure of the affinity of the ligand for the protein, with units of mol^{-1} dm^3 (or concentration^{-1}). A higher value of K_a corresponds to a higher binding affinity between ligand and protein.

It is also sometimes useful to consider the binding equilibrium from the standpoint of the fraction of ligand binding sites on the protein that are occupied:

$$\theta = \frac{\text{binding sites occupied}}{\text{total number of binding sites}} = \frac{[\mathrm{PL}]}{[\mathrm{PL}] + [\mathrm{P}]} \tag{1.3.4}$$

By rearranging equation 1.3.2 it is possible to substitute $K_\mathrm{a}[\mathrm{L}][\mathrm{P}]$ for [PL], yielding:

$$\theta = \frac{K_\mathrm{a}[\mathrm{L}][\mathrm{P}]}{K_\mathrm{a}[\mathrm{L}][\mathrm{P}] + [\mathrm{P}]} = \frac{K_\mathrm{a}[\mathrm{L}]}{K_\mathrm{a}[\mathrm{L}] + 1} = \frac{[\mathrm{L}]}{[\mathrm{L}] + 1/K_\mathrm{a}}. \tag{1.3.5}$$

The value of K_a can then be determined by plotting θ against the concentration of free ligand, [L].

Often in biological circles the dissociation constant, K_d, is used to characterise protein-ligand binding. The dissociation constant is the reciprocal of the association constant, $K_\mathrm{d} = 1/K_\mathrm{a}$ and given in concentration units. When looking at the dissociation constant, equation 1.3.5 becomes:

$$\theta = \frac{[\mathrm{L}]}{[\mathrm{L}] + K_\mathrm{d}}. \tag{1.3.6}$$

The dissociation constant, K_d can be described as the molar concentration of ligand needed for half of the available binding sites to be occupied (see figure 1.5).

From the association constant it is possible to calculate the molar Gibbs free energy of binding,

$$\Delta G^{\ominus} = -RT \ln K_\mathrm{a} c^{\ominus}, \tag{1.3.7}$$

where R is the universal gas constant, T is the temperature (usually 298 K or 300 K) and the standard reference concentration $c^{\ominus} = 1$ mol dm^{-3}.

As with any thermodynamic phenomenon, the Gibbs Free energy, ΔG, determines

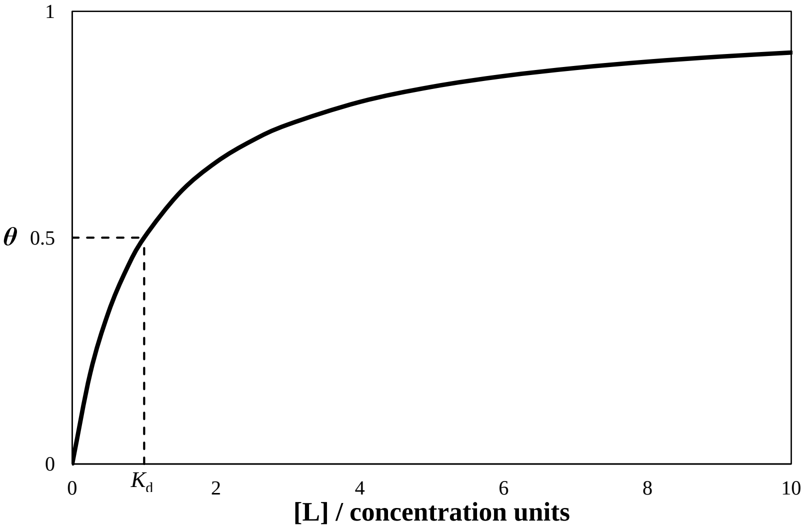


Figure 1.5: A Graphical representation of ligand binding, showing the fraction of occupied ligand binding sites, θ, plotted against free ligand concentration, [L]. The dissociation constant K_d (or $1/K_\mathrm{a}$) is equivalent to the value of [L] when half of the available ligand binding sites are occupied ($\theta = 0.5$)

how well the ligand binds to the protein. $\Delta G < 0$ indicates a favourable binding event and a $\Delta G > 0$ indicates an unfavourable binding event. Where proteins are involved the magnitude of ΔG is usually relatively small, as the ability for a protein to unbind a ligand once it has performed its specific purpose is necessary for the protein to continue to function. The Gibbs Free energy is composed of an enthalpic and an entropic part

$$\Delta G = \Delta H - T\Delta S. \tag{1.3.8}$$

The enthalpic contribution to the free energy, ΔH, is typically associated with changes to structure upon binding that are described in the induced fit model. An enthalpically favourable reaction, where $\Delta H < 0$, results from a net increase in favourable interactions (such as hydrogen bonds or van der Waals interactions) due to the structural change. Changes in the protein dynamics tends to have a larger effect on the entropy, with an entropically favourable binding event being one where $-T\Delta S < 0$. The entropy of a system is a measure of how disordered it is, measuring the number of ways the particles in the system can be arranged. If N is the total number of microstates available to a system and the ratio $P_i = n_i/N$ is the proportion of members of the ensemble occupying a certain microstate, the entropy of the system is given as:

$$S = -k_{\mathrm{B}} \sum_i \frac{n_i}{N} \ln \frac{n_i}{N} = -k_{\mathrm{B}} \sum_i P_i \ln P_i. \tag{1.3.9}$$

Here the index i is summed over all accessible microstates. If the total number of microstates that can be occupied by a system increases, the entropy increases. A ligand binding to a protein usually decreases the flexibility of the protein, decreasing the number of accessible microstates available to the system, which leads to a reduction in the entropy of the system. The less common case of ligand-protein binding increasing the flexibility of the protein causes a decrease in entropy using the same reasoning.

Enthalpy-entropy compensation can be observed during these binding events. Upon binding, the increase in attractive interactions contributing to an enthalpy increase, is compensated by a loss in entropy due to a reduction in the degrees of freedom of the system. There are a number of methods available, such as isothermal titration calorimetry (ITC) (see section 3.2), which can be used to determine the thermodynamic parameters of protein-ligand binding.

1.4 Allostery

Allostery is a process where binding at one site of a protein affects the binding of another ligand at a different site on the same protein. The cooperative binding in these allosteric events can either promote or inhibit the binding of other ligands at distant sites [28]. Being able to understand this process is, therefore, very important for many areas of molecular biology. For example, an allosteric site of a protein can be used as a target for drug design [29–31].

Homotropic allostery is when the second ligand binding to the protein is the same as the allosteric modulator ligand that binds. When the allosteric modulator is different to the second ligand that binds, the interaction is called heterotropic allostery. Proteins may have more than one modulator, so it is possible that they undergo both homotropic and heterotropic allostery. The system studied in this report, CAP, is one such protein. CAP undergoes homotropic allostery with the ligand cyclic adenosine monophosphate (cAMP). The binding of the second cAMP modulator then activates the protein, increasing its affinity for binding to DNA. The homotropic allostery in CAP will be discussed in more detail in section 1.5.5 and will be investigated throughout the thesis.

1.4.1 Cooperative Binding

Allostery is thermodynamic in nature, the cooperativity driven by the free energy difference between the individual binding events. If binding of one ligand makes the binding of the second more energetically favourable, the allosteric effect is said to have positive cooperativity. Negative cooperativity occurs when binding of the first ligand causes the binding of the second to be less favourable.

In 1910 Archibald Hill developed a general approach to study cooperative binding to multi-subunit proteins [32]. By extending the equilibrium of equation 1.3.1, it is possible to describe cooperative binding as:

$$\mathrm{P} + n\mathrm{L} \underset{k_\mathrm{d}}{\overset{k_\mathrm{a}}{\rightleftharpoons}} \mathrm{PL_n}. \tag{1.4.10}$$

The overall association constant in this case then becomes:

$$K_a = \frac{[\mathrm{PL}_n]}{[\mathrm{P}][\mathrm{L}]^n} \tag{1.4.11}$$

and the fraction of occupied binding sites, θ becomes:

$$\theta = \frac{[\mathrm{L}]^n}{[\mathrm{L}]^n + 1/K_a} = \frac{[\mathrm{L}]^n}{[\mathrm{L}]^n + K_d}. \tag{1.4.12}$$

Figure 1.6 shows how the graphical representation of ligand binding as seen in figure 1.5 changes with cooperative binding. Rearranging equation 1.4.12 gives the Hill equation:

$$\frac{\theta}{1-\theta} = \frac{[\mathrm{L}]^n}{K_d}, \tag{1.4.13}$$

which can also be given in its logarithmic form:

$$\log\left(\frac{\theta}{1-\theta}\right) = n\log[\mathrm{L}] - \log K_d. \tag{1.4.14}$$

A Hill plot is then created by plotting $\log(\theta/(1-\theta))$ against $\log[\mathrm{L}]$. Theoretically, this plot would have a gradient of n if all of the binding sites were filled simultaneously; this

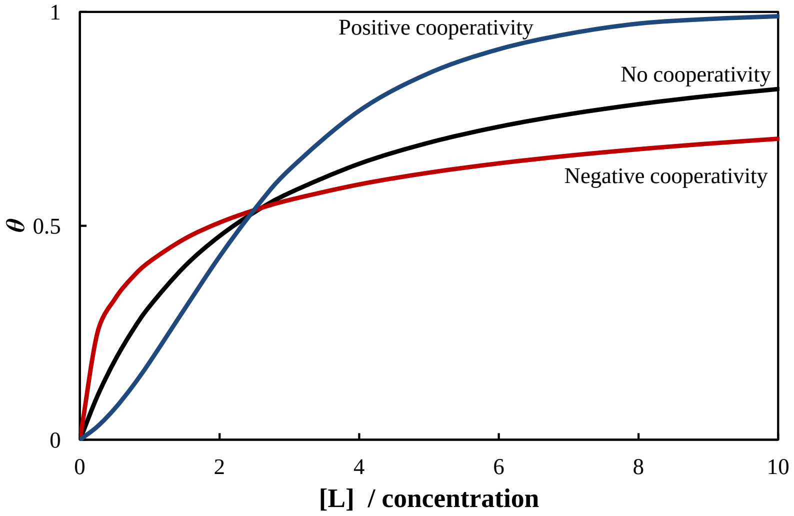


Figure 1.6: Binding curves describing noncooperative binding, positively cooperative binding and negatively cooperative binding. Noncooperative binding occurs when the affinity for binding does not change once an allosteric effector binds. Positively cooperative binding has a sigmoidal shape curve, whereas negatively cooperative binding has a curve similar to a protein with two classes of binding sites. Adapted from [33].

corresponds to perfectly cooperative binding. However, in practice the experimentally determined gradient reflects the level of cooperativity between the binding sites rather than the total number of binding sites (n). Thus, the gradient of the Hill plot, called the Hill coefficient and denoted n_H, reflects the degree of cooperativity. Values of n_H greater than one indicates positive cooperativity in the ligand binding, whereas values of n_H less than one implies that the ligand binding is negatively cooperative. If the value of n_H is equal to one, then the ligand binding is not cooperative. Figure 1.7 shows example Hill plot's for the binding of oxygen to myoglobin and haemoglobin.

When reducing the complexity of cooperative binding of n ligands to homotropic allostery with two consecutive binding events of equivalent ligands, L_1 and L_2, the equilibrium becomes:

$$\mathrm{P} + \mathrm{L}_1 + \mathrm{L}_2 \underset{k_{\mathrm{d}^1}}{\overset{k_{\mathrm{a}^1}}{\rightleftharpoons}} \mathrm{PL}_1 + \mathrm{L}_2 \underset{k_{\mathrm{d}^2}}{\overset{k_{\mathrm{a}^2}}{\rightleftharpoons}} \mathrm{PL}_1\mathrm{L}_2. \tag{1.4.15}$$

Here the rate constants for binding the first ligand (L_1) and the second ligand (L_2) can be differentiated by their indices, 1 and 2 respectively. The association constants for

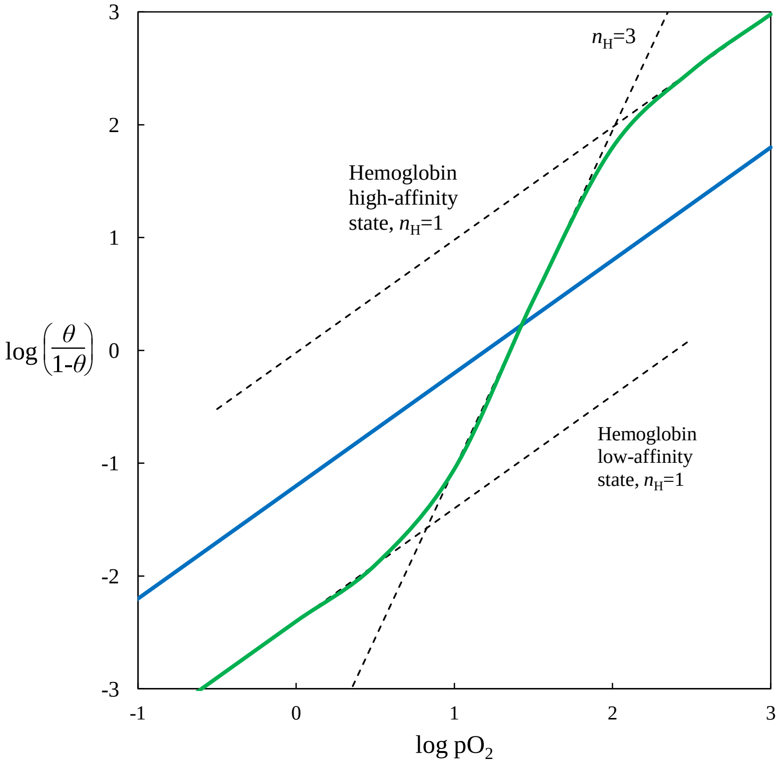


Figure 1.7: Hill plots for the binding of oxygen to myoglobin (blue line) and haemoglobin (green line). Myoglobin exhibits no cooperativity with $n_{\mathrm{H}} = 1$ due to it having a single binding site for oxygen. The maximum value of n_{H} for hemoglobin is $n_{\mathrm{H}} = 3$ indicating strong positive cooperativity, however with haemoglobin's four oxygen binding sites, it is not perfect cooperativity. In this plot, the partial pressure of oxygen, pO$_2$, is used rather than the free ligand concentration [L].

each binding event are then given as:

$$\begin{aligned} K_a^1 &= \frac{[\mathrm{PL}_1] + [\mathrm{L}_2]}{[\mathrm{P}] + [\mathrm{L}_1] + [\mathrm{L}_2]} \\ K_a^2 &= \frac{[\mathrm{PL}_1\mathrm{L}_2]}{[\mathrm{PL}_1] + [\mathrm{L}_2]} \end{aligned} \tag{1.4.16}$$

As with a single ligand binding event, the free energy change upon binding determines whether the binding event is energetically favourable or not. In homotropic allostery, the difference between the free energy of binding for two consecutive binding events, $\Delta\Delta G$, determines the cooperativity of the allostery:

$$\Delta\Delta G = \Delta G_2 - \Delta G_1. \tag{1.4.17}$$

Here ΔG_1 is the free energy for the first binding event and ΔG_2 is the free energy for the second binding event, so each of these can in turn be written as the difference between the free energy of the systems before and after each binding event:

$$\Delta\Delta G = (G_{\mathrm{P\cdot 2L}} - (G_{\mathrm{P\cdot L}} + G_{\mathrm{L_2}}) - (G_{\mathrm{P\cdot L}} - (G_{\mathrm{P}} + G_{\mathrm{L_1}}). \tag{1.4.18}$$

Here $\mathrm{P \cdot 2L}$ refers to the protein with the two equivalent ligands bound, $\mathrm{P \cdot L}$ refers to the protein with just one ligand bound and P refers to the free protein. These are frequently denoted holo$_2$, holo$_1$ and apo respectively and are referred to as such throughout this thesis. In homotropic allostery, the two ligands, $\mathrm{L_1}$ and $\mathrm{L_2}$ are the same, meaning that equation 1.4.18 can be simplified to:

$$\Delta\Delta G = G_{\mathrm{P\cdot 2L}} - 2G_{\mathrm{P\cdot L}} + G_{\mathrm{P}}. \tag{1.4.19}$$

A slightly different notation is adopted for heterotropic allostery; a protein that is not bound to the allosteric effector is called the apo-protein whereas a protein with the allosteric effector bound is called the holo-protein. The allosteric free energy can then be determined by finding the difference between the free energy of binding the substrate to the apo-protein and the holo-protein;

$$\Delta\Delta G = \Delta G_{\mathrm{holo}} - \Delta G_{\mathrm{apo}}. \tag{1.4.20}$$

For both homotropic and heterotropic allostery, if $\Delta\Delta G$ is negative, the two binding events are positively cooperative, meaning that the second binding event is more favourable than the first binding event. A $\Delta\Delta G$ greater than zero indicates negative cooperativity.

Another way to measure the cooperativity of binding in homotropic allostery is the cooperativity constant, given by the ratio of association constants between the consecutive binding events,

$$C = \frac{K_a^2}{K_a^1}, \tag{1.4.21}$$

where a C value greater than 1 indicates positive cooperativity and a C value less than 1 indicates negative cooperativity. $\frac{K_a^2}{K_a^1}$ can be related to the allosteric free energy difference by:

$$\Delta\Delta G = -RT \ln \frac{K_a^2}{K_a^1}. \tag{1.4.22}$$

Early models of allosteric signalling [34,35], explained any allosteric effects in terms of conformational changes of the protein. These models are described in greater detail in sections 1.4.2. However, the free energy of binding (ΔG) has an entropic contribution (ΔS) as well as the enthalpic contribution (ΔH) which should really be considered due to the thermodynamic nature of allostery; $\Delta G = \Delta H - T\Delta S$. Therefore, more recently it has been accepted that a change in the dynamics of a protein (primarily entropic) can contribute to or even drive allostery in some cases (see section 1.4.3). It has been observed experimentally that a change in the average backbone structure of a protein is not always necessary for allostery to occur, instead it can be solely explained by a change in the dynamics of the protein. It was Cooper and Dryden [36] who originally proposed a model for allosteric signalling driven by the effect of dynamics. They derived approximate expressions, allowing the magnitude of the allosteric effect to be measured by changes in normal mode frequencies and mean-square atomic displacements.

1.4.2 Models for Cooperative Binding

Many models have been proposed describing cooperative binding in proteins. There are, however, two very different models, inspired by the binding of oxygen to haemoglobin, that are widely accepted by the field.

MWC Model

The first of these two models, usually called either the MWC, or concerted model, was developed by Monod, Wyman and Changeaux in 1965 [34]. The MWC model assumes that the allosteric protein comprises of at least two identical subunits in identical conformations; each with one binding site. Each subunit exists in one of two states, however the symmetry of the protein is conserved, so although large structural changes may occur during the transition between states, all of the subunits must undergo the changes simultaneously. In one state, the T_i state, the binding site has a low affinity for ligand and in the other state, the R_i state, the binding site has a higher affinity for the ligand. The index i denotes the number of ligands bound to the protein in each state, however the protein's affinity for the ligand stays constant regardless of the value of i. This model assumes that there is an equilibrium between the two forms such that at any moment there exists a mixture of proteins with all their subunits in one of the two states (see figure 1.8). In the the absence of ligand the equilibrium between the

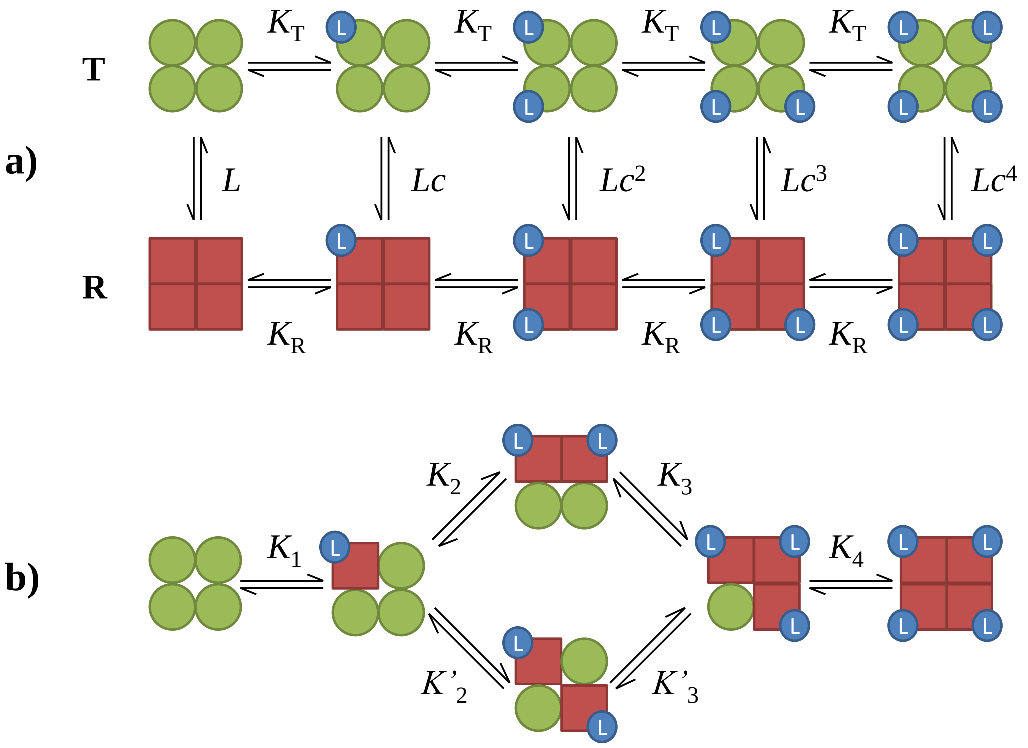


Figure 1.8: The MWC and KNF models for allostery. Adapted from [4, 33, 35]

two states is characterised by the equilibrium constant, $L = \frac{[T_0]}{[R_0]}$. In this model $L > 1$, meaning that most molecules are in the T_0 state, and the protein has a low affinity for ligand.

The model describes cooperativity by increasing the proportion of molecules in the R_i state as ligands bind to the subunits of the protein; a shift in the equilibrium to favour the R form more in accordance with Le Châtelier's principle. The association constants for the binding of one ligand to the T_i and R_i states are designated K_T and K_R respectively, staying constant regardless of how many ligands are bound to the protein. A cooperativity constant, c, is then given by the ratio of these association constants;

$$c = \frac{K_T}{K_R}, \tag{1.4.23}$$

with $0 < c < 1$ and smaller values of c corresponding to stronger positive cooperativity. When one ligand binds, the equilibrium constant between the T_1 and R_1 states becomes Lc. This extends to the case of i ligands bound, making the ratio between T_i and R_i states:

$$\frac{[T_i]}{[R_i]} = Lc^i. \tag{1.4.24}$$

The simplicity of this model with only two parameters L and c, means that it is best for describing homotropic positive allostery. It is unable to correctly model negative allostery, however the original authors suggested that heterotropic allostery could be modelled with the same mechanism, with the allosteric effector stabilising either the T or R states for positive or negative allostery respectively.

At the time of writing the paper was a success; the failure of the model to describe negatively cooperative homotropic allostery was not important, as no such systems had been observed.

KNF Model

The second model was proposed by Koshland, Nemethy and Filmer in 1966 [35]; called the KNF model or the sequential model. This model was devised for allosteric proteins composed of two or more subunits.

In the sequential model, each of the subunits are able to exist in one of two conformations, W a weak binding form, or S a strong binding form. The binding of the ligand to a subunit induces a structural change in that subunit to the S form. Cooperativity in this model arises because the pair interactions between adjacent subunit pairs change depending on the composition of the pairs. Such that the interactions between W-W,W-S and S-S pairs are assumed to be different. In this model, the association constant for binding one ligand, L, to an individual subunit is defined as:

$$K_{\mathrm{a}} = \frac{[\mathrm{SL}]}{[\mathrm{S}][\mathrm{L}]}. \tag{1.4.25}$$

The transformation constant for a subunit from W to S when the neighbouring subunits are the S form is also

$$K_{\mathrm{t}} = \frac{[\mathrm{S}]}{[\mathrm{W}]}. \tag{1.4.26}$$

The interaction between the subunits are represented by the constants K_{WS} and K_{SS} for W-S and S-S subunit pairs respectively. From these constants and the topology of the interactions between the subunits, the fraction of occupied binding sites, θ can be determined. The fact this model allows different topologies of interactions to be modelled is an advantage over the MWC model, for example besides a square topology seen in figure 1.8 where each subunit has two neighbours, tetrahedral topologies are allowed where subunits each have three neighbours.

The KNF model, unlike the MWC model creates the possibility of negative cooperativity by destabilising latter binding events as well as positive allostery by stabilising the latter binding events. However, the MWC model is mathematically much simpler with only two parameters that need to be fitted in comparison to the four parameters along with the interaction model that need to be fitted for the KNF model. Therefore, due to its mathematical complexity, the KNF model has not experienced the same level of popularity as the MWC model.

Ackers Model

The cooperative binding of O_2 to haemoglobin, has been a popular system for modelling allostery. Haemoglobin is a tetramer, containing two copies of α and β subunit pairs, which themselves are very similar. Haemoglobin is therefore often modelled as a dimer of α-β pairs. For this example, the MWC model was shown to be insufficient when X-ray diffraction [37–39] and NMR studies [40,41] showed the conformation of subunits changing independently upon binding, breaking the symmetry rule of the model. In 1992 Ackers *et al.* summarised that the cooperative binding of O_2 to haemoglobin has similarities to both the MWC model and the KNF model, but agrees completely with neither. A number of other models for allostery have been proposed to try to model this effect better than the MWC and KNF models; often drawing from themes in the MWC and KNF models [42–46].

A popular model for haemoglobin was introduced by Ackers *et al.* in 1992 [47] and is displayed in figure 1.9. In this model, it is assumed that the haemoglobin tetramer is comprised of two subunits, each with an α and a β monomer. Each monomer (regardless of whether it is an α or a β monomer) is in the t state when no ligand is bound and undergoes a structural change in the tertiary structure to r upon binding the ligand. The overall quaternary structure starts in the T state when no ligands are bound and only changes from the T to the R state once a ligand is bound to at least one of the binding sites on each α-β pair. Hence, this model draws themes from both the MWC model and the KNF model.

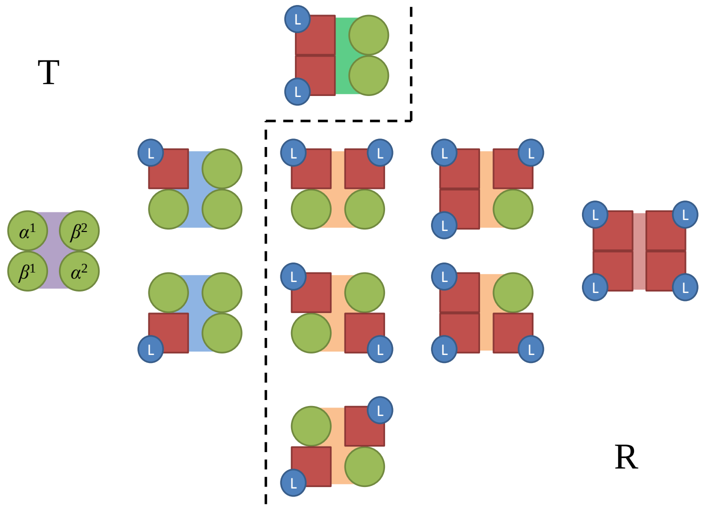


Figure 1.9: A model for allostery developed by Acker *et al.*. In this model, the four subunits of the tetramer are split into $\alpha - \beta$ pairs. Each of these subunits exists in either the t state (depicted as a green circle) or the r state (depicted as a red square). The quaternary structure of the tetramer has two states, the T state or the R state. The tetramer exists in the T state when no ligands are bound and only transitions to the R state when a ligand binds to at least one subunit of each of the $\alpha - \beta$ pairs.

1.4.3 Dynamic Allostery

All of the models of allostery previously described assume that allosteric signalling is solely caused by conformational changes caused by one or more ligands binding to the protein. Thermodynamically, this view is a solely enthalpic view, looking only at the strengths of interactions in the allosteric systems. However, as has previously been discussed, ligand binding also contains an entropic term, which can equally have a role in allosteric binding. Weber first made this point in 1972 [2], stating that proteins constantly fluctuate between a number of conformations and the structure observed is only the average. The importance of these conformational fluctuations in allosteric signalling have been studied with combinations of NMR and X-ray diffraction [48–51]. There is increasing evidence that allosteric cooperativity can be communicated between distant sites on proteins through modulation of their dynamic properties, even in cases with no structural change between the ligand bound (holo) and unbound (apo) forms [36, 52–59].

Cooper and Dryden view of Dynamic Allostery

It was Cooper and Dryden who first identified the potential of allostery without a macromolecular conformational change in 1984 [36], introducing a model to describe this dynamic allostery. They investigated the potential importance of dynamics in allosteric signalling by evaluating the thermodynamics of ligand binding. They hypothesised that changes in the amplitude and frequency of macromolecular thermal fluctuations could generate cooperative binding free energies amounting to several kJ mol^{-1}. They noted that this effect can involve the entire spectrum of dynamic behaviour from highly-correlated low frequency vibrational modes to the local anharmonic motion of individual atoms or residues. Protein motions over both timescales had been demonstrated experimentally [60–63] and theoretically [64–69] shortly before.

This view of allostery is predominantly entropic; a stark change to the previous models which explained allostery entirely by conformational changes, a phenomenon that is predominantly enthalpic. In this new view, Cooper and Dryden suggested that proteins may have evolved to take advantage of these thermal motions as well as their mean conformational states. They started by looking at the thermodynamics of protein binding, noting an almost universal decrease in heat capacity upon ligand binding [70], which suggests a reduction in the thermal energy fluctuations.

They therefore opted to study the canonical partition functions of the molecular species involved, and using them to evaluate the association constant for ligand binding:

$$K_\mathrm{a} = \frac{Z_\mathrm{P} Z_\mathrm{L}}{Z_{\mathrm{PL}_1}} \exp\left(\frac{-\Delta\varepsilon_1}{kT}\right) \tag{1.4.27}$$

where $\Delta\varepsilon_1 = \varepsilon_\mathrm{P} + \varepsilon_\mathrm{L} - \varepsilon_{\mathrm{PL}_1}$ is the difference in zero point energies, corresponding to the energy of ligand binding at 0 K. Z_P, Z_L and Z_{PL_1} are the partition functions for the free protein, the free ligand and the protein-ligand complex respectively. In the absence of significant changes in volume, the Gibbs free energy is then:

$$\begin{aligned} \Delta G_1 &= -kT \ln K_\mathrm{a} \\ &= \Delta\varepsilon_1 - kT \ln\left(\frac{Z_\mathrm{P} Z_\mathrm{L}}{Z_{\mathrm{PL}_1}}\right) \end{aligned} \tag{1.4.28}$$

The same argument can be applied to subsequent binding events to find ΔG_2 and so on. The allosteric free energy difference is then returned as:

$$\begin{aligned} \Delta\Delta G &= \Delta G_2 - \Delta G_1 \\ &= \Delta\varepsilon_2 - \Delta\varepsilon_1 - kT \ln\left(\frac{Z_{\mathrm{PL}_1}^2}{Z_\mathrm{P} Z_{\mathrm{PL}_2}}\right) \end{aligned} \tag{1.4.29}$$

For homotropic allostery in the absence of structural change, $\Delta\varepsilon_1 = \Delta\varepsilon_2$ as the same molecular contacts are involved at each site.

The partition function is separable into translational, rotational, vibrational, electronic and conformational states, so Cooper and Dryden investigated each contribution. The electronic, rotational and translational contributions were disregarded straight away. The electronic because at ambient temperatures there are no significant electronic excitations and the rotational and translational disregarded because the size of the protein compared to the ligand means that the terms are very similar for different ligated states, cancelling each other out. This leaves just the contribution due to changes in the vibrational spectrum of the system and conformational changes in the system.

The classical limit for the vibrational partition function of a vibration with frequency ν is:

$$q(\nu)_{\text{class}} = \frac{kT}{h\nu}, \tag{1.4.30}$$

and the free energy difference associated with a vibrational frequency shift $\nu_{\text{P}} \rightarrow \nu_{\text{PL}_1} \rightarrow \nu_{\text{PL}_2}$ during sequential binding is

$$\Delta\Delta G_{\text{vib}} = -kT\left(\ln q(\nu_{\text{P}}) + \ln q(\nu_{\text{PL}_2}) - 2\ln q(\nu_{\text{PL}_1})\right). \tag{1.4.31}$$

If the only modes affected by the sequential binding are high frequency modes then the quantum partition function is essentially unity, as $h\nu >> kT$, giving non cooperative binding ($\Delta\Delta G_{\text{vib}} = 0$). They then showed that for a low frequency global mode of motion the allosteric free energy difference becomes

$$\Delta\Delta G_{\text{vib}} = -kT\ln\left(\frac{\nu^2_{\text{PL}_1}}{\nu_{\text{P}}\nu_{\text{PL}_2}}\right) \tag{1.4.32}$$

within the classical limit. They showed that if a stiffening in the protein was induced each time a ligand bound, $\Delta\Delta G$ would be negative, indicating positive cooperativity. For example if each ligand binding event increases the normal mode frequencies by 10%, a $\Delta\Delta G$ of around -0.01 kT would be observed per mode. Summing over the hundreds of low-frequency normal modes proteins have, this effect can give rise to biologically relevant free energy changes around the order of a few kJ mol^{-1}.

As well as the small shifts in the frequencies of normal modes that could be caused by ligand binding, there is also the possibility that binding could freeze out entire global modes involving collective domain motions or even convert them to higher frequency modes. This effect was hypothesised for lysozyme [71] and later demonstrated experimentally [62, 72].

They then went on to study the hypothesis that the protein wanders in a haphazard and none periodic fashion amongst a number of conformational states, as supported by MD [68, 69]. The probability distribution of these conformational states can be used to determine the flexibility of a protein, with wider probability distributions corresponding to more flexible proteins. Ligand binding can then either cause a shift in the mean of this probability distribution, corresponding to a conformational change in the usual sense, or it can cause a change in the overall shape of the distribution. In this case, if a narrower distribution appears due to ligand binding, this represents a stiffening of the protein without a global structural change.

They hypothesised that if these distributions of fluctuations in coordinates for each atom were approximated as being Gaussian in shape with a width of σ. The atomic partition function for a Gaussian distribution about a fixed mean position is proportional to the width of the distribution, σ. Therefore for homotropic allostery with two equivalent binding sites, the contribution from atom i to the allosteric free energy is:

$$\Delta\Delta G = -kT\ln\left(\frac{\sigma_1(i)\sigma_1(i')}{\sigma_0(i)\sigma_2(i)}\right) \tag{1.4.33}$$

assuming i and i', equivalent sites from the two chains have no difference for the apo and holo$_2$ forms. The $\sigma_n(i)$ represent the root-mean-square fluctuation (RMSF) at i when n ligand molecules are bound to the protein. In this case, a shift in the RMSF of the order of 1% per atom, which would be difficult to observe experimentally could give rise to cooperativity free energy differences in the order of a few kT, which again is an entirely entropic effect.

The article was unable to make an immediate impact on the views of allostery, partly due to the lack of experimental techniques available to observe the changes in protein structural fluctuations upon ligand binding. It wasn't until 1993 that thermodynamic fluctuations were shown experimentally by NMR to contribute to the cooperative free energy of ligand binding [73]. Since then more examples of dynamic allostery, unable to be explained by structural changes in the classical sense have emerged, including:

allostery without a structural pathway [74–76], changes in cooperativity due to surface mutations that do not affect structure [77], and allostery without a conformational change [17]. This original article and the subsequent studies of dynamic allostery has sparked debate over the mechanism for which these dynamic fluctuations are communicated between distant allosteric sites.

Mechanisms for signalling between allosteric sites

One hypothesis, by Hilser, describes a protein as an ensemble of conformational states [78–80]. The statistical weights are computed for each of the conformational states in the ensemble by Boltzmann weighting the free energy of the conformation. This hypothesis investigates how a perturbation, such as ligand binding at distant sites can affect the cooperative network within the protein. Ligand binding is viewed to stabilise the state which has the highest affinity for the ligand, causing changes in the dynamics, conformational entropy or the population's mean conformation. This ultimate redistribution of the ensemble affects binding of another ligand, which could be identical to the first ligand for homotropic allostery or different for heterotropic allostery.

A second perspective is that physically connected pathways of increased (or decreased) thermal motions, coupled along their trajectories, connect the allosteric sites [81–83]. In this view there are a network of spatially distributed key residues which have cooperative interactions as a pathway for allostery or other signalling events. The exact nature of these cooperative interactions differ from case to case, but a number of examples with this kind of signalling have been observed, for example in G protein-coupled receptors [84,85], in the modulation of ligand binding in PDZ domains [86] and in controlling the catalytic rate for a number of enzymes [21,87–89]. Protein residues which are functionally important, such as those in an allosteric pathway, have been observed to have co-evolved throughout the long term evolution of a protein; whereas the majority of amino acids evolve almost independently [90–95], strengthening the argument for allosteric pathways.

In this thesis, the hypothesis that the global collective modes are important carriers of the allosteric signal and can act without a change in conformation is investigated.

1.5 Catabolite Activator Protein.

The catabolite activator protein (CAP), often referred to as the cAMP receptor protein (CRP), is an extensively studied transcription factor native to certain strains of *E. coli*. CAP that has been activated by the allosteric effector cAMP is a transcription activator for genes involved in the catabolism of lactose. In its apo form (without the ligand cAMP bound) it is inactive, but when two molecules of cAMP bind to the dimer (cAMP$_2$-CAP), a change in the protein's structure renders it active. CAP also has an intermediate form, where one cAMP is bound to one of the subunits, while the other subunit remains un-liganded (cAMP$_1$-CAP).

The allosteric regulation of CAP by the effector, cAMP is an example where the cooperativity cannot be described solely by a conformational change. It was recently shown that a truncated form of CAP exhibited negative cooperativity without a substantial change in conformation [17].

1.5.1 CAP's role in Catabolite Repression

When bacteria have more than one food source, for example glucose and lactose, it is more efficient for the bacteria to use the food source that is most quickly metabolisable to get its energy and carbon. In this example, the bacteria would preferably metabolise glucose than the lactose. This phenomenon (initially called the glucose effect) was first described by Monod when he observed a two stage growth of *E. coli* with a mixture of glucose and lactose as the food source [96]. The bacteria is able to do this by suppressing the genes necessary for the catabolysis of lactose. This process has more recently been named catabolite repression.

When the concentration of glucose in the cell is high, the concentration of cAMP in the cell decreases. This is because the presence of glucose in the cell inhibits the production of cAMP by specific mechanisms [97,98]. Firstly, glucose inhibits the intake of inducer molecules, which are necessary for the induction of repressed genes. One such inducer that is believed to be excluded from the cell by glucose is a dephosphorylated component (Factor III) of the phosphoenol pyruvate-dependent phototransferase system (PTS). Factor III of PTS is an activator of adenylate cyclase, which in turn is a catalyst for the synthesis of cAMP from adenosine triphosphate (ATP), thereby the overall concentration of cAMP in the cell is reduced [99, 100].

This means that when glucose is available to the cell, the concentration of cAMP is low and when it is not, the concentration of cAMP dramatically increases. When cAMP then binds to CAP it activates the protein and increases its affinity for binding to DNA to a very large extent.

CAP acts as a transcriptional activator, it is known to bind to a promoter of the *lac* operon gene as well up to 200 other promoters [101, 102]. The *lac* operon contains genes for three proteins (β-galactosidase, β-galactoside permease and β-galactoside transacetylase) necessary for the catabolism and transport of lactose. When CAP binds to the DNA it is known to bend the chain by approximately 90° [103]. It forms a complex with the ribonucleic acid (RNA) polymerase, thereby allowing it to bind to the DNA and initiate transcription of the genes in the *lac* operon.

1.5.2 Structure of CAP

CAP, like many prokaryotic transcription factors, exists as a homodimer with two-fold symmetry; each subunit consisting of 210 amino acids [104]. Each subunit in the CAP dimer consists of two domains, a C-terminal DNA binding domain (DNABD) (residues 138 to 210) and an N-terminal ligand binding domain (LBD) (residues 1 to 137) to which the allosteric effector, cAMP, binds. A hinge region (residues 134-138) connects the first α-helix (D-helix) of the DNABD to the last α-helix (C-Helix) of the N-terminal domain [105, 106]. The two-domain structure can easily be seen by looking at the 1.48 Å resolution structure of the CAP-cAMP complex, obtained by X-ray crystallography (see figure 1.10) [59, 107].

The dimerisation of CAP can be accounted for by the LBD, which is predominantly composed of β-sheets with an antiparallel β-roll structure from residues 19 to 99 and a large α-helix on the dimer interface. The DNA binding domain consists of three α-helices; two of which form a typical helix-turn-helix motif. The F-helix in this motif binds to specific sequences of DNA when cAMP is present (see figure 1.10). The orientation of this domain is believed to be affected by the binding of cAMP to the LBD of the protein [18, 108].

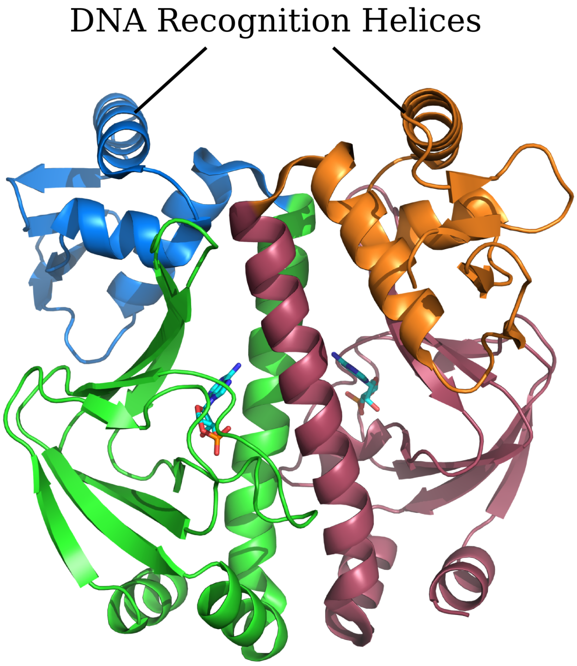


Figure 1.10: 1.48 Å resolution structure of CAP (PDB code 4HZF) [59], showing α-helices as coils and β-sheets as arrows. One monomer is coloured green and blue while the other is coloured red and orange. The DNABDs (coloured blue and orange) and LBDs (coloured green and red) are clearly two separate subunits. The DNA recognition (F) helix is labelled. The cAMP ligands (with a stick representation) can be seen nested in the LBD.

Structure of cAMP$_2$-CAP

Most structural information about CAP is for the cAMP$_2$-CAP bound state. When in this bound state, the hinge region separating the LBD and the DNABD sits between residues 134 and 138 and the DNABD on both monomers are oriented so that the DNA recognition F-helices point in the correct direction to bind to DNA [109]. The DNABDs of each subunit were observed in different orientations relative to their respective LBD in the original crystal structure by Steitz and Weber [110]. This demonstrates the inherent flexibility of the hinge region and the ability of the DNABD to rotate around this hinge to a certain extent.

Two cAMP molecules bind CAP in the anti-conformation at two equivalent sites in the LBD, one for each monomer. They sit between the β-roll structure and the coiled-coil at the dimer interface. The cAMP molecule forms a hydrogen bonding network within the binding site, as shown in figure 1.11. Hydrogen bonds form between the ribose and phosphate oxygens of cAMP and residues Gly72, Glu73 and Ser84. An additional hydrogen bond, or potentially an ionic interaction occurs between one of the phosphate oxygens of cAMP and Arg83. At the other end of cAMP the adenine

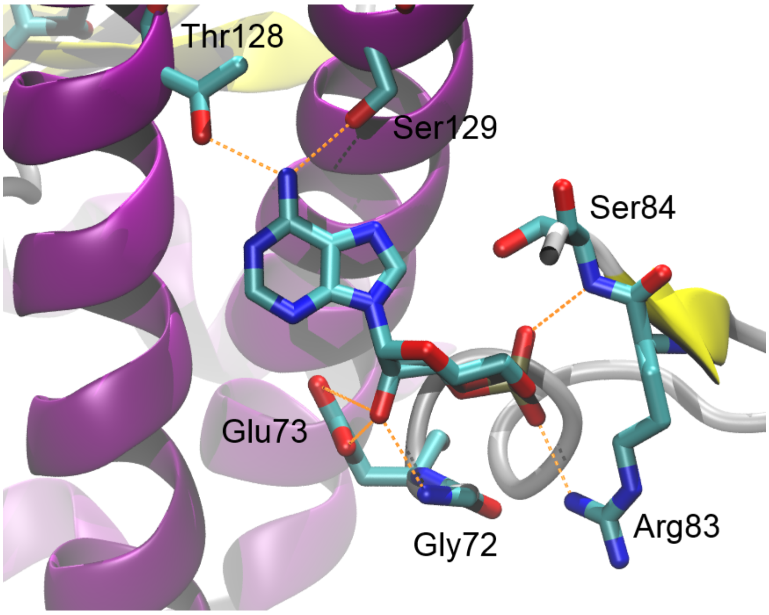


Figure 1.11: The binding site of cAMP in CAP (PDB code 4HZF) with the hydrogen bonds and ionic interactions between cAMP and CAP shown.

6-amino group forms one hydrogen bond with the OH of residue Thr128 in the large C-helix of the same chain and another with the OH of Ser129 of the C-helix in the opposite chain [110].

Structure of apo-CAP

A number of studies of apo-CAP have shown that in the absence of cAMP, CAP undergoes a slight shortening of the C-helices in the LBDs and a lengthening of the D-helices of the DNABD by a similar amount. This effect has been shown in X-ray [108] and NMR structures [111] of CAP and earlier by NMR NOESY spectra [112]. The structures of apo-CAP show a rigid rotation and translation of the DNABD when compared to the $cAMP_2$ structure. For example, Kalodimos *et. al* [111] used NMR to predict a rotation of roughly 60° and translation of 7 Å of the DNA binding domain compared to the $cAMP_2$ CAP.

Steitz *et. al* [108], while not giving magnitudes of rotation and translation of the DNA binding domain in apo-CAP relative to $cAMP_2$-CAP, showed the DNABD in a different orientation for a 2.3 Å resolution structure of a D138L apo-CAP mutant. They also provided a low resolution (3.6 Å) structure of apo-CAP, agreeing with the large reorientation of the DNABD seen in the mutant. They observed little difference between the internal structures of the DNABDs of the mutant and wild-type (WT); the main difference being slightly different orientations of the loops near the cAMP binding site. However, the extent to which the DNABD transformed with respect to the LBD was different between mutant and WT. The WT structure also contained three dimers per asymmetric unit, all of which showed the DNABD in a slightly different orientation with respect to the LBD. Figure 1.12 shows the structural change observed by Steitz et al. [108].

The lack of agreement between the structures in these studies suggest that the hinge region between the LBD and the DNABD is very flexible, allowing the DNABD to rotate and translate to a number of different preferred orientations. The rotation of the DNABD relative to that of cAMP$_2$-CAP in both these instances mean that the DNA-recognition helix are not in the correct position or orientation to fit into the major groove of the DNA. Hence apo-CAP is inactive and does not regulate transcription.

Regardless of the large structural rearrangement of the DNABD, the LBD responsible for the dimerisation of CAP and the cooperative binding of cAMP undergoes very little structural change. The main structural change observed is the shortening of the C-helices, which are stabilised in cAMP$_2$-CAP by the hydrogen bonds between residues THR127 and SER128 in the C-helices and cAMP [17, 18, 111].

Structure of cAMP$_1$-CAP

No crystal structures for cAMP$_1$-CAP are available. However, there is evidence that the structure of the full CAP dimer does not undergo major changes from the apo form upon binding one cAMP molecule, but changes on a larger scale on binding of the second cAMP. Investigations using Raman spectroscopy have shown that the secondary structures of apo-CAP and cAMP$_1$-CAP in solution have no noticeable differences [113, 114]. Whereas cAMP$_2$-CAP shows a 7% decrease in α-helix content and a 5% increase in β-sheet content when compared to the secondary structure of apo-CAP. In addition to this, investigations into the hydrodynamic properties of CAP have shown that the stokes radius of CAP decreases by a very small amount on the binding of one cAMP and by 2% on binding of the second cAMP [115]. The lack of a change in structure of CAP when the first ligand binds suggests that dynamics must hold a very important role in the allosteric regulation of CAP.

A recent NMR study, by Popovych et al. [17], of the N-terminal CAP dimer with

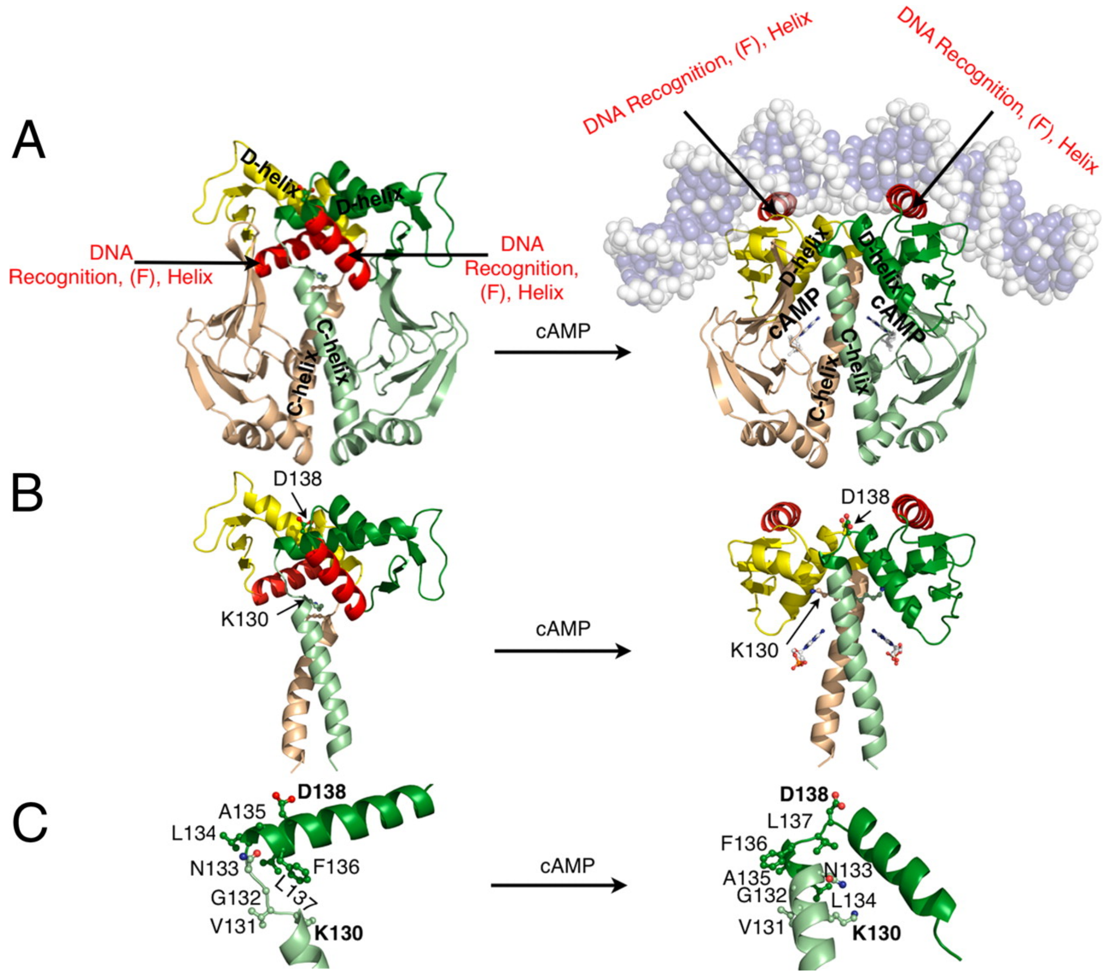


Figure 1.12: The structures of apo-CAP (left) and holo$_2$-CAP (right), demonstrating the structural change of CAP when cAMP binds. (A) The structures of full length CAP. (B) The LBDs have been excluded in these representations to emphasise the differences in orientations of the DNABDs; the DNA recognition helice for apo-CAP is not in the correct orientation to bind specifically to DNA. (C) Representations of just the hinge regions of one monomer. The location of the hinge region is different for apo-CAP to holo$_2$-CAP. Taken from [108].

the DNABDs truncated (CAP^{N}) showed that if a ligand bound to one subunit, then the structure of the subunit with which the cAMP was bound changed, but the structure of the unbound subunit remained the same as apo-CAP^{N}. The binding of the second cAMP ligand restored the symmetry of the CAP^{N} dimer, with both subunits having the same conformation as the liganded subunit in the cAMP$_1$-CAP^{N} complex. They were able to show this result using a 2D-NMR technique; only one set of resonances were present for apo-CAP^{N} and cAMP$_2$-CAP^{N}, whereas two resonances were present for the 2D-NMR spectrum of cAMP$_1$-CAP^{N}, hence the structure was only changed for the liganded subunit.

1.5.3 CAP with Four cAMP Binding Sites

Contrary to the view that two ligands of cAMP bind to CAP, crystal structures of a CAP-DNA complex with four cAMP molecules bound to each CAP have been published [116]. In addition to the two cAMP molecules buried in the N-terminus domains in the *anti*-conformation, they also contain an additional two *syn*-cAMP molecules bound to the surface of the DNA binding domain.

Passner and Steitz [116] postulated that the three states of CAP in its homotropic allostery were not apo-CAP, cAMP_1-CAP and cAMP_2-CAP. Instead they argued that the intermediate structure could actually be cAMP_2-CAP with the two cAMP molecules bound in the buried sites of the N-terminus. They also argued that the fully occupied CAP was cAMP_4-CAP with the two *anti*-cAMP bound in the N-terminus and the two *syn*-cAMP bound to the surface of the DNA binding domain. There is however no compelling experimental evidence, except for this crystal structure, that the activated form of CAP has four cAMP molecules bound. Most experimental evidence in fact points to the stoichiometry of the cAMP-CAP complex being 1:1 for the activated complex including structures of the CAP-DNA complex [103, 117, 118].

1.5.4 CAP binding to DNA

Crystal structures of CAP bound to DNA show that the F-helices in the CAP DNABDs fit into the major grooves of the DNA binding domain. This binding constrains the DNA, causing it to bend by approximately 90° [103, 117] allowing RNA polymerase to bind to the DNA and initiate transcription. A number of X-ray structures of the complex of CAP with DNA [103, 117, 118] and CAP with DNA and the C-terminal domain of RNA polymerase [119] show no significant conformational change in CAP from the cAMP_2-CAP form when compared to the conformational change seen between apo and holo_2-CAP.

1.5.5 Dynamic Allostery in CAP

Unlike most observed allosteric systems, the homotropic cooperative binding of cAMP to CAP is an example of negative cooperativity [17, 25]. This makes CAP an interesting protein to study because in contrast to positively cooperative systems, negative

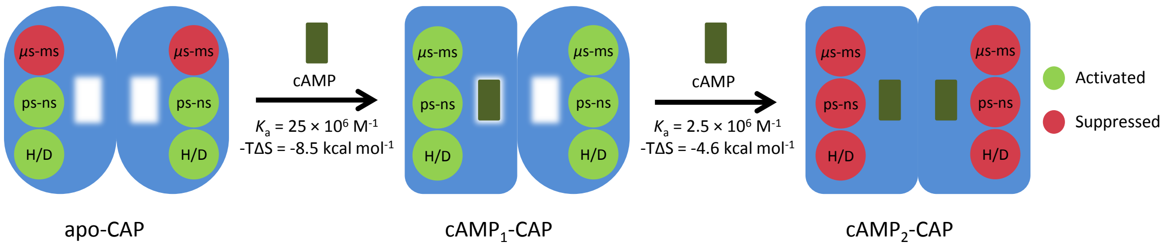


Figure 1.13: A representation of the effect the sequential binding of cAMP to CAP has on the structure and dynamics of CAP^{N}. For apo-CAP^{N}, motions on the μs to ms time scale are suppressed. Binding of the first cAMP does not change the structure of the subunit to which it is not bound, but does activate motions on this time scale. Binding of the second cAMP, suppresses the motions on all the timescales. A change in backbone conformation is indicated by a change in the shape of the monomer. Adapted from [17].

allostery allows intermediate states to be studied in the WT protein. This was shown to be possible in a study of allostery in the N-terminal domain of CAP (CAP^{N}) by Kalodimos *et al.*, which showed a strong negative cooperativity when binding cAMP without a structural change [17]. The measured values of the cAMP association constants to CAP^{N} in this study were $K_{\text{a}}^{1} = 25 \times 10^{6}\ \text{M}^{-1}$ and $K_{\text{a}}^{2} = 2.5 \times 10^{6}\ \text{M}^{-1}$, with $\Delta\Delta G = 11.8\ \text{kJ mol}^{-1}$. ITC data showed that this negative cooperativity was entirely entropically driven, with the enthalpic term alone corresponding to positive cooperativity.

The structural and dynamical properties of the monomers in the dimeric CAP^{N} were studied using NMR. They observed that the structure of $\text{cAMP}_2\text{-CAP}^{\text{N}}$ and apo-CAP^{N} were symmetric and the structure of the ligated and un-ligated monomers of $\text{cAMP}_1\text{-CAP}^{\text{N}}$ had structures similar to $\text{cAMP}_2\text{-CAP}^{\text{N}}$ and apo-CAP^{N} respectively. They observed that motions on the microsecond to millisecond timescale not observed in apo-CAP^{N} were activated in both monomers upon binding the first ligand. Upon binding the second ligand, a tightening of the entire protein occurs, with motions on the microsecond to millisecond and the picosecond to nanosecond timescale were suppressed. Figure 1.13, demonstrates the changes in structure and dynamics observed in the study.

As the binding of the first cAMP did not cause a change in the mean conformation of the second ligand binding site they concluded that the negative allostery observed in CAP^{N} was driven by changes in protein dynamics upon ligand binding. There is however, no such mechanistic description for how cooperative binding can occur in the full-length protein.

A number of experimental techniques , such as ITC, have been used to study the thermodynamic parameters of the cooperative binding events in full-length CAP. These have shown negative cooperativity in most instances although not to the extent seen in CAP^{N} [59, 120–124]. This weaker negative cooperativity unfortunately means that studies isolating cAMP_1-CAP are much less likely to be possible for full-length CAP than for the truncated form. Interestingly, there is evidence that the cooperativity of cAMP binding is dependent on the ionic strength of the solution. Takahashi et al. [120] found that increasing the concentration of KCl switched the cooperativity from being negative at low concentrations to being strongly positive at high concentrations.

Toncrova and McLeish recently developed a coarse grained model for allosteric protein homodimers and applied this model to the cooperative binding in CAP [54, 55]. They modelled the motions in the protein homodimer as a single slow harmonic breathing mode for each subunit. Each slow mode was also coupled to a number of localised fast modes.

They assumed that the binding of a cAMP molecule scaled the spring constant of the slow mode of local subunit by a factor, β, and the coupling between the subunits by a different factor, α. Hence, binding of the second cAMP did the same for the other subunit (meaning the spring constant coupling the two monomers was scaled by α^2). They were then able to show that the cooperativity of binding could be changed from negative to positive by adjusting the spring constant of the breathing modes, the coupling spring constant between the subunits and the factors by which the binding of cAMP affects the spring constants. The theory behind this model is discussed further in section 2.4

1.6 Outline of the Current Work

This thesis explores the hypothesis that changes to global low frequency fluctuations of the protein carry an allosteric signal in CAP. This is approached with the following structure. Chapter 2 introduces the basic theory behind the computational methods used, such as MD, normal mode analysis (NMA) and principal component analysis (PCA). Likewise, chapter 3 explores the experimental techniques, ITC and X-ray crystallography, used in this thesis to study protein structure and thermodynamic properties.

Chapter 4 explores this hypothesis by using a variety of coarse grained methods in an attempt to reduce the complexity of the system and explore the possibility of manipulating the allostery in CAP. Chapter 5, complements the previous chapter by studying the allostery in CAP using atomistic modelling techniques such as MD to validate the coarse grained models and investigate potential pathways for the allosteric signalling. Chapter 6 is the experimental chapter with results and discussion from experiments to support the computational results in the previous two chapters, investigating both the structure and thermodynamics of CAP during these allosteric events. These chapters are followed by conclusions and closing remarks.

Part II

Theory and Background of Computational and Experimental Techniques

Chapter 2

Theory

A range of computational techniques are available for studying the dynamics of biomolecules. The techniques used in this thesis range from atomistic modelling to coarse graining the protein to reduce the degrees of freedom of the protein as much as possible. In this chapter, these techniques are introduced and the theory behind them discussed.

2.1 Molecular Dynamics Background and Theory

MD is a very powerful tool for modelling atoms and molecules (in this case biomolecules) and studying their movement [125]. To run any MD simulation, two things are necessary. First a molecular model is needed: This consists of the coordinate positions of the atoms in the system and information about which atoms are bonded to each other. Secondly, a mathematical model of the interactions between atom pairs, termed a force field, is needed (see 2.1.1).

For large biomolecules such as proteins, it is not possible to predict their tertiary and quaternary structures using purely computational techniques, so the input coordinates are taken from crystal structures that are determined experimentally.

Using the force field, it is then possible to calculate the force acting on each individual atom (at any point in time and with any initial velocity) by the relationship:

$$\mathbf{F}_i(t) = -\nabla_{\mathbf{r}_i} V, \tag{2.1.1}$$

where $\mathbf{F}_i$ gives the total force acting on atom i, $\mathbf{r}_i$ represents the atom's Cartesian coordinates and V is the potential energy of the system. Using Newton's laws of motion the acceleration, $\ddot{\mathbf{r}}_i$, of each individual atom can then be calculated as shown in the equation:

$$\ddot{\mathbf{r}}_i(t) = m_i^{-1}\mathbf{F}_i(t) \tag{2.1.2}$$

From integration of the acceleration with respect to time, it is then possible to calculate the velocities and positions of the atoms after discrete time steps (see 2.1.2). Repeating this process allows one to build a trajectory, describing the positions, velocities and accelerations of the individual atoms in the system as a function of time.

2.1.1 Force Fields

As mentioned above, a force field is a potential energy function characterising the pairwise interactions of the atoms in the system. There are two main types of force fields used in MD simulations. These are all atom (AA) force fields, which treat every atom individually; and united atom (UA) force fields, which group non-polar hydrogen atoms with the heavier atom to which they are bonded, treating this group as a single "united atom". [126, 127]

A typical force field, such as the one in equation 2.1.3 (the AMBER force field) is made up of four terms, the first three terms being bonded terms and the fourth being a non bonded term.

$$\begin{aligned} V &= \sum_{bonds} K_r(r-r_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} \frac{v_n}{2}[1+\cos(n\phi-\delta)] \\ &+ \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{4\pi\varepsilon_0 R_{ij}}\right]. \end{aligned} \tag{2.1.3}$$

The first term in the potential function describes bond lengths as harmonic springs with spring constant K_r and a deviation from the equilibrium bond length (r_0) of ($r - r_0$). The bond angles are also described by a harmonic function (second term) with K_θ representing the bending force constant and ($\theta - \theta_0$) giving the deviation of the bond angle from the equilibrium angle, θ_0. The third term gives the potential energy associated with the dihedral angles in the system. Here, $v_n/2$ is the force constant, ϕ is the dihedral angle and δ gives the phase angle. The fourth term in the potential function contains both a Lennard-Jones potential modelling the Van der Waals interaction and the electrostatic term given by Coulomb's law. In this term,

A_{ij} and B_{ij} are parameters in the Lennard-Jones potential that describe the strength and repulsion between pairs of non-bonded atoms, q_i and q_j are the partial electronic charges of atoms i and j, and R_{ij} is their interatomic distance.

Other terms can sometimes be added to the force field, to model hydrogen bonding in the system, polarisation terms or implicit water (see sections 2.1.10 and 2.1.11).

2.1.2 Molecular Dynamics Integrators

For a molecular dynamics simulation, a potential function, $V(\mathbf{r}_i)$ (section 2.1.1) determines the interactions between the particles of an N particle system with coordinates, $\mathbf{r}_i$. The force acting on atom i is then given as [128]:

$$\mathbf{F}_i = -\frac{\partial V(\mathbf{r}_i)}{\partial \mathbf{r}_i}. \tag{2.1.4}$$

This force can be used with Newton's equations of motion for a system of N particles (equation 2.1.2) along with initial conditions to solve the acceleration, $\ddot{\mathbf{r}}_i(t)$, of particle i. Subsequently, the velocity, $\dot{\mathbf{r}}_i(t)$, and the future position, $\mathbf{r}_i(t+\delta t)$, of particle i can be calculated by solving these coupled ordinary differential equations.

Solving these ordinary differential equations is achieved through finite difference methods. The molecular dynamics integrator defines the form and method of the finite difference equations used. For an integrator to be successful, a number of conditions need to be met. Firstly, at the limit of the step size, $\Delta t \to 0$, the integrator should reproduce the original differential equations and should reproduce the analytical solution to a tolerable accuracy at more reasonable time steps. It should also be efficient, so larger time steps can be used and computational time can be reduced. The algorithm for an integrator also needs to be time reversible to mirror Newton's equations. Finally, the volume in phase space and the total energy of the system need to be conserved throughout the simulation.

Verlet Algorithm

The Verlet algorithm [129] returns the position of atom i, at time $t + \Delta t$ by summing the Taylor expansions of $\mathbf{r}(t - \Delta t)$ and $\mathbf{r}(t + \Delta t)$ at time t and rearranging, to give:

$$\mathbf{r}_i(t+\Delta t) = 2\,\mathbf{r}_i(t) - \mathbf{r}_i(t-\Delta t) + m_i^{-1}\mathbf{F}_i(t)(\Delta t)^2 \tag{2.1.5}$$

Therefore, for each step, the position, $\mathbf{r}_i(t)$, and force, $\mathbf{F}(t)$, at time t and position $\mathbf{r}_i(t-\Delta t)$ at time $t-\Delta t$, are needed to calculate the position $\mathbf{r}_i(t+\Delta t)$, at time $t+\Delta t$. The velocity is not directly returned by the algorithm, but can be calculated for time t with:

$$\mathbf{v}_i(t) = \frac{\mathbf{r}_i(t+\Delta t) - \mathbf{r}_i(t-\Delta t)}{2\ \Delta t} \tag{2.1.6}$$

This algorithm encounters a problem at $t = 0$, where the position at $-\Delta t$ is required. The solution to this problem is either to use a Taylor expansion about $\mathbf{r}(t)$, or to use a different algorithm for the first time step.

Leapfrog Algorithm

The leapfrog algorithm is algebraically equivalent to the Verlet algorithm, however it makes adjustments to how some of the quantities are calculated [128]. This algorithm uses the position and force at time t and the velocity at the previous half time step, $\mathbf{v}_i\left(t-\frac{\Delta t}{2}\right)$, to calculate the velocity at the next half time step $\mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)$. These half-step velocities are subsequently used to calculate the positions at the next full time step $\mathbf{r}_i(t+\Delta t)$. The leapfrog algorithm can be written as:

$$\begin{aligned} \mathbf{v}_i\left(t+\frac{\Delta t}{2}\right) &= \mathbf{v}_i\left(t-\frac{\Delta t}{2}\right) + m_i^{-1}\mathbf{F}_i(t)(\Delta t) \qquad (2.1.7) \\ \mathbf{r}_i\left(t+\Delta t\right) &= \mathbf{r}_i(t) + \mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)\Delta t \qquad (2.1.8) \end{aligned}$$

The velocity at time t is the average of $\mathbf{v}_i\left(t-\frac{\Delta t}{2}\right)$ and $\mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)$.

The AMBER simulation package [130] uses an algorithm similar to the leapfrog algorithm, however it is more complicated due to the use of bond length constraints, thermostats and barostats.

2.1.3 Energy Minimisation

The crystal structure of a protein taken from the PDB can safely be assumed to be at a local free energy minimum. However, the size of hydrogen atoms and their ability to diffract X-rays is normally not sufficiently large for their positions to be resolved. The positions of the hydrogens are therefore included by a computer program that adds them using typical bond length and angles for amino acid residues. This can lead to a few bad contacts between atoms appearing close together in the resulting 3D structure. When combined with a knowledge that the force field used (although parametrised by experimental data) does not model all of the interactions completely accurately, one can safely assume the starting protein structure will not be in a local minimum of the force field. This creates a need to minimise the energy of a starting protein structure prior to the an MD simulation to remove any high energy contacts, which would have lead to failure of the simulation.

A number of algorithms can be employed to perform the energy minimisation such as a steepest descent algorithm, a conjugate gradient algorithm [131] or the L-BFGS algorithm [132]. A minimisation can either be run for a chosen number of steps or until the force acting on an atom is less than a chosen tolerance for all atoms in the system.

Steepest Descent

Steepest descent minimisation works on the principle that a function, $V(x)$, decreases fastest at a point **a** in the direction of the negative gradient of V, $-\nabla V(a)$. So a path to a local minimum can be found by the algorithm:

$$x_{n+1} = x_n - \gamma_n \nabla V(x_n), n >= 0 \tag{2.1.9}$$

for small enough values of γ_n, such that $V(x_n) >= V(x_{n+1})$. The value of γ_n is variable upon each iteration, so eventually convergence to the local minimum will be achieved.

One limitation of steepest descent minimisation is that convergence can be slow when ∇V is very small, for example when the system is close to a local minimum.

Conjugate Gradient

The conjugate gradient method is an iterative method for solving systems of linear equations that is frequently used to minimise functions [131]. With the conjugate gradient method it is possible to minimise a system with an $N \times N$ sparse system matrix in a maximum N steps when there are no rounding errors. This method starts with an initial estimate, x_0, of the minimum. Successive new estimates, $x_0, x_1, x_2...$ of the minimum are made, such that each estimate x_i is closer to the minimum than x_{i+1}.

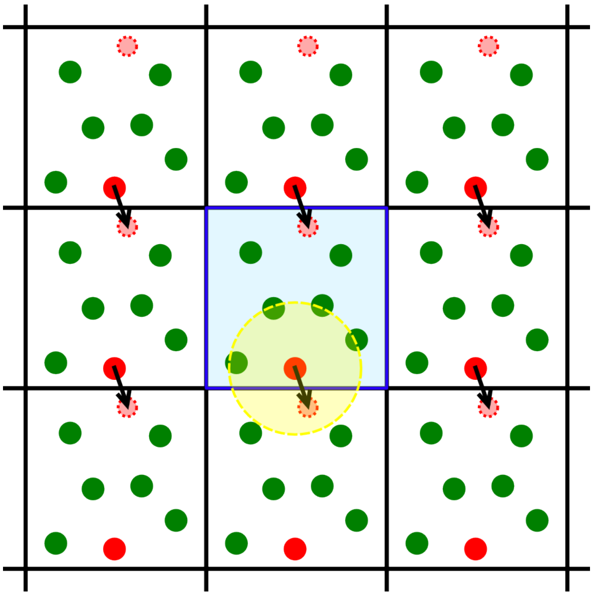


Figure 2.1: Graphical representation of periodic boundary conditions in a two dimensional system. The illustration shows a snapshot of the whole system and the movement of the red particle during the next time step. A (yellow) cut-off radius is shown around the red particle in the blue central box. This particle interacts with any particle whose centre of mass falls within this radius, whether it is in the same box or across the periodic boundary. [128]

L-BFGS Minimisation

The limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) is a quasi-Newton method for minimisation of a multivariable function. L-BFGS implicitly approximates the inverse Hessian matrix, using vectors. It then uses this Hessian approximation to steer the search through variable space for the local minimum. The L-BFGS method stores the positions and gradients calculated for a number of previous steps, which are used to implicitly guide the implicit inverse Hessian search of variable space.

2.1.4 Periodic Boundary Conditions

Periodic boundary conditions are used for molecular dynamics simulations of proteins that use explicit water. For periodic boundary conditions, the simulation box is surrounded by images of itself; particles near one face of the box interact with particles near the opposite face of the box. Also, if a particle leaves the simulation box at one side, it will instantaneously enter the box at the opposite side, keeping the number of particles in the simulation box constant (Figure 2.1)

The reason for using periodic boundary conditions is to ensure the atoms at the edge of the box undergo the same interactions as those in the centre of the box and that the particles in the box do not run off into the space surrounding it.

The word 'box' has been used to describe the shape of the repeating unit for the periodic boundary conditions. It can in fact be any shape that can tessellate perfectly in three dimensions such as a cube or a truncated octahedron.

2.1.5 Ensembles

There are a number of thermodynamic ensembles that can be used to derive the properties of thermodynamic systems in equilibrium [133]. Without temperature and pressure regulation, molecular dynamics simulations would naturally be performed in the microcanonical ensemble (NVE) of microstates, assuming no loss of precision. This ensemble has a constant number of particles (N), constant volume (V) and constant system energy (E). However, this ensemble does not match the conditions that most experiments are performed in, so instead MD simulations can be performed in a number of different ensembles which better match experimental conditions using devices such as temperature and pressure regulation (described in sections 2.1.6 and 2.1.7 respectively).

One such ensemble is the canonical ensemble (NVT), which has a constant number of particles (N), constant volume (V) and constant temperature (T). The characteristic state function for this ensemble is the Helmholtz free energy, $A = -k_{\mathrm{B}}T \ln Z(N, V, T)$, where $Z(N, V, T)$ is the canonical partition function. In this thesis, this ensemble is used during the temperature equilibration stage of the molecular dynamics simulations. The volume is kept constant by not allowing the volume of the periodic box to change, whereas the temperature is controlled by using a weak coupling thermostat as described in 2.1.6.

Another ensemble is the isothermal-isobaric (Gibbs) ensemble (NPT), where both the pressure and temperature have specified average values and the number of particles is fixed; the volume of the system is allowed to fluctuate. The characteristic state function for this ensemble is the Gibbs free energy, $G = -k_{\mathrm{B}}T \ln Z(N, P, T)$. For this thesis this ensemble is used for all production run MD simulations.

2.1.6 Thermostat

To regulate the temperature of MD simulations a thermostat is needed. Thermostat algorithms are modifications to the Newtonian MD scheme which scale the velocities of the molecules in the system so that the average kinetic energy of the system matches that of the system at a target temperature [133]. Thermostat algorithms modify Newton's second law (equation 2.1.2) to

$$\ddot{\mathbf{r}}_i(t) = m_i^{-1}\mathbf{F}_i(t) - \gamma(t)\dot{\mathbf{r}}_i(t), \tag{2.1.10}$$

introducing a pseudo friction coefficient $\gamma(t)$. A positive value of $\gamma(t)$ indicates heat being drawn from the system into the heat bath, whereas a negative value indicates heat flowing in the opposite direction. This thermostat, however does not take the stochastic nature of thermal fluctuations into account so should not be used in a stochastic MD simulation.

The thermostat used in the MD simulations presented in this thesis, the Langevin thermostat, makes a few changes to equation 2.1.10, which becomes

$$\ddot{\mathbf{r}}_i(t) = m_i^{-1}\mathbf{F}_i(t) - \gamma_i(t)\dot{\mathbf{r}}_i(t) + m_i^{-1}\mathbf{R}_i(t). \tag{2.1.11}$$

In the Langevin Thermostat algorithm the atomic friction coefficient, $\gamma_i(t)$ is positive definite, hence it draws heat from the system. A stochastic force, $\mathbf{R}_i$ is introduced, giving heat to the system. This stochastic force is uncorrelated with the previous velocities and forces, has a time average of zero and a mean square value of

$$\left\langle \mathbf{R}_i^2 \right\rangle = 2\ m_i\gamma_i k_B T_0, \tag{2.1.12}$$

where T_0 is the target temperature for the simulation. The stochastic force is also independent for each atom, Cartesian coordinate axis and time step.

2.1.7 Barostat

In constant pressure simulations such as the NPT ensemble, a barostat is needed to regulate the pressure. This barostat adjusts the volume of the unit cell at each timestep to make the observed pressure of the system approach the target pressure. The barostat used is a weak coupling barostat, called the Berendsen barostat [134]. The pressure of the system given by

$$P = \frac{2}{3V}(E_k - \Xi), \tag{2.1.13}$$

where E_k is the kinetic energy of the system and Ξ is the internal virial for pair additive potentials, given by:

$$\Xi = -\frac{1}{2}\sum_{i<j} \mathbf{r}_{ij} \cdot \mathbf{F}_{ij}. \tag{2.1.14}$$

Here $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{F}_{ij}$ is the force from particle i acting on particle j. The barostat scales the distances between particles, $\mathbf{r}_{ij}$, which in turn scales the internal virial and the pressure of the system.

2.1.8 TIP3P water model

TIP3P is a model to explicitly express water in molecular dynamics simulations [135, 136]. Three point water models, such as TIP3P, represent water as a rigid molecule with fixed bond lengths and angles. Each water molecule has three interaction sites, corresponding to each of the three atoms in a water molecule. The potential function between two water molecules, m and n is then represented by

$$\sum_{i}^{m}\sum_{j}^{n} \frac{q_i q_j e^2}{r_{ij}} + \frac{A}{r_{\mathrm{OO}}^{12}} - \frac{B}{r_{\mathrm{OO}}^{6}}, \tag{2.1.15}$$

where i and j are the charged sites on m and n. As with the force field in equation 2.1.3, A and B are the Lennard-Jones parameters, q_i are the partial charges and r_{ij} is the distance between charged sites. The distance between oxygen atoms of any two molecules of water is also given as r_{OO}. For TIP3P water these parameters are given in table 2.1 along with the geometry used for a water molecule.

Using rigid water molecules such as TIP3P speeds up MD simulations as bonded terms do not need to be calculated. Using a three point water molecule such as TIP3P also benefits from a reduction in the calculation time when compared to 4-site (TIP4P) or 5-site (TIP5P) water models. These reductions in calculation time are significant when a large number of water molecules are simulated, such as in biomolecular simulations. Hence, the TIP3P water model was used for the simulations in this study.

Parameter	Value
$r(\text{OH})$ / Å	0.9572
$\angle$HOH / deg	104.52
$A \times 10^{-3}$ / kcal Å^{12} mol^{-1}	582.0
B / kcal Å^{6} mol^{-1}	595.0
$q(\text{O})$	-0.834
$q(\text{H})$	0.417

Table 2.1: Geometry and parameters used for the TIP3P water model.

2.1.9 SHAKE Bond Length Constraints

SHAKE is an algorithm used in MD to constrain the length of certain bonds during the simulation [137]. Using SHAKE during MD simulations allows longer time steps to be used as the size of the time step is determined by the fastest motion in the system, which in an unconstrained simulation is bond stretching. Using longer time steps in turn increases the total time a simulation can be run for. For the molecular dynamics simulations in this thesis, SHAKE bond length constraints are used for all bonds involving hydrogen, therefore a timestep of 2 fs is employed.

2.1.10 Poisson-Boltzmann Implicit Solvent Model

Poisson-Boltzmann (PB) solvents are frequently used to model the effect solvation has on the electrostatic interactions and have been shown to reliably reproduce the energetics and conformations of a variety of systems compared to explicit solvents [130, 138]. PB implicit solvent models describe all solvent and dissolved salt as a continuum, while representing the solute as a dielectric body with a shape defined by the coordinates and cavity radii of its atoms [139]. An electrostatic field in the solvent and the solute region is created by point charges at the atomic positions of the solute and is computed by solving the PB equation [140, 141]:

$$\nabla \cdot [\varepsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] = -4\,\pi\rho(\mathbf{r}) - 4\,\pi\lambda(\mathbf{r}) \sum_i z_i c_i \exp\left(-z_i \phi(\mathbf{r})/k_B T\right). \tag{2.1.16}$$

Here, ε represents the position dependent dielectric constant, $\phi(\mathbf{r})$ the electrostatic potential, $\rho(\mathbf{r})$ the solute charge, z_i the charge of ion type i, c_i the bulk number of ion type i and $\lambda(\mathbf{r})$ the Stern layer masking function, determining the accessibility of position $\mathbf{r}$ to the ions in solution. In assisted model building with energy refinement (AMBER) the PB equation is solved using a finite difference approach on a numerical grid.

The electrostatic contribution to the free energy of solvation is then [142]:

$$\Delta G_{\mathrm{el}} = \frac{1}{2} \sum_i q_i \left(\phi_{\mathrm{sol}}(\mathbf{r}_i) - \phi_{\mathrm{vac}}(\mathbf{r}_i) \right), \tag{2.1.17}$$

where ϕ_{sol} and ϕ_{vac} are the electrostatic potentials of the solute partial charges (q_i) in implicit solvent and in vacuum respectively.

2.1.11 Generalised-Born Implicit Solvent Model

Due to the computational cost of solving equation 2.1.16 for the PB implicit solvent model, often an approximation called the Generalised-Born (GB) implicit solvent model is used [130,142,143]. This method models the atoms in the protein as spheres of radius R_i (the effective Born Radii). The dielectric constant inside these spheres is 1 and the dielectric constant outside matches that of the target solvent (80 for water at 300 K). For the GB model to be successful the effective Born radii need to be accurately estimated [142].

The GB model then approximates the value of ΔG_{el} with [143, 144]:

$$\begin{aligned} \Delta G_{\mathrm{el}} &\approx \Delta G_{\mathrm{GB}} \\ &= -\frac{1}{2} \sum_i \sum_j \frac{q_i q_j}{f_{\mathrm{GB}}(r_{ij}, R_i, R_j)} \left(1 - \frac{\exp(-\kappa f_{\mathrm{GB}})}{\varepsilon} \right). \end{aligned} \tag{2.1.18}$$

Here the R_i and R_j are the effective Born radii of atoms i and j and r_{ij} the distance between them. The Debye-Huckel screening parameter, κ, introduces the electrostatic screening effects of salt [144], while f_{GB} is a smooth function, which makes the GB equation mimic the relevant equations of electrostatics. It is often defined as [143]

$$f_{\mathrm{GB}} = \left(r_{ij}^2 + R_i R_j \exp\left(\frac{-r_{ij}^2}{4 R_i R_j} \right) \right)^{1/2}, \tag{2.1.19}$$

although it can be defined in other ways [142, 145].

2.1.12 Nonpolar implicit solvent models

The nonpolar (NP) component of the free energy of solvation is often estimated using the solvent accessible surface area (SASA) [146,147]. This model yields a nonpolar free energy of:

$$\Delta G_{\mathrm{np}} = \gamma \cdot \mathrm{SASA} + c, \tag{2.1.20}$$

where γ is the surface tension coefficient and c is the free energy of solvation for a point solute (i.e. when SASA=0).

Recently, a few studies have shown that the NP solvation free energy is better approximated by splitting it into an attractive and a repulsive part [130, 148, 149], ΔG_{att} and ΔG_{rep}, such that;

$$\Delta G_{\mathrm{np}} = \Delta G_{\mathrm{att}} + \Delta G_{\mathrm{rep}}. \tag{2.1.21}$$

In one such model, proposed by Tan *et al.* the repulsive contribution is modelled with SASA

$$\Delta G_{\mathrm{rep}} = \gamma \cdot \mathrm{SASA} + c \tag{2.1.22}$$

and the attractive contribution can be approximated by the van der Waals attractive interaction between the solvent (v) and solute (u):

$$\Delta G_{\mathrm{att}} \approx \langle U^{\mathrm{uv}}_{\mathrm{att}} \rangle \tag{2.1.23}$$

as was first observed by Chandler *et al.* [150, 151]. The solute-solvent van der Waals interaction energy, $U^{\mathrm{uv}}_{\mathrm{att}}$, itself can be expressed as:

$$U^{\mathrm{uv}}_{\mathrm{att}} = \sum_{a=1}^{N_s} \int \rho_{\mathrm{a}w}(\mathbf{r}_{aw}) V_{\mathrm{att}}(\mathbf{r}_{\mathrm{aw}}) \mathrm{d}\mathbf{r}_{aw}. \tag{2.1.24}$$

Here, $\rho_{aw}(\mathbf{r}_{\mathrm{aw}})$ is a solvent distribution function around solute atom a, $\mathbf{r}_{aw}$ is the solvent distance and $V_{\mathrm{att}}(\mathbf{r}_{aw})$ is the attractive van der Waals potential. The sum is over all of the solute atoms, N_s, and the integration is over the volume occupied by the solvent. Equation 2.1.24 is often simplified further, by approximating $\rho_{aw}(\mathbf{r}_{aw})$ as a constant density distribution.

2.2 Normal Mode Analysis

A normal mode of vibration of a protein involves an oscillation of the atoms in the protein. In a normal mode, all of the atoms vibrate at the same frequency and in phase. For proteins the normal modes that are believed to be most important for protein function are the low frequency vibrations, which cause the largest deviations in the protein structure and have the largest contributions to the heat capacity and entropy.

NMA is a technique that can often probe the large structural rearrangements of proteins better than MD, due to its ability to identify motions on slow time-scales. It is less computationally expensive in terms of CPU time than MD, but more costly in terms of the memory required.

NMA finds the normal modes of a protein at a local minimum on its potential energy surface. In NMA, a couple of approximations are made; firstly that the protein cannot cross energy barriers to another local minimum and secondly that the shape of the energy well can be approximated as a harmonic potential.

In NMA the conformation of a system containing N particles is described as a $3N$-dimensional position vector,

$$\mathbf{q} = (x_1, y_1, z_1, ..., x_N, y_N, z_N)^T. \tag{2.2.25}$$

The harmonic approximation looks at small displacements of the particles from their equilibrium conformation, $\mathbf{q}_0$, where the net force acting on all N particles in the system is zero.

Consider first the Taylor expansion of the potential energy $V(\mathbf{q})$ about a minimum. [152–154]

$$\begin{aligned} V(\mathbf{q}) &= V(\mathbf{q}^0) + \sum_i^{3N} \left(\frac{\partial V}{\partial q_i}\right)^0 (q_i - q_i^0) \\ &+ \frac{1}{2}\sum_i^{3N}\sum_j^{3N} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 (q_i - q_i^0)(q_j - q_j^0) + ... \end{aligned} \tag{2.2.26}$$

Here, i and j represent N 3D coordinates for the particles. The potential energy function used can be any potential energy function, such as the one used in MD (see equation 2.1.3), as long as it is at a minimum. By definition, at a local minimum, the first derivative of the potential is equal to zero, so this term in the expansion of the potential disappears. The first term of the Taylor expansion can also be set to zero by studying the potential relative to this minimum potential. If we say the displacement of the coordinate i from equilibrium is $\Delta q_i = (q_i - q_i^0)$, equation 2.2.26 simplifies to:

$$V(\mathbf{q}) = \frac{1}{2}\sum_i^{3N}\sum_j^{3N}\left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 \Delta q_i \Delta q_j$$

$$V(\mathbf{q}) = \frac{1}{2}\mathbf{\Delta q}^T\mathbf{H}\mathbf{\Delta q} \tag{2.2.27}$$

Where $\mathbf{H}$ is the $3N \times 3N$ Hessian matrix given by the Laplacian of the potential, $\nabla^2 V(\mathbf{q})$, containing the force constants between neighbouring particles. The velocities of the particles are then given by:

$$\dot{q}_i = dq_i/dt = d(\Delta q_i)/dt = (\dot{\Delta q_i}). \tag{2.2.28}$$

It is possible to solve the equations of motion of the system by looking at the kinetic energy, T, as a quadratic function of the velocities:

$$T(\dot{\mathbf{q}}) = \frac{1}{2}\sum_i^{3N}\sum_j^{3N} m_{ij}\dot{q}_i\dot{q}_j = \frac{1}{2}(\dot{\mathbf{\Delta q}})^T\mathbf{M}(\dot{\mathbf{\Delta q}}) \tag{2.2.29}$$

Here m_{ij} gives the elements of the $N \times N$ diagonal matrix, $\mathbf{M}$, containing the masses of the particles; each mass repeated three times due to its three Cartesian coordinates.

The Lagrangian, $L = T - V$, of the system is given by

$$L = \frac{1}{2}\left((\dot{\mathbf{\Delta q}})^T\mathbf{M}(\dot{\mathbf{\Delta q}}) - \mathbf{\Delta q}^T\mathbf{H}\mathbf{\Delta q}\right). \tag{2.2.30}$$

Then, solving the Euler-Lagrange equation

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

yields

$$\mathbf{M}(\ddot{\mathbf{\Delta q}}) + \mathbf{H}\mathbf{\Delta q} = 0. \tag{2.2.31}$$

The modes of motion for the system, represented by $\mathbf{u}_k$, are 3N-dimensional vectors that are solutions to equation 2.2.31:

$$\mathbf{u}_k = \mathbf{a}_k \exp(-i\omega_k t). \tag{2.2.32}$$

Where for each mode of motion, ω_k is the angular frequency and $\mathbf{a}_k$ is a complex vector containing both the amplitude and phase factor. Differentiating $\mathbf{u}_k$ twice gives:

$$\frac{d^2\mathbf{u}_k}{dt^2} = -\omega_k^2 \mathbf{u}_k. \tag{2.2.33}$$

Combining this with equation 2.2.31 gives the equation of motion as

$$\mathbf{H}\mathbf{u}_k = \omega_k^2 \mathbf{M}\mathbf{u}_k, \tag{2.2.34}$$

which can be expressed in matrix form as

$$\mathbf{HU} = \mathbf{MU\Lambda} \tag{2.2.35}$$

where $\mathbf{U}$ is the $3N \times 3N$ matrix. Each column is a solution, $\mathbf{u}_k$, and $\mathbf{\Lambda}$ is the diagonal matrix with the diagonal elements $\lambda_k = \omega_k^2$. Converting the solutions of the equation of motion $\mathbf{u}_k$ into mass-weighted coordinates (by mass-weighting the Hessian ($\tilde{\mathbf{H}} = \mathbf{M}^{-1/2}\mathbf{H}\mathbf{M}^{-1/2}$)) is then possible, giving the solutions $\tilde{\mathbf{u}}_k = \mathbf{M}^{1/2}\mathbf{u}_k$, which are the normal modes of the system. Each $\tilde{\mathbf{u}}_k$ is the eigenvector for normal mode k (where $k = 1, ..., 3N$) and each λ_k is the frequency of the normal mode squared.

For systems where every particle has the same mass, m, the mass matrix, $\mathbf{M}$, becomes the identity matrix multiplied by m. So in this case, the eigenvector for the normal mode is scaled by the square root of m, such that $\tilde{\mathbf{u}} = \sqrt{m}\mathbf{u}$.

2.2.1 Atomistic Normal Mode Analysis

Atomistic normal mode analysis uses an AA force field, as is seen in equation 2.1.3 as the potential for which a harmonic approximation is found. In this case, the particles mentioned in section 2.2 are the atoms of the protein, so if the protein has N atoms the Hessian matrix created from its potential is of size $3N \times 3N$. NMA needs to be performed at a local minimum, and the tolerance for it being performed away from this minimum is very low (the maximum force acting on any atom shouldn't exceed around 0.001 kJ mol^{-1} nm^{-1}) [155]. Therefore, before it can be performed a rigorous L-BFGS minimisation (see 2.1.3) is performed to obtain a local energy minimum for the force field used [132].

A draw-back for using atomistic NMA is that it can be computationally expensive in terms of memory. The memory needed to perform a fully atomistic normal mode analysis is proportional to N^2 for a macromolecule with N atoms. To reduce the amount of memory needed for normal mode calculations, force fields where bond lengths and bond angles have been fixed can be used [65]. Elastic network models were later introduced, reducing the memory needed further [156].

2.2.2 Elastic Network Model Background and Theory

An elastic network model (ENM) is a simplified model for computing the normal modes of a macromolecule. The simplicity of ENMs means that computational cost is greatly reduced; hence the normal modes of much larger molecules can be computed.

The first ENMs, inspired by the work of Tirion [156], modelled the protein's structure as a network of point masses connected by simple Hookean spring potentials for point mass pairs separated by a distance less than a specified cut-off r_c. Tirion noticed that using this potential, the low frequency normal modes were insensitive to small changes in protein structure, therefore it was safe to assume that the experimental crystal structure was a local minimum. This removed the need for minimisation, which is necessary for fully atomistic normal mode analysis.

The simplest ENMs, such as the Gaussian network model (GNM) and the anisotropic network model (ANM) use one point mass per amino acid located at the position of the C$_\alpha$ atoms and use the same spring constant between all connected amino acid pairs [157–160]. There are ENMs that use more complex potentials however, such as Hinsen's model, which uses distance dependent spring constants [161, 162].

2.2.3 Gaussian Network Model

In the GNM it is assumed that the position of the point masses undergo Gaussian distributed fluctuations around the equilibrium position, [157, 158] inspired by Flory's theory of polymer networks [163]. The potential used in the GNM is given by

$$V_{GNM} = \frac{1}{2}\sum_i \sum_j \gamma_{ij}(\mathbf{R}_{ij} - \mathbf{R}^0_{ij})(\mathbf{R}_{ij} - \mathbf{R}^0_{ij}) = \frac{1}{2}\sum_i \sum_j \gamma_{ij}(\Delta\mathbf{R}_{ij})^2, \tag{2.2.36}$$

where $\mathbf{R}^0_{ij}$ is the distance vector between the equilibrium position vectors, $\mathbf{R}^0_i$ and $\mathbf{R}^0_j$, of nodes i and j. $\mathbf{R}_{ij}$ is the instantaneous distance vector between the position vectors, $\mathbf{R}_i$ and $\mathbf{R}_j$ and $\Delta\mathbf{R}_{ij}$ is the difference between the equilibrium and instantaneous distance vectors between node pair i and j. Also, γ_{ij} is the spring constant between nodes i and j, but this is usually treated as a constant γ between all node pairs that are within the specific cutoff distance $\mathbf{R}_c$.

For the GNM the fluctuations of the C_α positions and their cross-correlations are controlled by the $N \times N$ Kirchoff or connectivity matrix, Γ_{ij}, defined as

$$\Gamma_{ij} = \begin{cases} -1 & \text{for } i \neq j, \text{ and } \mathbf{R}_{ij} < \mathbf{R}_c \\ 0 & \text{for } i \neq j, \text{ and } \mathbf{R}_{ij} > \mathbf{R}_c \\ -\sum_{k,k\neq i}\Gamma_{ik} & \text{for } i = j. \end{cases} \tag{2.2.37}$$

Here, the Kirchoff matrix multiplied by γ is the counterpart of the Hessian matrix, so diagonalisation gives N eigenvalues and eigenvectors. In this instance the eigenvectors are one dimensional distance vectors meaning it is not possible to obtain information describing the 3-dimensional directionality of motion using the GNM. However, they still provide very useful information about the normal modes of macromolecules. For example, the inverse of the Kirchoff matrix can be used to see the correlation between fluctuations Δr_i and Δr_j

$$\langle \Delta r_i \cdot \Delta r_j \rangle = \frac{k_B T}{\gamma}\left[\Gamma^{-1}\right]_{ij} \tag{2.2.38}$$

A web server called El Nemo that allows a user to submit protein coordinates in a PDB file and runs the GNM on the protein is available [164].

2.2.4 Anisotropic Network Model

The anisotropic network model succeeded the Gaussian network model, incorporating 3D directions of the fluctuations of proteins [159, 160]. The ANM, like the GNM, usually uses the positions of the C$_\alpha$ atoms as the positions of the nodes in the network and treats all residue masses as equal. In the ANM, the potential used is of the form

$$V_{ij} = \begin{cases} \frac{\gamma_{ij}}{2}(R_{ij} - R^0_{ij})^2 & R_{ij} \leq R_\mathrm{C} \\ 0 & R_{ij} > R_\mathrm{C} \end{cases}. \tag{2.2.39}$$

In this case, the instantaneous distance vector, R_{ij}, is in one of the three directions in the 3D space such that the position of every node is described by three distance vectors. This is of course the same for the equilibrium position, R^0_{ij}, and the difference between the two, ΔR^0_{ij}. The spring constant between pairs of nodes, γ_{ij}, is usually treated as a universal spring constant, γ, for all node pairs. The Hessian matrix, **H**, for a system of N nodes is then made up of $N \times N$ off diagonal 3×3 sub-matrices of **H**, which take the form

$$\mathbf{H}_{ij} = -\frac{\gamma_{ij}}{(R^0_{ij})^2} \begin{bmatrix} (x^0_{ij})^2 & x^0_{ij}y^0_{ij} & x^0_{ij}z^0_{ij} \\ y^0_{ij}x^0_{ij} & (y^0_{ij})^2 & y^0_{ij}z^0_{ij} \\ z^0_{ij}x^0_{ij} & z^0_{ij}y^0_{ij} & (z^0_{ij})^2 \end{bmatrix}, \tag{2.2.40}$$

where each $x_{ij} = (x_j - x_i)$. The diagonal sub-matrices of **H** are given by

$$\mathbf{H}_{ii} = -\sum_{j,j\neq i} \mathbf{H}_{ij}. \tag{2.2.41}$$

As illustrated in section 2.2, the equations of motion for this Hessian are solved to calculate the normal modes.

A web server utilising the anisotropic network model is available [165]. A user can submit a PDB file containing a protein's coordinates to this server and the server will run the ANM on the protein.

The elastic network model used in this study has the same potential function as the ANM, however the Hessian matrix is mass-weighted before diagonalisation, using the mass of the residue as the C$_\alpha$ mass [156].

2.2.5 Rotational-Translational Block Model

The rotational-translational block (RTB) method for normal mode analysis treats a protein as a sequence of n_b rigid-body blocks of consecutive amino acids. The rigid-body blocks can be made of one, or a few consecutive residues [166]. Each block is allowed six degrees of freedom, three rotational and three translational when calculating the normal modes.

To do this the $3N \times 3N$ Hessian matrix, $\mathbf{H}$, is projected into a $6n_b \times 6n_b$ Hessian, $\mathbf{H}_b$, with a basis defined by the rotations and translations of the rigid blocks. This is given by

$$\mathbf{H}_b = \mathbf{P}^T \mathbf{H} \mathbf{P}, \tag{2.2.42}$$

where $\mathbf{P}$ is a $3N \times 6n_b$ matrix, whose columns contain the vectors associated with the local translations and rotations of each block. The eigenvectors, $\mathbf{U}$, associated with the normal modes are then given by:

$$\mathbf{U} = \mathbf{P}\mathbf{U}_b. \tag{2.2.43}$$

Here, the rows of $\mathbf{U}_b$ are the eigenvectors of the projected Hessian, $\mathbf{H}_b$. The frequencies of the modes are the eigenvalues of $\mathbf{H}_b$.

2.3 Principal Component Analysis

Principal component analysis (PCA) is a technique applied in the post-processing of an extended MD trajectory [167–170]. This method, originally proposed by Karplus and Kushick [167], allows one to study functionally important modes of motion, without the need to make the assumption that the system is harmonic. These important modes are found for a macromolecule of N atoms through diagonalisation of the $3N \times 3N$ covariance matrix, defined as

$$\sigma_{ij} = \langle (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle) \rangle$$

$$\boldsymbol{\sigma} = \left\langle (\mathbf{q} - \langle \mathbf{q} \rangle)(\mathbf{q} - \langle \mathbf{q} \rangle)^{\mathrm{T}} \right\rangle, \tag{2.3.44}$$

where $\mathbf{q}$ is the $3N$-dimensional vector describing the macromolecule's atomistic coordinates. $\langle\ \rangle$ denotes an average over time, such that $\langle \mathbf{q} \rangle$ is the average coordinates for all the snapshots in the MD trajectory. The covariance matrix is mass-weighted using the mass matrix [171, 172], $\mathbf{M}$,

$$\tilde{\boldsymbol{\sigma}} = \mathbf{M}^{1/2}\boldsymbol{\sigma}\mathbf{M}^{1/2}, \tag{2.3.45}$$

which ultimately gives a more accurate description of the modes. The same analysis with a non-mass-weighted covariance is called Quasiharmonic analysis. $\tilde{\boldsymbol{\sigma}}$ is a symmetric semi-definite matrix, so it can be diagonalised to

$$\mathbf{u}^{-1}\tilde{\boldsymbol{\sigma}}\mathbf{u} = \left\langle \mathbf{Q}^2 \right\rangle. \tag{2.3.46}$$

Again, the full motion of the system is approximated to a harmonic motion, with eigenvectors, $\mathbf{u}$, that describe the principal modes of motion for the system. The eigenvalues $\langle Q_{ii}^2 \rangle$ are the classical variances for the resulting new coordinates, Q_i. Here, larger eigenvalues correspond to larger fluctuations in the systems structure. Depending on the methodology used, the first six modes can be rotational and translational modes, as is the case for NMA. If this is the case, they are ignored when calculating the entropy.

It is possible to determine the frequency, ω_i, of mode i by extending the harmonic approximation further [173]. Looking at the classical partition function for each mode:

$$Z_i = \int\int \exp -\frac{1}{kT}\left(\frac{1}{2}P_i^2 + \frac{1}{2}\omega_i^2 Q_i^2\right) \mathrm{d}P_i\, \mathrm{d}Q_i. \tag{2.3.47}$$

Here, P_i represents the momentum of the mode. Thermodynamic averages of function, X of Q_i and P_i can be computed by the equation:

$$\langle X(Q_i, P_i) \rangle = \frac{\int\int X(Q_i, P_i) \exp -\frac{1}{kT}\left(\frac{1}{2}P_i^2 + \frac{1}{2}\omega_i^2 Q_i^2\right) \mathrm{d}P_i\, \mathrm{d}Q_i}{Z_i}. \tag{2.3.48}$$

which can be used to show that:

$$\left\langle Q_i^2 \right\rangle = \frac{kT}{\omega_i^2}. \tag{2.3.49}$$

Treating the system as a harmonic oscillator, however, does not take into account that some of the motions observed in a molecular dynamics simulation may be caused by random diffusion. Methods for identifying whether modes are a consequence of random diffusion are discussed in section 2.5 along with other methods for analysing normal modes from PCA and NMA.

2.4 Super Coarse Graining Proteins Background and Theory

The secondary structures of proteins such as α-helices and β-sheets often create regions of greater rigidity (than the rest of the protein), connected by short residue chains that create, what can be described as a hinge. Hawkins and McLeish suggested a coarse-grained (CG) model, treating these rigid bodies as solid objects coupled together by Hookean springs. This is mathematically equivalent to representing the protein as a small number of point masses coupled by springs. Therefore, it is a simple model of the protein undergoing coupled oscillations [56, 57]. Toncrova and McLeish later applied this method to protein homodimers, taking CAP as the example system [54, 55].

These CG models (termed super coarse grained (SCG) models) assume that the minimum energy of the system occurs when none of the elements of the system undergo any displacement, $\mathbf{x} = \mathbf{0}$. A force (or elasticity) matrix, $\mathbb{K}$, is then built for the system of coupled oscillators, whose elements are related to the spring constants in the CG model.

The energy of the CG model (regardless of the exact model) can then be expressed in terms of the Lagrangian,

$$L = \frac{1}{2}\left(\dot{\mathbf{x}}^T\mathbf{M}\dot{\mathbf{x}} - \mathbf{x}^T\mathbb{K}\mathbf{x}\right). \tag{2.4.50}$$

Here $\mathbf{x}$ is the vector showing the displacement of the rigid domains from equilibrium and $\dot{\mathbf{x}}$ is the velocity vector. $\mathbf{M}$ is the inertia matrix. For the model of ligand binding, it is assumed that the changes in mass are negligible, so the term containing the inertia matrix is ignored.

By comparing this Lagrangian to the one in section 2.2, one can see that the elasticity matrix is analogous to the Hessian matrix for the system. Therefore the slow vibrational modes can be solved in the same way as described in section 2.2, ignoring the mass-weighting of the system. The free energy and entropy of the system can then either be computed as described below.

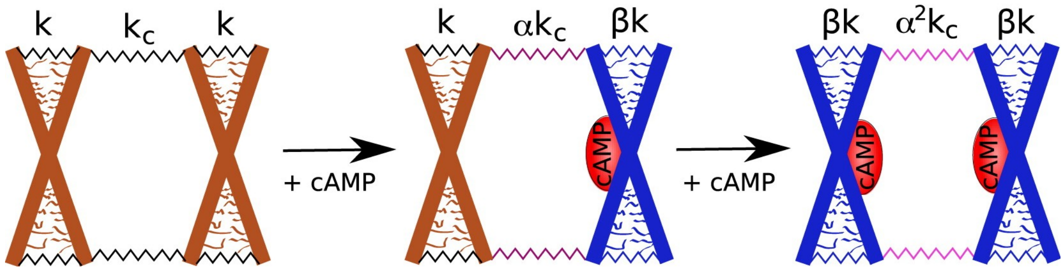


Figure 2.2: The CAP CG model devised by Toncrova and McLeish [54, 55]. In this picture, the crosses represent the CAP monomers as hinged rods and the fine jagged lines represent the springs connecting the rods. k is the internal spring constant of each subunit and k_c is the coupling spring constant between the subunits and (x_1, x_2) represents the displacement of the two monomers. The binding of cAMP to a monomer scales k_c by α and k for the bound monomer by β. Taken from [54].

It is possible to calculate the partition function, Z, for the system using the theorem for multidimensional Gaussian integrals. Using this method, the partition function is given as

$$Z = \frac{(2\pi k_\mathrm{B}T)^n}{|\mathbb{K}|^{\frac{1}{2}}}, \tag{2.4.51}$$

where n is the number of coordinates in the system and $|\mathbb{K}|$ is the determinant of $\mathbb{K}$. The conformational free energy is then

$$G_\mathrm{conf} = -k_\mathrm{B}T \ln Z = \frac{1}{2} k_\mathrm{B}T \ln |\mathbb{K}| + C \tag{2.4.52}$$

with C being a constant that disappears when calculating free energy differences,

$$\Delta\Delta G_\mathrm{conf} = G^\mathrm{c}_{\mathrm{holo}_2} - 2G^\mathrm{c}_{\mathrm{holo}_1} + G^\mathrm{c}_\mathrm{apo}. \tag{2.4.53}$$

Where $G^\mathrm{c}_{\mathrm{holo}_2}$, $G^\mathrm{c}_{\mathrm{holo}_1}$ and $G^\mathrm{c}_\mathrm{apo}$ represent the conformational free energy for a doubly bound, singly bound and unbound protein respectively. Combining equations 2.4.52 and 2.4.53 yields:

$$\Delta\Delta G_\mathrm{conf} = \frac{1}{2} k_\mathrm{B}T \ln \frac{|\mathbb{K}_\mathrm{apo}||\mathbb{K}_{\mathrm{holo}_2}|}{|\mathbb{K}_{\mathrm{holo}_1}|^2}. \tag{2.4.54}$$

Toncrova and McLeish, applied a CG model to the allostery of CAP [54, 55]. They modelled each subunit of the CAP dimer as solid rods, connected by springs with spring constant k, to make a ‘scissor’ domain. The two subunits were fixed in space at their hinge and were also coupled together by a spring constant k_c and the displacement of the rods from equilibrium is given by the vector $\mathbf{x} = (x_1, x_2)$ (see figure 2.2). The elasticity matrix, $\mathbb{K}$, is then built by looking at the springs that displacements along the directions in $\mathbf{x}$ are coupled to, and is given as:

$$\mathbb{K} = \begin{pmatrix} k + k_c & -k_c \\ -k_c & k + k_c \end{pmatrix}. \tag{2.4.55}$$

One pitfall of this model is that it allows only 2 degrees of freedom, because it allows no change in the centre of mass of the monomers. This results in only 2 normal modes; one in phase motion where both scissors open simultaneously and one out of phase motion, where one scissor opens while the other closes. Another weakness is that it models each monomer of CAP as a single scissor domain, however the flexibility between the LBD and the DNABD suggests that a model with two scissor domains to each monomer would perhaps model CAP more accurately. This new model would therefore imply that the model by Toncrova and McLeish models just the LBDs of CAP.

The work in chapter 4 of this thesis, will expand on the work by Toncrova and McLeish, tackling some of the issues with the model mentioned above.

2.5 Analysis of Normal Modes and Principal Components

The modes of motion of a protein as calculated by techniques such as NMA and PCA are characterised by a set of eigenvectors, which define the shapes of the normal modes and a set of eigenvalues, which contain information about the frequency of the motion in the case of NMA or the magnitude of the motion in the case of PCA. There are a number of techniques for analysing the shape of the normal modes of motion. The techniques used in this thesis are described in this section.

2.5.1 B-factors

It is possible to calculate the B-factor (B_i) of an atom i in a system from its normal modes by utilising the relationship between B-factors and RMSDs:

$$B_i = \frac{8kT\pi^2}{3m_i} \sum_{v}^{n} \frac{|\mathbf{u}_i^2(v)|}{\omega_v^2}. \tag{2.5.56}$$

Where $\mathbf{u}_i^2(v)$ is the normal mode vector and $\omega_{\mathbf{v}}$ is the frequency of the normal mode. The sum is performed over a set of normal modes, usually less than the total number, as the higher frequency modes contribute much less to the B-factor than the low frequency modes.

2.5.2 Cross Correlation

The cross correlation, C_{ij}, shows whether or not atoms i and j move in similar directions in a correlated manner for a set of normal modes:

$$C_{ij} = \sum_{v} \left(\frac{\mathbf{u}_i(v) \cdot \mathbf{u}_j(v)}{(|\mathbf{u}_i(v)|^2 |\mathbf{u}_j(v)|^2)^{0.5}} \right) \tag{2.5.57}$$

A value of $C_{ij} = 1$ signifies that the motion of atoms i and j is perfectly correlated, whereas $C_{ij} = -1$ signifies perfectly anti-correlated motion. The cross correlation is therefore useful for identifying regions of the protein that undergoes correlated motion.

When two different modes, $\mathbf{u}(v)$ and $\mathbf{v}(v)$, calculated by different methods are being compared the cross correlation equation becomes:

$$C_{ij} = \frac{\mathbf{u}_i(v) \cdot \mathbf{v}_j(v)}{(|\mathbf{u}_i(v)|^2 |\mathbf{v}_j(v)|^2)^{0.5}}. \tag{2.5.58}$$

This is most useful when the atoms $i = j$, so the similarities in the motion of individual atoms can be compared between different modes. This is how this analysis is performed in this thesis.

2.5.3 Overlaps

The overlap between two normal modes, $I_{v_1v_2}$, is a measure of the extent of similarities between two normal modes, v_1 and v_2:

$$I_{v_1v_2} = \left(\frac{|\mathbf{u}(v_1) \cdot \mathbf{u}(v_2)|}{(|\mathbf{u}(v_1)|^2 |\mathbf{u}(v_2)|^2)^{0.5}} \right). \tag{2.5.59}$$

An overlap of 1 indicates that the motion for normal mode v_1 is identical to the motion for normal mode v_2, whereas a value of 0 indicates that the motion is completely different (or orthogonal). As the eigenvectors calculated by NMA and PCA form an orthonormal basis, the overlap of an eigenvector with another from the same set of normal modes will always equal 1 if $v_1 = v_2$ or 0 if $v_1 \neq v_2$. Therefore the overlap is mainly useful for comparing normal modes from two different systems (with the same number of atoms included in the eigenvector) or from two different analyses.

2.5.4 Subspace Overlaps

When comparing multiple sets of eigenvectors, looking at the individual overlaps of modes separately can become overwhelming. The subspace overlap quantifies the similarities between the subspaces spanned by two sets of n orthonormal vectors, $\mathbf{u}$ and $\mathbf{v}$. It is defined as:

$$s(\mathbf{u}, \mathbf{v}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\mathbf{u}_i \cdot \mathbf{v}_j}{(|\mathbf{u}_i|^2 |\mathbf{v}_j|^2)^{0.5}} \right)^2, \tag{2.5.60}$$

with a subspace overlap of 1 corresponding to $\mathbf{u}$ and $\mathbf{v}$ occupying the same subspace.

It is also possible to quantify the extent to which a single eigenvector $\mathbf{W}$ exists within a subset of eigenvectors, $\mathbf{u}$, by reducing equation 2.5.60 to:

$$s(\mathbf{u}, \mathbf{W}) = \sum_{i=1}^{n} \left(\frac{\mathbf{u}_i \cdot \mathbf{W}}{(|\mathbf{u}_i|^2 |\mathbf{W}|^2)^{0.5}} \right)^2. \tag{2.5.61}$$

This is useful for seeing the extent to which the differences between two PDB structures can be accounted for by a subset of the normal modes or principal components.

2.5.5 Normalised Subspace Overlaps

In principal component analysis, when looking at the principal components from two simulations of the same system, the overlap of the fluctuations can provide a measure for the convergence of the sampled space [174]. It is possible to perform this analysis in terms of covariance matrices. The normalised subspace overlap, s_{norm}, between two covariance matrices A and B is defined as:

$$s_{\text{norm}}(A,B) = 1 - \frac{\sqrt{\text{tr}[A^{1/2} - B^{1/2}]^2}}{\sqrt{\text{tr}A + \text{tr}B}}, \tag{2.5.62}$$

where tr$A$ denotes the trace of covariance matrix A. The normalised overlap, s_{norm}, is 1 iff the two matrices are identical and 0 if the subspaces sampled by the two matrices are orthogonal.

This method has a couple of advantages over the subspace overlap (section 2.5.4) of the first n_A and n_B eigenvectors of matrix A and B. Firstly by using the covariance matrices, all of the eigenvectors are included in the calculation. Secondly, the normalised subspace overlap, by using the covariance matrices includes the information contained in the eigenvalues. The subspace overlap on the other hand ignores the eigenvalues, meaning that differences between eigenvectors with large eigenvalues reduce the overlap by the same amount as the same differences between eigenvectors with small eigenvectors.

2.5.6 Cosine Content of Principal Components

In PCA, the principal component (PC) of the eigenvector, $\mathbf{u}_i$, with index i is defined as the projection ($p_i(t)$) of the MD trajectory onto the eigenvector. Where:

$$p_i(t) = \mathbf{u}_i \cdot (\mathbf{x}(t) - \bar{\mathbf{x}}(t)). \tag{2.5.63}$$

Here $\mathbf{x}(t)$ indicates the positional vector of the system at time t and $\bar{\mathbf{x}}(t)$ indicates the average orientation.

The shape of the PCs contains information about whether good sampling has been achieved in the MD simulation. Hess showed in a study of the PCs of a potential-less Brownian motion simulation that the shape of the PCs of random diffusion are cosine shaped with the number of periods equal to half the index of the principal component [174,175]. Therefore cosine shaped PCs in protein simulations are indicative that the the motion captured by the PCA for that specific mode is caused by diffusive motion rather than the motion of a protein trapped within an energy well. The cosine content, c_i, is a measure of how much a PC resembles a cosine, and hence how much of its motion is caused by random diffusion:

$$c_i = \frac{2\left(\int_0^T p_i(t)\cos(i\pi t)\ \mathrm{d}t\right)^2}{T\int_0^T p_i(t)^2\ \mathrm{d}t} \tag{2.5.64}$$

Here the integral is over the length, T, of the simulation used to calculate the covariance matrix. A cosine content of 1 indicates that the PC is a perfect cosine, so a cosine content in the region of 1 would indicate that the PC is driven by random diffusion. The fluctuation of the cosine content can be large between different PCs from the same PCA, so a low value of the average cosine content can not be used as an indicator of good sampling. However, it can be said that if the first PC strongly resembles a cosine, the sampling in the simulation is far from converged [174].

2.6 Entropy and Free Energy of Normal Modes

Using statistical mechanics it is possible to calculate an estimate to the conformational free energy and conformational entropy of a system using the frequency of vibrations such as the normal modes. For this method, the partition function, ζ_k, for normal mode k is given as

$$\zeta_k = \frac{1}{1-\exp\left(\frac{-\hbar\omega_k}{k_{\mathrm{B}}T}\right)} \tag{2.6.65}$$

and the total partition function is the product of the partition functions for all k normal modes

$$\zeta_{\mathrm{tot}} = \prod_k^{3N}\zeta_k. \tag{2.6.66}$$

When looking at one mole of the system, the total partition function, Z_{tot}, is the product of the partition functions, ζ_{tot}, for all the molecules, such that

$$Z_{\mathrm{tot}} = (\zeta_{\mathrm{tot}})^{N_{\mathrm{A}}}$$

The conformational free energy, E_{conf}, of a system with internal energy, U, and entropy, S, is given by

$$E_{\mathrm{conf}} = U - TS_{\mathrm{conf}} \tag{2.6.67}$$

and can be calculated for a mole of the system from the total partition function with

$$E_{\mathrm{conf}} = -N_{\mathrm{A}} k_{\mathrm{B}} T \ln(\zeta_{\mathrm{tot}}) = -k_{\mathrm{B}} T \ln(Z_{\mathrm{tot}}) \tag{2.6.68}$$

By combining equations 2.6.67 and 2.6.68, the entropy of a mole of the system is

$$S_{\mathrm{conf}} = \frac{U}{T} - k_{\mathrm{B}} \ln(Z_{\mathrm{tot}}). \tag{2.6.69}$$

With the internal energy given by

$$U = U(0) + k_{\mathrm{B}} T^2 \frac{\partial(\ln(Z_{\mathrm{tot}}))}{\partial T}. \tag{2.6.70}$$

Here $U(0)$ is the zero point energy of the system. Some generalisations are frequently used when calculating the conformational entropy for the system these are discussed along with the methods for calculating the entropies in 2.6.1 and 2.6.2

Equation 1.4.19 in section 1.4 shows the allosteric free energy for homotropic allostery. The same relationship applies for the entropy, giving an allosteric entropy of:

$$\Delta\Delta S = S_{\mathrm{holo}_2} - 2S_{\mathrm{holo}_1} + S_{\mathrm{apo}}. \tag{2.6.71}$$

If this value is positive, the allostery of the system has positive entropic cooperativity and if it is negative, it exhibits negative entropic cooperativity.

To get a more accurate picture of the cooperativity of the binding events, however, one should consider the Gibbs free energy difference, $\Delta\Delta G$ (see 1.4.19), which takes into account the enthalpic contribution to binding as well as other entropic contributions.

The cooperativity of binding can also be measured with the cooperativity constant, C, as shown in equation 1.4.21 of section 1.4. The cooperativity constant can also be determined by

$$C = \frac{k_{\mathrm{holo}_2}/k_{\mathrm{holo}_1}}{k_{\mathrm{holo}_1}/k_{\mathrm{apo}}}, \tag{2.6.72}$$

where k_s are the effective spring constants of the system, taken as the product of the eigenvalues of the Hessian matrix (λ_i) over a chosen number of modes for each of the systems, $k = \prod_i \lambda_i$.

2.6.1 Classical Method

The entropy of a one dimensional quantum harmonic oscillator with a frequency ω is given by:

$$S_{\mathrm{ho}} = k_\mathrm{B}\left(\frac{\alpha}{e^{\alpha}-1} - \ln\left[1-e^{-\alpha}\right]\right), \tag{2.6.73}$$

where $\alpha = \frac{\hbar\omega}{k_\mathrm{B}T}$.

For a harmonic system of N atoms in Cartesian coordinates, the system has $3N$ normal modes. In this case, each vibration contributes to the entropy, therefore equation 2.6.73 becomes:

$$S_{\mathrm{ho}} = k_\mathrm{B}\sum_{i}^{3N}\frac{\frac{\hbar\omega_i}{k_\mathrm{B}T}}{e^{\frac{\hbar\omega_i}{k_\mathrm{B}T}}-1} - \ln\left[1-e^{-\frac{\hbar\omega_i}{k_\mathrm{B}T}}\right], \tag{2.6.74}$$

where ω_i is the frequency of normal mode i. This harmonic entropy provides an upper bound to the conformational entropy of the system, such that $S_{\mathrm{ho}} > S$.

2.6.2 Schlitter Method

Starting from the entropy of a one-dimensional oscillator (equation 2.6.73), Schlitter introduced an approximation to S_{ho} [171],

$$S' = \frac{k_\mathrm{B}}{2}\ln\left[1+\frac{e^2}{\left(\frac{\hbar\omega}{k_\mathrm{B}T}\right)^2}\right] \tag{2.6.75}$$

with $e = \exp(1)$; showing that $S' > S_{\mathrm{ho}}$. For this method, every degree of freedom is treated as a quantum harmonic oscillator, with the equipartition theory (equation 2.3.49) used to link the variance to the frequency of a quantum harmonic oscillator to give:

$$S' = \frac{k_\mathrm{B}}{2}\ln\left[1+\frac{k_\mathrm{B}Te^2}{\hbar^2}\right]\left\langle q^2\right\rangle, \tag{2.6.76}$$

where Q is the classical variance.

In a multidimensional system, the Schlitter method becomes a way of calculating a maximum entropy estimate for a macromolecule directly from the covariance matrix

[171]. The eigenvalues, $\langle Q_{ii}^2 \rangle$, of the covariance matrix, $\tilde{\boldsymbol{\sigma}}$, as calculated by equation 2.3.46 are used to calculate the entropy:

$$\begin{aligned} S' &= \frac{1}{2} k_\mathrm{B} \sum_{i=1}^{3N} \ln \left[1 + \frac{k_\mathrm{B} T e^2}{\hbar^2} \langle Q_{ii}^2 \rangle \right] \\ &= \frac{1}{2} k_\mathrm{B} \ln \left(\prod_{i=1}^{3N} \left[1 + \frac{k_\mathrm{B} T e^2}{\hbar^2} \langle Q_{ii}^2 \rangle \right] \right) \end{aligned} \tag{2.6.77}$$

Taking the product of the diagonal elements of a diagonalised matrix is equivalent to finding the determinant of that matrix, which itself will not have changed upon diagonalisation, therefore equation 2.6.77 can be rewritten as

$$S' = \frac{1}{2} k_\mathrm{B} \ln \det \left[\mathbf{1} + \frac{k_\mathrm{B} T e^2}{\hbar^2} \mathbf{M} \boldsymbol{\sigma} \right]. \tag{2.6.78}$$

Like the classical method, this approximation only holds for $\hbar\omega << k_\mathrm{B}T$, so it fails for the high frequency motions. However, these high frequency motions contribute least to the entropy, meaning that the Schlitter approximation is likely to be good approximation for the entropy.

The Schlitter method has been used to calculate the absolute entropy of ideal gases and Lennard-Jones Fluids [176] and entropy of cooperative ligand binding in DNA [177, 178], with sufficient agreement to experiment or analytical methods.

2.7 MM/PBSA

Molecular mechanics/Poisson Boltzmann surface area method (MM/PBSA) is a method for calculating the free energy of protein ligand-binding events [179, 180]. In this method, the free energy of a complex system is estimated using a set of structures, typically collected from a molecular dynamics simulation. These structures are stripped of any solvent and counter-ions and the solvated free energy for the system, $\bar{G}_\mathrm{solvated}$, calculated by:

$$\bar{G}_\mathrm{solvated} = \bar{E}_\mathrm{MM} + \bar{G}_\mathrm{PBSA} - TS_\mathrm{MM}. \tag{2.7.79}$$

Here, $\bar{E}_\mathrm{MM}$ is the average molecular mechanical energy, $\Delta\bar{G}_\mathrm{PBSA}$ is the solvation free energy determined by the implicit PB model (section 2.1.10) and a non-polar model (section 2.1.12). The entropy is often determined using either NMA (section 2.2) or

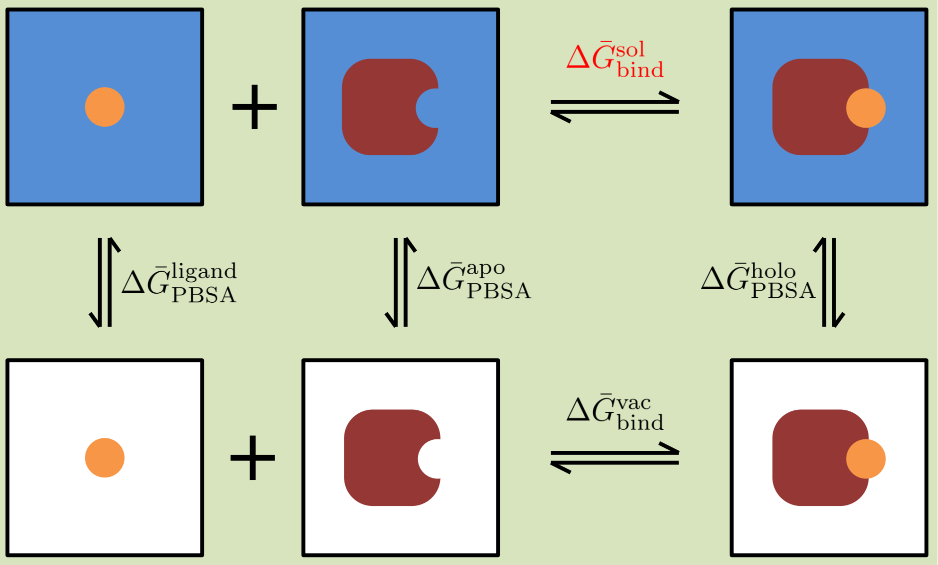

$\Delta\bar{G}^{\text{sol}}_{\text{bind}}$
$\Delta\bar{G}^{\text{ligand}}_{\text{PBSA}}$
$\Delta\bar{G}^{\text{apo}}_{\text{PBSA}}$
$\Delta\bar{G}^{\text{holo}}_{\text{PBSA}}$
$\Delta\bar{G}^{\text{vac}}_{\text{bind}}$

Figure 2.3: Thermodynamic cycle for calculating the binding free energy of a protein-ligand complex. Systems are shown in blue boxes if they are solvated, or white boxes if they are in vacuum. The free energy of binding is shown in red. Adapted from [180]

PCA (section 2.3).

The free energy of binding of a solvated protein-ligand complex can be determined by subtracting the solvated free energies of the protein ($\Delta\bar{G}^{\text{apo}}_{\text{solvated}}$) and the ligand ($\Delta\bar{G}^{\text{ligand}}_{\text{solvated}}$) from the solvated free energy of the protein-ligand complex ($\Delta\bar{G}^{\text{holo}}_{\text{solvated}}$):

$$\Delta\bar{G}^{\text{sol}}_{\text{bind}} = \Delta\bar{G}^{\text{holo}}_{\text{solvated}} - (\Delta\bar{G}^{\text{apo}}_{\text{solvated}} + \Delta\bar{G}^{\text{ligand}}_{\text{solvated}}), \tag{2.7.80}$$

Where $\Delta\bar{G}^{\text{sol}}_{\text{bind}}$ is the average free energy of binding in solvent. This process is shown as a thermodynamic cycle in figure 2.3. To determine the allosteric free energy, the difference between the free energy of binding for the two consecutive binding events is calculated (see section 1.4).

Chapter 3

Experimental Techniques used to study Protein Structure and Dynamics

While using computer modelling can be useful for studying the dynamics and thermodynamics of cooperative binding in proteins, without experimental validation no confidence can be put in the computational results. This section looks at a couple of experimental techniques, X-ray crystallography (section 3.1) and ITC (section 3.2), used to validate the computational studies in this thesis.

3.1 X-ray Crystallography

X-ray crystallography is the most important tool for the determination of a macromolecule's structure. It also provides useful information about the macromolecules dynamics with the B-factors, however it does not provide any information about the timescale of the motion observed. This section looks at X-ray crystallography as a method for studying protein structure and dynamics, looking at some of the challenges that need to be overcome and theory used.

3.1.1 The Crystal Lattice

In a crystal, the molecules are arranged in a lattice, a construct of regular periodic units, repeated in three dimensions. These periodic units contain a motif, which in

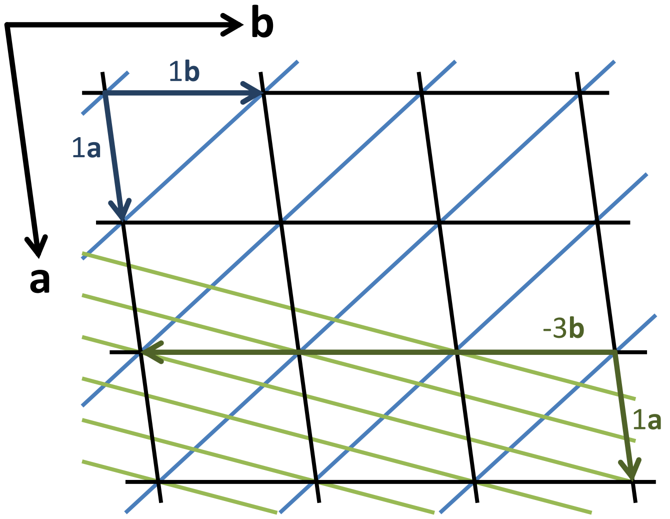


Figure 3.1: 2-dimensional representations of two sets of lattice “planes” and their Miller indices for a 2-dimensional lattice with dimensions **a** and **b**. The Miller indices $(h\,k)$ of the sets of lattice “planes” are $(1\,1)$ for the blue lines and $(3\,\text{-}1)$ for the green lines. Adapted from [181].

the case of protein crystallography, consists of one or more protein subunits, water molecules and often other molecules from the crystallisation cocktail. The periodic unit and its constituent motif is called the unit cell (see section 3.1.3).

The lattice has a set of lattice planes, which intersect the lattice in a periodic and distinct manner. Each set of lattice planes can be defined by its Miller indices, hkl, a set of integers indicating the reciprocal of the number of intercepts a set of lattice planes have with the unit cell. Figure 3.1 shows 2-dimensional representations of two sets of lattice planes and their Miller indices. The interplanar distance vector $\mathbf{d}_{hkl}$, is the distance vector between two adjacent planes from the same set of planes with Miller indices hkl.

3.1.2 Protein Crystallisation

For a protein to crystallise, the free energy of crystallisation;

$$\Delta G_{\mathrm{c}} = \Delta H_{\mathrm{c}} - T\Delta S_{\mathrm{c}} \tag{3.1.1}$$

needs to be less than zero, therefore the entropy lost from the removal of the rotational and translational degrees of freedom needs to be overcome (around -30 to -100 kJ mol^{-1} at 300 K). The enthalpic contribution to the crystallisation free energy, ΔH_{c}

is favourable, but not to the extent that is needed to overcome the entropic penalty as there are generally very few contacts between the protein molecules in a crystal. The value of ΔH_c is typically either moderately negative such as for lysozyme (-70 kJ mol^{-1}) [182] or insignificant such as for ferritin, apoferritin and lumazine synthase (around 0 kJ mol^{-1}) [183–185]. Therefore, another factor must contribute to the free energy of crystallisation for it to occur. This is the entropy gained upon releasing structured water from the surface of the protein upon crystallisation. Typically between 5 and 30 ordered water molecules are released from the surface of the protein upon crystallisation, which contributes between 30 and 200 kJ mol^{-1} at 300 K [181, 186].

Taking all of these contributions of the free energy of crystallisation, into account:

$$\Delta G_\mathrm{c} = \Delta H_\mathrm{c} - T(\Delta S_\mathrm{protein} + \Delta S_\mathrm{solvent}), \tag{3.1.2}$$

which has are two large entropy terms of opposite sign with a small difference between their magnitude. Therefore fairly small changes to each of these entropy terms can have drastic effects on the overall entropy change and hence whether or not the protein crystallises.

A negative ΔG_c, although necessary for crystallisation, is not the only condition that need to be met for crystallisation to occur. The kinetic barriers associated with nucleation and self assembly also need to be traversed. Phase transitions necessary for crystallisation such as these are initiated by “critical events” which are kinetically driven and hard to predict for complex systems such as protein solutions. Nevertheless, all established physical principles of phase formation are applicable to protein crystallisation, as has been shown by atomic force microscopy (AFM) [181].

The nucleation phase transition occurs when the protein solution becomes supersaturated. At this point the protein solution is metastable, so will eventually separate into protein plus a saturated solution. If the conditions are correct, and provided a nucleation event occurs, the protein will separate out into the form of a crystal. The solubility phase diagram for a protein solution can be seen in figure 3.2.

There are two types of nucleation events that can occur in protein crystallisation. For the first, homogeneous nucleation, proteins in the solution occasionally collide and form favourable interactions that overcome the entropy lost from the removal of degrees of freedom. These small groups of proteins, or nuclei, can either be broken apart by

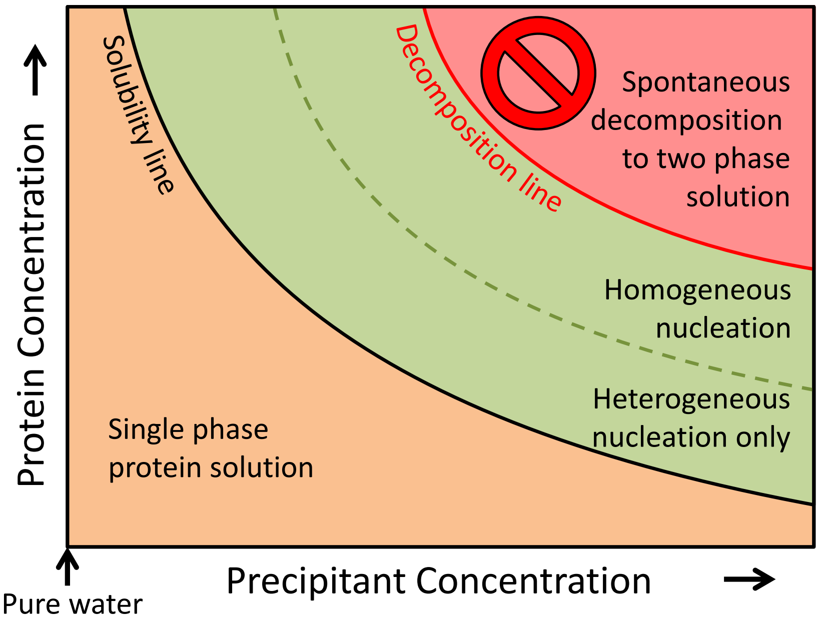

Protein Concentration
Solubility line
Decomposition line
Spontaneous decomposition to two phase solution
Homogeneous nucleation
Heterogeneous nucleation only
Single phase protein solution
Pure water
Precipitant Concentration

Figure 3.2: An Idealised solubility phase diagram for a protein solution. Below the solubility line there is a stable single phase protein-precipitant solution. The higher the precipitant concentration, the lower the protein concentration that can be achieved and vice versa. Above the decomposition line, the solution spontaneously decomposes to two phases. Between the solubility line and the decomposition line is the metastable region where the protein solution is supersaturated. It is in this region that crystallisation happens given nucleation occurs. Homogeneous nucleation tends need higher levels of supersaturation than heterogeneous nucleation. Adapted from [181].

collisions with other molecules or go on to form a larger nuclei. These events occur continuously in the supersaturated solution, with the probability of forming a larger and more stable nuclei increasing with degree of supersaturation. Once a nucleus has reached a critical size, the fluctuations in the solution cease to break it up, at which point additional collisions with other protein molecules are energetically favourable and crystallisation occurs. The second type of nucleation event is heterogeneous nucleation, which occurs at a surface of the supersaturated solution. Heterogeneous nucleation can be exploited with seeding techniques, where nucleation is artificially induced by using previously formed crystals or crystal fragments as a starting point to grow larger crystals. For spontaneous homogeneous nucleation to occur, generally a greater extent of supersaturation is needed than for heterogeneous nucleation.

To achieve protein solution supersaturation, high purity protein is needed at a sufficient concentration. Typically this has been noted to be around 10 mg/ml, however there are examples of crystallisation at much lower concentrations [181, 187]. Protein solution supersaturation can then be achieved in a number of ways, such as by the

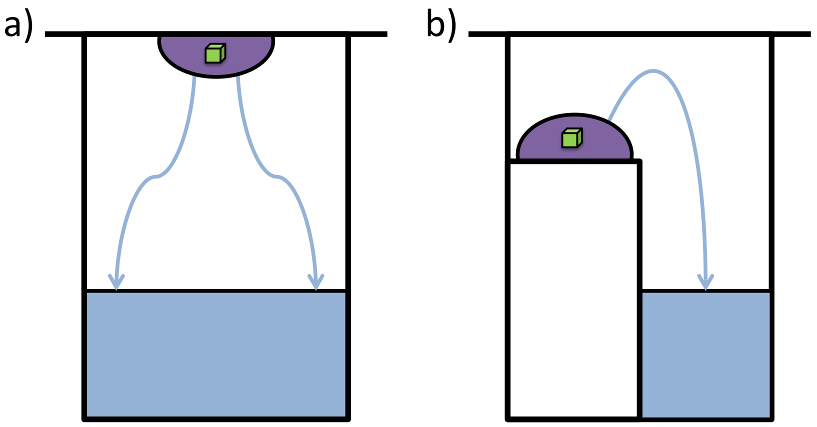


Figure 3.3: Two Vapour diffusion techniques: a) hanging drop vapour diffusion is a common method used for manual setups. The drop of protein crystallisation cocktail (purple) is typically made of 1 μl pure protein solution and 1 μl crystallisation solution from the reservoir (blue). b) Sitting drop vapour diffusion has been optimised for robotic automation, with droplets of the protein crystallisation solution typically 100 nl + 100 nl. Both methods work on the principle that the water vapour diffuses from the protein droplet to the reservoir. Adapted from [181].

addition of precipitant to the protein solution, removal of water by vapour diffusion, solvent exchange or a change of pH, among others.

The method used to reach supersaturation in this thesis was vapour diffusion. Two common vapour diffusion methods are sitting drop vapour diffusion and hanging drop vapour diffusion, which are summarised in figure 3.3. Both of these methods work on the principal that there is a drop of protein solution separate from a reservoir of precipitant solution in a closed system. The drop typically contains half the concentration of the precipitant in comparison to the reservoir, resulting in water vapour diffusing from the drop containing the protein solution to the well. This increases both the protein and the precipitant concentration, causing the solution to become supersaturated. Ideally, the protein solution gets supersaturated to the extent that homogeneous nucleation occurs. An idealised phase diagram showing the path of a successful vapour-diffusion experiment can be seen in figure 3.4.

3.1.3 Crystallisation Cocktail

The phase diagrams shown earlier are a highly idealised picture of the crystallisation problem, and in reality finding the correct conditions for crystallisation is a multidimensional problem, with a number of variables controlling whether or not crystallisation

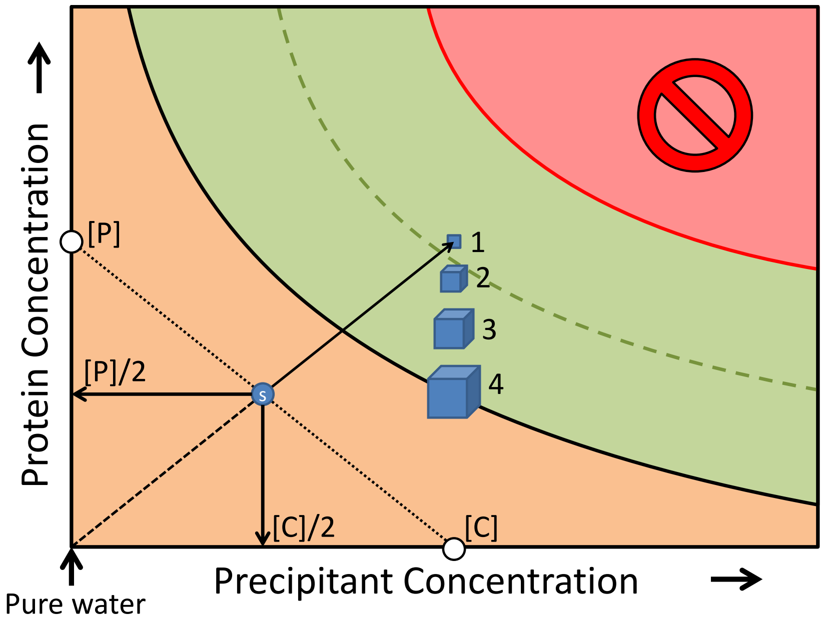


Figure 3.4: Idealised Phase diagram showing the path of a successful vapour diffusion crystallisation trial. The starting droplet (denoted S) has half the precipitant concentration of the solution in the reservoir, causing water vapour to diffuse from the droplet to the reservoir. The loss of water causes a rise in the protein and precipitant concentrations in the droplet, pushing the system into the metastable region of the phase diagram where spontaneous nucleation occurs (1). As the crystal grows (2, 3) the protein concentration in the solution reduces until the crystals reach equilibrium with the saturated solution (4), where they cease to grow. Adapted from [181].

will occur. Therefore a number of different things go into the crystallisation cocktail to try to cause the protein to crystallise.

One of these variables that contributes to whether crystallisation occurs is the precipitant used. Salts and organic compounds can both be used as precipitants. The most commonly used organic precipitant is polyethylene glycol (PEG), which was first introduced as a precipitant in the 1970s [188]. PEGs work by competing for the water molecules around the protein, thereby forcing the protein molecules to interact with each other.

Salts work in a different way; the charged ions in the salt interact with the charged residues on the surface of proteins, affecting the solubility of proteins in a complicated manner. Small quantities of salts actually increase the solubility of a protein, this is called the *salting in* effect. Larger quantities, however, do reduce the solubility; the desired effect of a precipitant.

Another control of crystallisation is the pH used; this is controlled by the buffer used for the crystallisation cocktail. A protein's minimum solubility can be found at its isoelectric point (pI) (where it has zero net charge) however evidence from the PDB shows no direct correlation between the pI and crystallisation pH, in fact the average pH of crystallisation is around 7.4 [181, 189, 190]. This lack of correlation between pH and pI implies that in many cases a net charge on the protein can increase a protein's likelihood of crystallising.

Finally, additives can be added to the crystallisation cocktail; often this is done in the process of optimising crystallisation. Additives are often added to crystallisation cocktails to tackle problems such as small crystals, poor morphology, weak diffraction patterns or poor reproducibility of results. An additive can essentially be anything that improves crystallisation, therefore there are a lot of mechanisms for how different additives work. Additives are an important strategy for optimising crystallisation, as is investigated in the study by McPherson and Cudney in a study that looks at the effects of different additives on crystallisation [191].

3.1.4 Crystallisation Screens

To try and solve the multivariable problem of finding the correct conditions that crystallisation occurs, crystallisation screens are used. Initial *a priori* knowledge about a protein, allows us to delete detrimental factors from the screen such as pH values that would denature the protein and excessive precipitant concentrations. Despite this knowledge, the crystallisation space is still extremely vast, so crystallisation screens are used.

There are a number of different types of screens that can be used in protein crystallography, such as footprint screens [192] and oversampled sparse screens [193]. The type of screen used in this thesis is a grid screen, which can commonly be found in the 24 well or 96 well format. Grid screens frequently vary two different variables in the crystallisation space, such as pH and PEG size (pH/PEG) or salt cation and PEG concentration (cation/[PEG]), but can technically be any two different variables from the crystallisation space. If a screen is successful, protein crystals form which can then be used for X-ray diffraction to solve the crystal structure.

3.1.5 X-ray Diffraction

In the classical description, X-ray diffraction is a result of the electrons in the atoms of the protein interacting with the X-rays causing them to scatter. Electrons interact with the electric field vector of an X-ray, causing it to oscillate with the same frequency as the X-ray. For the elastic Thompson scattering, utilised in X-ray crystallography the oscillating electron then emits radiation at the same frequency as the incoming wave. This is called a scattering event and can be described by a scattering vector

$$\mathbf{S} = \mathbf{s}_1 - \mathbf{s}_0, \tag{3.1.3}$$

where $\mathbf{s}_0$ is the incoming wave vector and $\mathbf{s}_1$ is the wave vector of the scattered wave. Wave vectors are vectors of length $1/\lambda$ in the propagation direction of the wave. The direction of the scattered electrons is not isotropic, but determined by the polarisation of the incoming radiation.

Scattering of X-rays by an Atom

When the X-ray diffracting electrons are in atoms, they move around the nucleus in orbits defined by their probability distributions. Therefore the amplitude of a scattered wave depends on the electron density, $\rho(r)$, of the atom. Each of the electrons in the atom are excited by the electric field vector of the X-ray causing them to scatter partial waves. These partial waves recombine to create the outgoing wave.

When looking at the scattering of two separate volume elements of the electron density, O at the origin, and P which is separated from O by the vector $\mathbf{r}$, the path difference between partial waves scattered by O and P is:

$$\Delta p = \lambda \mathbf{r} \cdot \mathbf{s}_1 - \lambda \mathbf{r} \cdot \mathbf{s}_0 = (\mathbf{s}_1 - \mathbf{s}_0) \cdot \mathbf{r}\lambda = \mathbf{S} \cdot \mathbf{r}\lambda. \tag{3.1.4}$$

The multiplication factor of λ is present because the wave vectors $\mathbf{s}_0$ and $\mathbf{s}_1$ are in units of $1/\lambda$. This is represented in figure 3.5. The path difference corresponds to a phase difference of:

$$\Delta\phi = \frac{2\pi}{\lambda}\Delta p = 2\pi \mathbf{S} \cdot \mathbf{r} \tag{3.1.5}$$

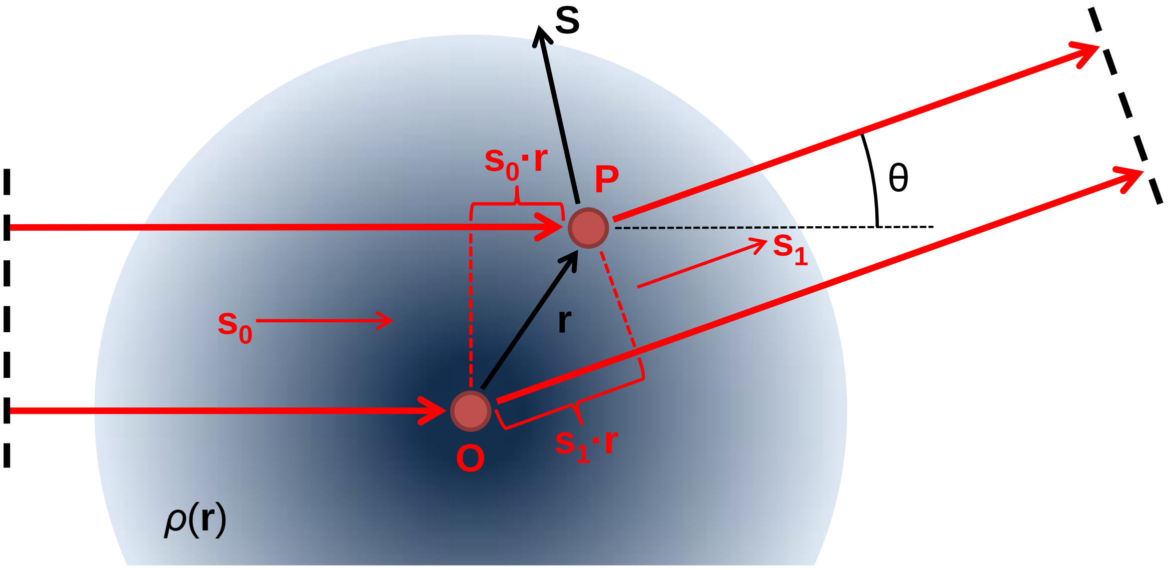


Figure 3.5: X-ray scattering by electrons in an atom. The figure shows two partial waves being scattered by two volume elements, O (at the origin) and P in the electron density, $\rho(r)$, of an atom. The path difference between the two partial waves is given by the difference between $\mathbf{s}_1 \cdot \mathbf{r}$ and $\mathbf{s}_0 \cdot \mathbf{r}$. This translates to a phase difference between the two waves of $\Delta p = 2\pi \mathbf{S} \cdot \mathbf{r}\lambda$. Figure adapted from [181].

Depending on the positions of the volume elements within the electron density of the atom, constructive or destructive interference can occur between the partial waves. The scattering of the whole atom can then be described by the atomic scattering factor, $f_{\mathbf{S}}$, which integrates over all of the partial waves emanating from the entire atom volume (V):

$$f_{\mathbf{S}} = \int_{\mathbf{r}}^{V} \rho(\mathbf{r}) \exp(2\pi i \ \mathbf{S}\mathbf{r}) \ d\mathbf{r}. \tag{3.1.6}$$

This is the Fourier transform of the electron density of the atom. For macromolecular crystallographic programs, the scattering factor uses an empirical approximation of the following form:

$$f_{\mathbf{S}}^{0} = \sum_{i=1}^{4} a_i \cdot \exp\left(\frac{-b_i \left|\mathbf{S}\right|^2}{4}\right) + c = \sum_{i=1}^{4} a_i \cdot \exp\left(-b_i \left(\frac{\sin\theta}{\lambda}\right)^2\right) + c, \tag{3.1.7}$$

which is denoted f_s^0 to indicate that it is wavelength independent. The a_i, b_i and c are the Cromer Mann coefficients for each atom, which are listed in the International Tables for Crystallography [194].

Scattering of X-rays by a Molecule

In a molecule, the partial waves scattered by neighbouring atoms interfere, leading to a modulating scattering function. This scattering function is a superposition of all atomic scattering factors $f^0_{\mathbf{S}j}$:

$$\mathbf{F_S} = \sum_{j=1}^{\text{atoms}} f^0_{\mathbf{S},j} \cdot \exp(i\phi) \tag{3.1.8}$$

where $\phi = 2\pi\mathbf{S}\mathbf{r}_j$ is the relative phase of each contribution. In a single protein molecule, as there is no periodicity, this function is a continuous decaying complex function with an irregularly structured modulation to it. In practice, the scattering off individual molecules is too small to detect, whereas the arrangement of many molecules in a crystallographic lattice allows amplification to occur through constructive interference [181]. The amplification factor is proportional to the number of repeats in the crystal.

Scattering of X-rays by a Crystal

A simple interpretation of X-ray diffraction in a crystallographic lattice introduced by Sir William Lawrence Bragg, treats the X-ray diffraction as reflection on the lattice plane of a crystal. Using this interpretation, the Bragg equation:

$$n\lambda = 2d_{hkl} \sin\theta \tag{3.1.9}$$

relates the scattering angle, θ, to the interplanar distance for the set of lattice planes hkl. This equation works on the principal that the path difference between two interfering partial waves is a multiple (n) of the wavelength (λ) of the wave for maximum constructive interference to occur. As a consequence of this, the scattered X-rays only interfere constructively if the unit cells are correctly aligned, thus scattering ceases to occur in directions that do not correspond to integer hkl values.

Looking again at the scattering factor for unit cell direction **a** only for simplicity, assuming that there are u repeats of the unit cell in that direction, the scattering factor is:

$$\mathbf{F^a_S} = \sum_{u=0}^{\text{u-1}} \mathbf{F}^{\text{cell}}_{\mathbf{S}} \cdot \exp(2\pi i \ u\mathbf{Sa}) \tag{3.1.10}$$

where $\mathbf{F}^{\text{cell}}_{\mathbf{S}}$ is the scattering factor for all of the atoms in the unit cell. The phase difference between unit cells is $2\pi\mathbf{Sa}$, therefore summing over all the unit cells, the scattered wave becomes extinct through destructive interference when $\mathbf{Sa}$ is not an integer. As a consequence of this, no scattering occurs for none integer values of $\mathbf{Sa}$, and scattering peaks sharpen around integer values. Thus, these integer values correspond to the h indices from the reciprocal lattice. At these peaks, the scattering function for the $\mathbf{a}$ direction becomes $F_{\mathbf{S}}^{\mathbf{a}} = u \cdot \mathbf{F}_{\mathbf{S}}^{\mathrm{cell}}$.

Extending this to the three unit cell directions, the scattering function at the peaks for N unit cell repeats is $F_{\mathbf{S}}^{\mathrm{cryst}} = N \cdot \mathbf{F}_{\mathbf{S}}^{\mathrm{cell}}$ Therefore, the expression for molecular scattering can be used to describe crystal scattering, summing over all of the atoms in the unit cell rather than the atoms in the molecule:

$$\mathbf{F}_{\mathbf{S}} = \sum_{j=1}^{\mathrm{atoms}} f_{\mathbf{S},j}^{0} \cdot \exp(2\pi i \; \mathbf{Sr}_j) \tag{3.1.11}$$

This equation can be made general by first expressing the real space lattice vector $\mathbf{r}_j$ for each atom, j in terms of fractional coordinates:

$$\mathbf{r}_j = \mathbf{A} \cdot \mathbf{x}_j = (\mathbf{a}x_j + \mathbf{b}y_j + \mathbf{c}z_j) \tag{3.1.12}$$

Therefore by applying the relationships $\mathbf{S} \cdot \mathbf{a} = h, \mathbf{S} \cdot \mathbf{b} = k$ and $\mathbf{S} \cdot \mathbf{c} = l$ (the Laue equations), it follows that:

$$\mathbf{Sr}_j = \mathbf{Sa}x_j + \mathbf{Sb}y_j + \mathbf{Sc}z_j = hx_j + ky_j + lz_j = \mathbf{hx}_j \tag{3.1.13}$$

The total scattering from the crystal in lattice direction $\mathbf{h}$ or the structure factor of reflection $\mathbf{h}$ is therefore proportional to the sum of all of the scattering contributions in the unit cell:

$$\mathbf{F}_{\mathbf{h}} = \sum_{j=1}^{\mathrm{atoms}} f_{\mathbf{S},j}^{0} \cdot \exp(2\pi i \; \mathbf{hx}_j) \tag{3.1.14}$$

It is therefore possible to compute the scattering function for the diffracted X-ray corresponding to the reflection in the lattice planes with reciprocal Miller index hkl. When collecting the diffraction data, the structure factor amplitude, $F_{\mathbf{h}}$, of a diffraction peak in direction $\mathbf{h}$ is reconstructed from the intensity of the reflection, $I_{\mathbf{h}}$ by:

$$F_{\mathbf{h}} = k \cdot k' \cdot I_{\mathbf{h}}^{1/2}. \tag{3.1.15}$$

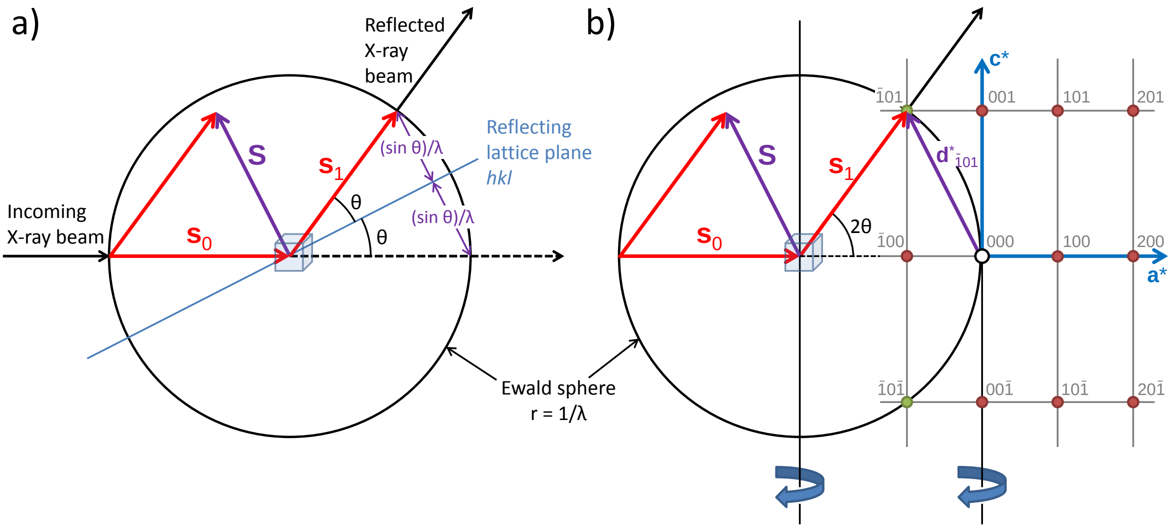


Figure 3.6: Constructing the Ewald sphere and the significance of the reciprocal lattice on diffraction conditions. a) Sketching a scattering event as a reflection on a set of lattice planes and encircling with an Ewald sphere of radius $1/\lambda$, it can be seen that $\mathbf{S} = (2\sin\theta)/\lambda$. b) Drawing the reciprocal lattice with origin displayed as an unfilled circle (∘) it is possible to see that for diffraction to occur for lattice plane hkl, the condition $\mathbf{S} = \mathbf{d}^*_{hkl}$ needs to be met. For this crystal orientation, the diffraction conditions are met for the $\bar{1}01$ and $\bar{1}0\bar{1}$ lattice planes. Rotation of the crystal and therefore the simultaneous rotation of the reciprocal lattice, changes which lattice points intersect the Ewald sphere. Adapted from [181].

where k and k' are correction terms.

A convenient way to visualise the geometric conditions necessary for X-ray diffraction to occur is an Ewald sphere (see figure 3.6). This method takes advantage of the reciprocal lattice. Each set of lattice planes in real space with Miller indices hkl has a reciprocal lattice vector $\mathbf{d}^*_{hkl}$, where the length of $\mathbf{d}^*_{hkl}$, d^*_{hkl} is reciprocal to the interplanar distance in the Bravais lattice, d_{hkl}. It is then possible to rewrite the Bragg equation so that:

$$d^*_{hkl} = \frac{1}{d_{hkl}} = \frac{2\sin\theta}{\lambda} \tag{3.1.16}$$

for the case where $n = 1$. Placing a scattering diagram into an Ewald sphere with radius $1/\lambda$ with the origin of the reciprocal lattice at the point where the incoming beam direction exits the sphere shows the connection between the reciprocal lattice and the diffraction direction. Any reciprocal lattice point that lies on the Ewald sphere fulfils the diffraction condition $d^*_{hkl} = (2\sin\theta)/\lambda$ and it also follows that $\mathbf{S}_{hkl} = \mathbf{d}^*_{hkl}$. The symmetry of the diffraction pattern reflects the symmetry of the crystal structure.

By closer inspection of figure 3.6, a shorter wavelength leads to a larger Ewald sphere. This corresponds to there being more reflections at lower angles, which can be useful as the size of the detector is finite.

Debeye Waller Factor and B-factors

The structure factor described in the previous section assumes that the atom positions are fixed. However, in reality atoms vibrate around a mean position and the average positions of atoms can be different in different unit cells. These two effects are generally indistinguishable from a single data set in protein crystallography, but can be described as an additional Gaussian, wavelength dependent term called the Debye-Waller factor:

$$T_s = \exp\left(-B_{\text{iso}}\left(\frac{\sin\theta}{\lambda}\right)^2\right) \tag{3.1.17}$$

Where B_{iso} is the isotropic B-factor, which is related to the mean square displacement of the atom from its mean position, $\langle q^2 \rangle$, by:

$$B_{\text{iso}} = 8\pi^2 \left\langle q^2 \right\rangle \tag{3.1.18}$$

For proteins, the B-factors of individual atoms depend on the static disorder of the atom, its dynamic behaviour and the overall dynamic behaviour of the protein. Residues in an inflexible region of the protein will have much lower B-factors than residues in flexible regions of the protein. These flexible regions do not diffract as well as the inflexible regions.

3.1.6 Phase Problem

Recall from 3.1.5 that the complex structure factor is the Fourier transform of the electron density. Fourier transforms (FT) are reversible, so it follows that:

$$\text{FT}^{-1}\left[\text{FT}\left(\rho(\mathbf{r})\right)\right] = \rho(\mathbf{r}) \tag{3.1.19}$$

Therefore performing a Fourier transform on the complex structure factor in reciprocal space, returns the electron density in real space:

$$\rho(\mathbf{x}) = \frac{1}{V}\sum_{\mathbf{h}=-\infty}^{+\infty} \mathbf{F}(\mathbf{h}) \cdot \exp(2\pi i \mathbf{h}\mathbf{x}). \tag{3.1.20}$$

In this instance, the Fourier transform is expressed as a discrete sum as there are finite reflections measured. Therefore, if the structure factors are known for each reflection it is possible to calculate the electron densities and hence the structure of the protein.

Unfortunately as only the intensity, $I_{\mathbf{h}}$, of the reflections is known, with $I_{\mathbf{h}} \propto F_{\mathbf{h}}^2$, the magnitude of the structure factor. The phase information from the structure factor is missing which contains most of the structural information. By using the relationship:

$$\mathbf{F_h} = F_{\mathbf{h}} \cdot \exp(i\alpha_{\mathbf{h}}), \tag{3.1.21}$$

where $\alpha_{\mathbf{h}}$ is the phase of the Fourier transform, it is possible to reformulate equation 3.1.20 as:

$$\begin{aligned} \rho(\mathbf{x}) &= \frac{1}{V}\sum_{\mathbf{h}=-\infty}^{+\infty} F(\mathbf{h}) \exp(2\pi i\mathbf{hx}) \exp(i\alpha_{\mathbf{h}}) && (3.1.22) \\ &= \frac{1}{V}\sum_{\mathbf{h}=-\infty}^{+\infty} F(\mathbf{h})\ \exp(2\pi i\mathbf{hx} + i\alpha_{\mathbf{h}}) && (3.1.23) \end{aligned}$$

The problem is now how to obtain the phases $\alpha_{\mathbf{h}}$.

Until 1934 and the introduction of the Patterson function [195], the only way to solve the phase problem was by trial and error. The Patterson function $P(\mathbf{u})$ is defined at any point $\mathbf{u} = u, v, w$ by:

$$P(\mathbf{u}) = \int_R \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})d\mathbf{r}, \tag{3.1.24}$$

where the integral is over the whole unit cell in real space (R). The Patterson function has a large value at $\mathbf{u}$ when both $\rho(\mathbf{r})$ and $\rho(\mathbf{r}+\mathbf{u})$ have large values. Therefore the Patterson function has maxima where the distance vector $\mathbf{u}$ corresponds to interatomic distances.

To obtain the Patterson function for a particular crystal, a Fourier transform is performed on the reflection intensities $I_{\mathbf{h}}$, or $F_{\mathbf{h}}$:

$$P(\mathbf{u}) = \sum_{\mathbf{h}} F_{\mathbf{h}}^2 \exp(-2\pi i\mathbf{hu}) \tag{3.1.25}$$

For a relatively small number of atoms, the position of the atoms can be determined from the maxima in the Patterson function. For larger molecules such as proteins, however, this soon becomes an impossible task due to there being $N(N-1)$ peaks in the Patterson function and many of the Patterson peaks are unresolvable from each other; even at high resolutions.

Regardless, Patterson functions are still used in certain steps of a number of number of methods, such as multiple isomorphous replacement (MIR), single wavelength anomalous dispersion (SAD), multiple wavelength anomalous dispersion (MAD) or a combination of these methods. In these methods it is used in conjunction with "direct methods" to locate the position of the heavy atoms. The technique used in this thesis, however is molecular replacement, so only this method will be described in greater detail.

Molecular Replacement

Molecular replacement is a technique that uses a model that is structurally similar to a molecule in the target crystal structure to try solve the phase problem. This idea behind this technique is that it places the model structure in the correct place and orientation within the unknown unit cell. This makes the problem a 6-dimensional problem, with 3 rotation dimensions and 3 translation dimensions.

The solution to this problem can be solved by looking closer at the Patterson function. The peaks on the Patterson function can be split into intramolecular peaks and intermolecular peaks. As the intramolecular peaks are not affected by the position of the molecules within the unit cell, just their orientation, the Patterson function for the model molecules can be rotationally fit to the experimental function before a translational fit is used to determine the positions of the molecules within the unit cell. In general, the peaks of the Patterson function closest to the origin are more likely to correspond to the intramolecular vectors, so the rotation function can be fitted with these peaks only.

In this thesis, the maximum likelihood rotational and translational functions are used to fit the molecular model to the experimental data [196,197]. Maximum likelihood methods works on the principal that the best model is the one that is most consistent with the data and is defined as the probability that the data would be observed given the model (p(data;model)). In crystallography, the data are the individual reflection intensities, which are assumed to be independent. In this case, the joint probability of the data is the product of the individual probability distributions. To avoid numerical problems of working with extensive data sets, the log of the likelihood is used, which is optimised at the same set of parameters as the likelihood:

$$\mathrm{LL}\left(\text{model};\{\text{data}_i\}\right) = \sum_i \ln\left[p(\text{data}_i;\text{model})\right] \tag{3.1.26}$$

An overview of the algorithms used in the maximum likelihood method for molecular replacement can be found at [196].

3.1.7 Model Building and Refinement

Model building is the construction of an atomic structure model of the macromolecule that fits into the electron density. This process is alternated with global restrained reciprocal space refinement of the atomistic positions and B-factors of the model.

The model building step is performed in real space using computational visualisation of electron density maps to fit the model. There are two types of electron density maps frequently used in model building, the combined electron density map $(2F_{\text{obs}} - F_{\text{calc}})\exp(i\phi_{\text{calc}})$ and the basic difference map $(F_{\text{obs}} - F_{\text{calc}})\exp(i\phi_{\text{calc}})$. The $(2F_{\text{obs}} - F_{\text{calc}})exp(i\phi_{\text{calc}})$ map is used as it displays regions of electron density where atoms are missing from the model at normal density levels. $F_{\text{obs}}\exp(i\phi_{\text{calc}})$ density maps, on the other hand only recreate half (or less) of the density for unmodelled regions. The combined map is therefore preferred for model building, whereas the difference map is preferred for model correction as it displays regions where the model is incorrect.

Models originating from molecular replacement, as is the case in this thesis have a number of challenges to overcome during the refinement stage of solving the crystal structure. The initial phases used to determine the electron density originate solely from the model, so can be incorrect or even missing in regions. This causes the model phases to be biased, leading to the electron density maps to reconstruct the model's electron density. The model will also be incomplete and contain large errors, which need to be accounted for when building the model. To minimise the model bias and take into account other errors, maps constructed from the Fourier coefficients $(2mF_{\text{obs}} - DF_{\text{calc}}) \cdot \exp(i\phi_{\text{calc}})$ for acentric reflections and $mF_{\text{obs}} \cdot \exp(i\phi_{\text{calc}})$ for centric reflections are used [198]. The coefficient m is the figure of merit, which takes into account

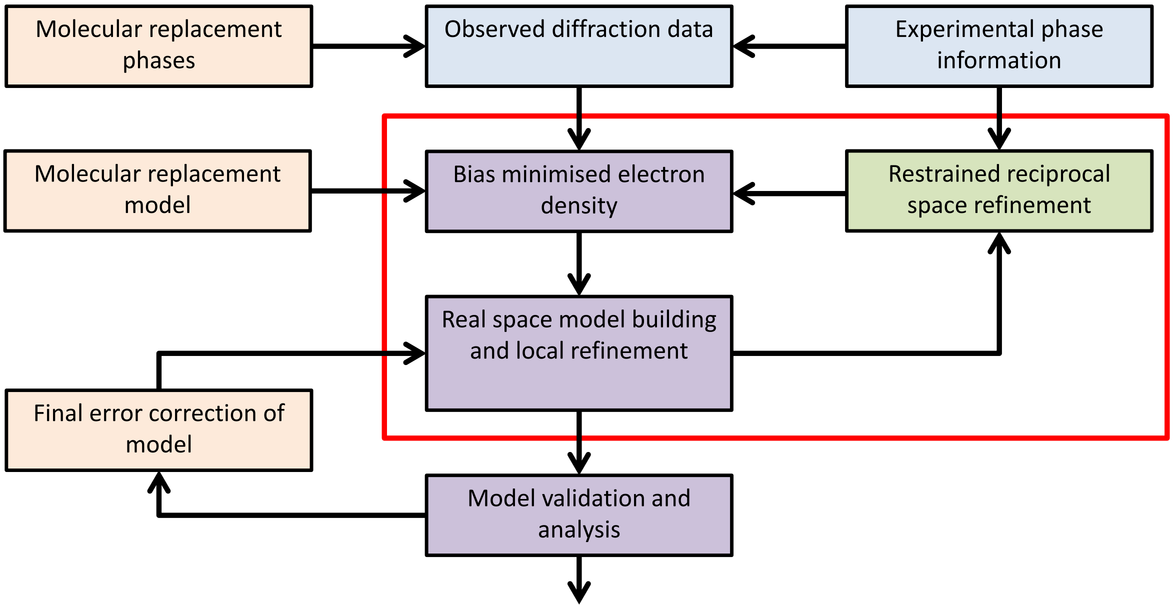

Molecular replacement phases
Observed diffraction data
Experimental phase information
Molecular replacement model
Bias minimised electron density
Restrained reciprocal space refinement
Real space model building and local refinement
Final error correction of model
Model validation and analysis

Figure 3.7: Flow chart showing the model building and refinement process. Initial electron density map is produced from the observed diffraction data and either the molecular replacement phases or experimental phase information. Model building then occurs in real space, which is alternated with reciprocal space refinement which minimises the differences between the observed and the calculated structure factor amplitudes. This cycle is contained within the red box. The reciprocal space refinement is restrained using prior knowledge. Adapted from [181]

the completeness of the model, while D is the Luzzati function, derived from σ_A; an estimate of the error in the coordinates of the model.

While building and modifying the model, other properties of proteins are checked, such as whether there are any disallowed backbone torsions (by observing the Ramachandran plot), unlikely side chain rotamers or improbable bond lengths and angles.

Local geometry errors after real space model building are corrected during the global reciprocal space refinement by optimising the fit between observed and calculated structure factor amplitudes. In this stage, the parameters describing the structural model, such as the atomic coordinates and the B-factors are refined against the experimental data, minimising the difference between the observed structure factor amplitudes and the calculated model structure factor amplitudes using maximum likelihood methods [197, 199]. A flow diagram of the model building and refinement process can be seen in figure 3.7.

The quality of the model in comparison to the diffraction data is quantified by the R-value. This is the global linear residual between the observed and calculated scaled structure factor amplitudes F_{obs} and F_{calc} defined as:

$$R = \frac{\sum_{\mathbf{h}} |F_{\text{obs}} - F_{\text{calc}}|}{\sum_{\mathbf{h}} F_{\text{obs}}}. \tag{3.1.27}$$

During the model building and refinement, a small amount of the data is discarded (typically 5%) and is not included in calculating the R-value. This neglected "free" data is used to calculate a free R-value, R_{free}, which helps determine if the model is being over fit with too many parameters [200]. The corresponding R-value calculated using all data except the free data is denoted R_{work}. A large difference between R_{work} and R_{free} gives a rough indication that too many parameters are being used in the model.

As a result of their often not being enough experimental data being used in the refinement relative to the number of parameters, prior knowledge about proteins is used. This helps prevent nonsensical models caused by over-fitting the data. This prior knowledge can be in the form of chemical information, such as the amino acids found in proteins or structural information such as knowledge about interatomic distances and backbone dihedral angles.

The prior knowledge is imposed upon the model in the form of constraints and restraints. Constraints reduce the number of refinement parameters by exploiting known geometric relationships, such as using rigid models for certain chemical groups, such as phenyl groups. Restraints also improve the data to parameter ratio, this time by acting as additional observations. Variables such as bond lengths and bond angles are restrained to values determined experimentally. The functional form of the standard restraint is harmonic; similar to the bond length and bond angle force field terms seen in Molecular dynamics:

$$Q_b = \sum_{j=1}^{\text{restr.}} \frac{1}{\sigma_{b(j)}^2} (b_{\text{calc}(j)} - b_{\text{i}(j)})^2. \tag{3.1.28}$$

Here $b_{\text{calc}(j)}$ represents the restrained geometric parameter, such as the bond length or bond angle, as calculated from the model and $b_{\text{i}(j)}$ is the ideal value of this geometric parameter. σ_b^2 represents the variance in the ideal value of the geometric parameter. $b_{\text{i}(j)}$ and σ_b^2 are determined experimentally [201, 202] and are typically kept in a dictionary of restraints that is utilised by automated refinement programs [199].

It is also possible to restrain the B-factors during the refinement. The B-factors are related to the displacement undertaken by the atoms, therefore it is not usually possible to distinguish partially occupied static conformations from dynamic fluctuations in atomic positions. It can also be seen that there is a correlation between the fluctuations of adjacent atoms (if one atom is static, the adjacent covalently bonded atoms cannot move around too much) making this a sensible restraint to use. The form of the restraint is again harmonic, restraining the B-factor of a target atom, $B_{\mathrm{target}(n)}$ to the B-factor of an adjacent originating atom $B_{\mathrm{origin}(n)}$:

$$Q_B = \sum_{n=1}^{\mathrm{pairs}} w_{B(n)} (B_{\mathrm{origin}(n)} - B_{\mathrm{target}(n)})^2. \tag{3.1.29}$$

Here $w_{B(n)}$ is the weight of the restraint, which is dependent on the position of the atom in the macromolecule. For example, the B-factors of long side chains on the surface diverge much quicker than those of adjacent backbone atoms, therefore they require stronger weights.

After the reciprocal space refinement round, the electron density map is recalculated using the phases from the new model. The iterative cycling of real space model building and reciprocal space refinement is continued until their is enough confidence in the crystal structure observed.

The crystal structure and B-factors contain a lot of information about the structure and the dynamics of the biomolecule; it is a technique that is unrivalled in that respect. It however does not contain information about the thermodynamics of protein-ligand binding; information that is very useful when studying an allosteric system. For this purpose a technique called ITC was used, a technique that is unrivalled for measuring the thermodynamics of reversible reactions between biomolecules.

3.2 Isothermal Titration Calorimetry

ITC allows the determination of the binding affinity, K_{a}, enthalpy, ΔH, and stoichiometry, n, of a protein-ligand binding event. From these values it is then possible to determine the Gibbs free energy, ΔG and entropy, ΔS of the reaction [203]. It is useful for reactions with multiple binding events such as allostery because different binding models can be used to investigate successive binding events.

Titration calorimetry first described as a method for determining K and ΔH in the

1960s and was originally used to study weak acid-base equilibria and metal ion complexation reactions [204–207]. At this time, the technique was limited by the sensitivity of the instruments to measure K values less than around 10^5 M^{-1}. Calorimetric binding studies were first used to study biological systems [208] in the late 1970s. 10 years later MicroCal created the first commercially available titration calorimeter [209]. The basic principles and design of a modern titration calorimeter have not changed much since these first machines.

In the calorimeter there are two cells, a reference cell containing water and a sample cell containing both typically containing around 250 μl of protein solution (in the case of MicroCal ITC$_{200}$ [210]). These cells are kept at exactly the same temperature, any changes are detected and the cells then returned to the same temperature. These cells are set to the desired experimental temperature, which is typically 25 °C.

A series of 2 μl aliquots of ligand solution are injected into the protein solution. If the ligand binds to the protein, very small heat changes occur, which are detected and measured. In the case of an exothermic reaction, the binding event causes a small increase in temperature of the sample cell, whereas in an endothermic reaction a small decrease in the sample cell temperature occurs. The heat change of the binding event, i, is related to the enthalpy of binding through the relationship [211]

$$q_i = V \cdot \Delta H \cdot \Delta[\mathrm{L}^{\mathrm{B}}]_i. \tag{3.2.30}$$

Here q is the change in heat due to the reaction, V is the volume of the sample cell and ΔH is the enthalpy change of the binding event and $\Delta[\mathrm{L}_i^{\mathrm{B}}]$ is the change in concentration of the bound ligand between the $(i-1)$th and ith injection. The magnitude of the heat change decreases after each consecutive ligand injection as the concentration of free protein reduces until the sample cell contains an excess of ligand to protein binding sites where the protein is saturated. Upon saturation, a small heat change caused by mixing and mechanical factors is present, which is subtracted from the rest of the experiment.

As we saw in chapter 1, the protein-ligand binding equilibrium for a binding site, j can be expressed as the fraction of j that are occupied by ligand:

$$\theta_j = \frac{[\mathrm{L}^{\mathrm{F}}]K_j}{1 + [\mathrm{L}^{\mathrm{F}}]K_j} \tag{3.2.31}$$

where K_j is the affinity constant for binding site j and $[\mathrm{L^F}]$ is the concentration of free ligand. The total ligand concentration can now be expressed as:

$$[\mathrm{L^T}] = [\mathrm{L^F}] + [\mathrm{P^T}]\sum_{j=1}^{k}(n_j\theta_j). \tag{3.2.32}$$

In the simplest case with only one binding site, this allows the concentration of bound ligand to be expressed as:

$$[\mathrm{L^B}] = [\mathrm{P^T}]\cdot\theta = [\mathrm{P^T}]\cdot\frac{[\mathrm{L^F}]K_\mathrm{a}}{1+[\mathrm{L^F}]K_\mathrm{a}}, \tag{3.2.33}$$

so equation 3.2.30 becomes:

$$q_i = V\cdot\Delta H\cdot[\mathrm{P^T}]\cdot\left(\frac{[\mathrm{L^F}]_iK_\mathrm{a}}{1+[\mathrm{L^F}]_iK_\mathrm{a}} - \frac{[\mathrm{L^F}]_{i-1}K_\mathrm{a}}{1+[\mathrm{L^F}]_{i-1}K_\mathrm{a}}\right). \tag{3.2.34}$$

With only the total concentration of ligand ($[\mathrm{L^T}]$) being known in the experiment, this equation needs to be rewritten in terms of the free ligand concentration, which can be determined after each injection from the total protein and ligand concentrations by the equation [203, 212]:

$$[\mathrm{L^F}] = \frac{-\left(1+K_\mathrm{a}\left([\mathrm{M}-\mathrm{L^T}]-[\mathrm{L^T}]\right)\right)+\sqrt{\left(1+K_\mathrm{a}\left([\mathrm{M}-\mathrm{L^T}]-[\mathrm{L^T}]\right)\right)^2+4K_\mathrm{a}[\mathrm{L^T}]}}{2K_\mathrm{a}}. \tag{3.2.35}$$

Modern ITC instruments operate on the heat compensation principal, so the signal measured is the amount of power needed to maintain a constant temperature difference between the sample cell and the reference. This is initially plotted against time, with each injection event causing a peak in the plot. The area under each peak is then integrated to return the heat released or absorbed and plotted against the molar ratio of ligand to protein. The resulting isotherm can be fitted to the binding model to directly return the association constant of binding, K_a, the stoichiometry and the enthalpy. Figure 3.8 shows a schematic of a typical ITC data set and the resulting isotherm for a simple 1:1 binding reaction with only one binding event per protein.

The free energy of ligand binding event i can then be determined from the associ-

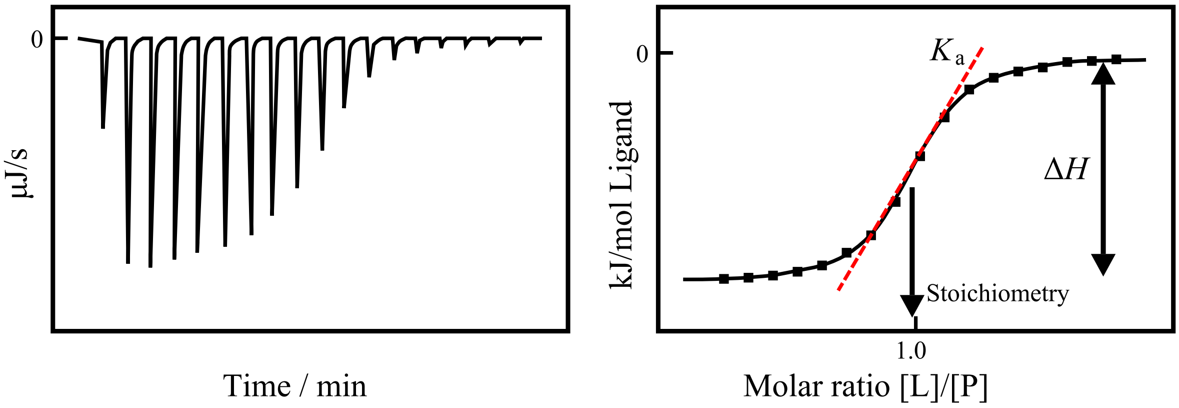


Figure 3.8: A schematic of a typical ITC experiment for an exothermic binding of ligand to single binding site protein. a) Shows the power output by the ITC machine to keep the temperature of the reference and sample cell the same. Each injection of ligand causes a spike in this graph. b) The area under the peaks of a) is plot against the molar ratio. The enthalpy and stoichiometry of the binding are determined by linear values on the graph and K_a is determined by fitting the curve to the binding model [213].

ation constant;

$$\Delta G_i = -RT \ln K_i$$

and the entropy from the difference between the free energy and the enthalpy;

$$\Delta G_i = \Delta H_i - T\Delta S_i.$$

ITC is therefore, an ideal technique to study protein binding events, such as cooperative binding; allowing the investigation of the enthalpic and entropic contributions separately.

Part III

Methods, Results and Discussion

Chapter 4

Coarse-Grained Models of CAP

The earliest examples of allosteric signalling were interpreted as being caused by changes in structure, which gave rise to the classic MWC and KNF (chapter 1) [34,35]. The allosteric signalling in CAP, however has been hypothesised to be driven by dynamics rather than by a structural change [59, 111]. The models discussed in this chapter focus on the modification of the thermal fluctuations upon ligand binding, and investigate how this can contribute to cooperative binding. They are coarse grained; reducing the degrees of freedom of the problem and averaging the many local interactions in the system into simpler CG potentials. This allows for the study of longer timescales and the study of large protein systems much quicker than atomic simulations would allow. With the slow global modes of motion contributing most to dynamic allostery, these CG models may be able to calculate these relevent motions to a sufficient accuracy.

The first models discussed in this chapter are CG models of CAP, which reduce the problem of allosteric signalling to as few parameters as possible. These models are termed SCG models throughout this thesis. They could be described as dynamic analogues to the static conformation based MWC and KNF allostery models [34,35]. These models look at how global properties of the protein, such as the average interactions within a domain or across domains can affect allosteric signalling.

A second model, the elastic network model is briefly discussed, which represents the protein structure as a network of C_α atoms and the interactions between residues as Hookean springs [156, 158]. This model allows mutations to be made by modifying the spring constants of the springs between C_α atoms [172].

4.1 Methods

4.1.1 Super Coarse Grained Models of CAP

Toncrova and McLeish's CG model for the allostery of CAP [54, 55] was motivated by the NMR experiments of Popovych et al. [17] on the N-terminal domain of CAP. Their model, however, did not allow any motion between the two monomers of CAP and they did not extend their model to the full untruncated form of CAP with a DNABD as well as the LBD. This section therefore looks at three new CG models of increasing complexity.

The first models only the LBDs again, this time allowing for motion between the center of masses of the two monomers with an additional degree of freedom describing the motion in the protein as three one-dimensional vector coefficients, x_1, x_2 and x_3. This model will be referred to as the truncated monomer model (TMM) in this thesis.

The second model includes the DNABDs as well as the LBDs, however it does not allow motion between the center of masses of the monomers. The protein motion in this model is described by four motion vector coefficients x_1, x_2, x_3 and x_4, where x_1 and x_2 correspond to motion within the two LBDs and x_3 and x_4 correspond to motion within the two DNABDs. This model will be referred to as the static monomer model (SMM).

The third and final model combines the improvements of both of the previously described models, including the DNABDs and allowing the center of masses of each domain to move. The motion in this model is described by seven one-dimensional vector coefficients x_1 to x_7. This model (first seen in [59] and developed alongside Dr. Tom Rodgers) will be referred to as the mobile monomer model (MMM). These last two models distinguish the LBDs and the DNABDs by labelling them the X-domains and Y-domains respectively. See figure 4.1 for a graphical representation of these models and how the vector coefficient describe the motion in the model.

The TMM discussed above has two spring constants controlling the interactions within the protein dimer; k_x, the internal spring constant of each monomer and k_xx, the coupling spring constant between the two monomers. As with the previous model by Toncrova and McLeish [54, 55], binding of one molecule of cAMP causes scaling of the spring constant, k_x, of the bound monomer by β to become βk_x and the coupling

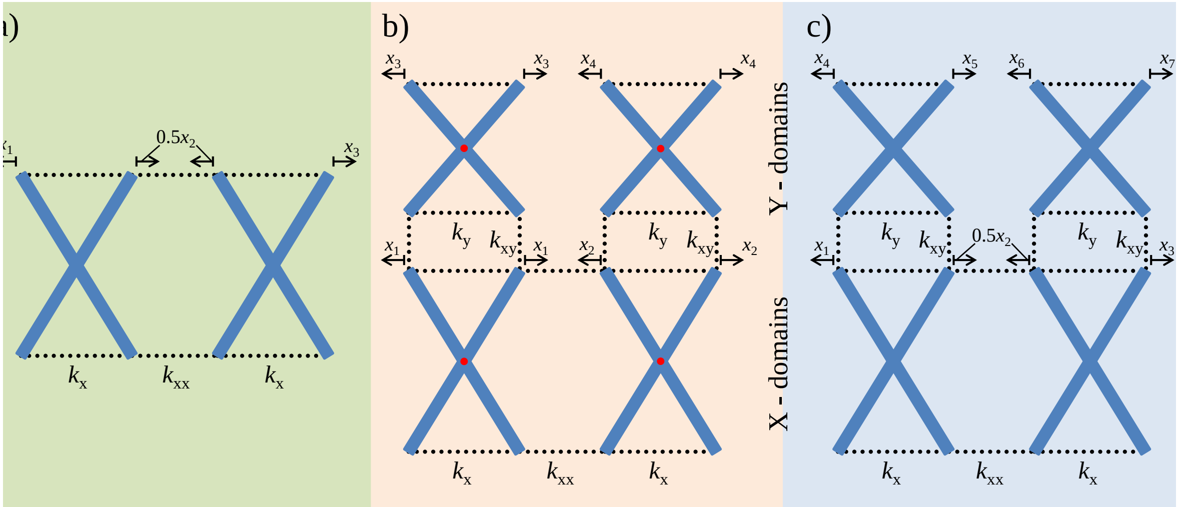


Figure 4.1: Super Coarse Grained models of CAP. (a) Model of the ligand binding (X) domains of CAP, where the center of mass of the monomers are allowed to move. (b) Model of the full length CAP, with the ligand binding (X) domain and the DNA binding (Y) domain. In this model the center of mass of each domain is technically fixed (represented by a red circle in the figure), as each scissor domain opens and closes symmetrically. (c) The final model allows the center of mass of each domain to move by adding three extra degrees of freedom.

spring constant, k_{xx}, to be scaled by α to become αk_{xx}. Binding the second molecule of cAMP to the second monomer again causes the internal spring constant to scale to βk_{x} and the coupling spring constant to scale to $\alpha^2 k_{\mathrm{xx}}$.

Along with the extra domains, the other two new models naturally introduce a couple of extra spring constants controlling the interactions within the homodimer. These are k_{y}, the internal spring constant of the Y domain (DNABD) and k_{xy}, the spring constant between the X domain (LBD) and the Y domain of the same monomer. When binding occurs to a monomer, as well as the changes to spring constants that occur for the X domain only model, the spring constant k_{xy} for the bound monomer is also scaled by γ. These models, as well as the changes in spring constants upon binding cAMP can be seen in figure 4.2.

Two more springs could have also potentially been added to the models, which would have been the spring between the DNABDs of opposite monomers and the spring from the LBD of one monomer to the DNABD of the other monomer. These springs were excluded from the models for a few reasons. Firstly from visual inspection of the holo$_2$-CAP structure, it was decided that these spring constants would be negligible in comparison to the other spring constants in the model. This was later found to be the case when determining the spring constants for WT-CAP using the RTB model as

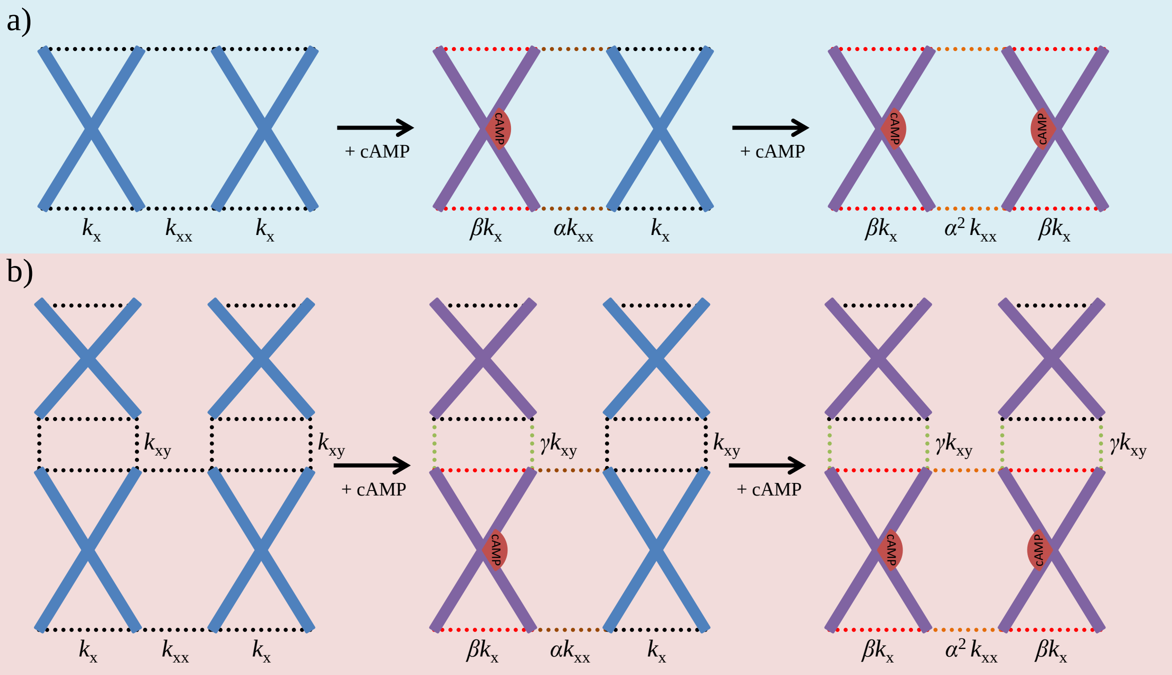


Figure 4.2: cAMP binding in the SCG models of CAP. (a) For the TMM, each binding event scales the internal spring constant of the bound monomer k_x by β and the coupling spring constant k_{xx} by α. (b) In addition to this scaling, the SMM and MMM scale the spring constant k_{xy} of the bound monomer by γ.

described in 4.1.2. And finally, leaving out these spring constants in the calculation of $\Delta\Delta G$ returns a slightly less complex term, with fewer variables, keeping a more manageable analytical solution. These spring constants could easily be reincorporated into the model if other proteins were being modelled, where these interactions were significant enough to include.

The elasticity matrices for each model were determined with the following method. Firstly, the potential energy function of the system was determined by:

$$V(\mathbf{x}) = \frac{1}{2}\sum_{i}^{\text{springs}} k_i \Delta x_i \tag{4.1.1}$$

where Δx_i represents the change from equilibrium length of the spring with constant k_i. Rearrangement of the resulting formula allows it to be written in the matrix format:

$$V(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbb{K} \mathbf{x} \tag{4.1.2}$$

As an example, this process is shown here in greater detail for the SMM. The potential energy of this system is given as:

$$\begin{aligned} V(\mathbf{x}) = {} & k_{\mathrm{x}}(x_1 - (-x_1))^2 + k_{\mathrm{x}}(x_2 - (-x_2))^2 + k_{\mathrm{xx}}(x_1 - (-x_2))^2 \\ & + k_{\mathrm{y}}(x_3 - (-x_3))^2 + k_{\mathrm{y}}(x_4 - (-x_4))^2 + k_{\mathrm{xy}}(x_1 - x_3)^2 \\ & + k_{\mathrm{xy}}(x_2 - x_4)^2 + k_{\mathrm{yy}}(x_3 - (-x_4))^2 \end{aligned} \tag{4.1.3}$$

The $\frac{1}{2}$ disappears in this model because each spring has an equivalent spring that is displaced by the same amount at the other end of the scissor. Rearranging 4.1.3 gives:

$$\begin{aligned} V(\mathbf{x}) = {} & x_1^2(4k_{\mathrm{x}} + k_{\mathrm{xx}} + k_{\mathrm{xy}}) + x_2^2(4k_{\mathrm{x}} + k_{\mathrm{xx}} + k_{\mathrm{xy}}) \\ & + x_3^2(4k_{\mathrm{y}} + k_{\mathrm{xy}}) + x_4^2(4k_{\mathrm{y}} + k_{\mathrm{xy}}) + 2x_1x_2(k_{\mathrm{xx}}) \\ & + 2x_1x_3(-k_{\mathrm{xy}}) + 2x_2x_4(-k_{\mathrm{xy}}) \end{aligned} \tag{4.1.4}$$

which can be written in matrix format as:

$$V(\mathbf{x}) = \mathbf{x}^T \begin{bmatrix} 4k_{\mathrm{x}} + k_{\mathrm{xx}} + k_{\mathrm{xy}} & k_{\mathrm{xx}} & -k_{\mathrm{xy}} & 0 \\ k_{\mathrm{xx}} & 4k_{\mathrm{x}} + k_{\mathrm{xx}} + k_{\mathrm{xy}} & 0 & -k_{\mathrm{xy}} \\ -k_{\mathrm{xy}} & 0 & 4k_{\mathrm{y}} + k_{\mathrm{xy}} & 0 \\ 0 & -k_{\mathrm{xy}} & 0 & 4k_{\mathrm{y}} + k_{\mathrm{xy}} \end{bmatrix} \mathbf{x} \tag{4.1.5}$$

Hence, the elasticity matrix can be determined. The same process was used to determine the elasticity matrices of the other models in this thesis.

The elasticity matrices were then evaluated in Maple, using equation 2.4.54 in section 2.4, allowing $\Delta\Delta G_{\mathrm{conf}}$ to be calculated for each model. As there are few parameters for each model, the parameter space of the models can be explored to see under what conditions cooperative binding occurs. Plots of the models' free energy surfaces were then created using the python module matplotlib.

4.1.2 Rotational Translational Block (RTB) Model

The RTB model splits the protein into rigid blocks that undergo rotational and translational motion (as described in section 2.2.5). The RTB model analysis was performed by Dr. Tom Rodgers and has previously been published in [59].

The blocks were determined using the hingefind plugin of VMD [214] on the lowest energy modes determined by principal component analysis from MD simulations of WT

CAP and corresponded to one block for each LBD and one block for each DNABD. See chapter 5 for MD protocol and results.

The resulting matrix from the RTB reduction contained 3 dimensional spring constants between blocks, which were averaged to create one dimensional spring constants for the SCG models. These spring constants determine the position of WT-CAP on the free energy surfaces of the SCG models and are; $k_{\mathrm{x}} = 57.36$, $k_{\mathrm{xx}} = 113.38$, $k_{\mathrm{xy}} = 21.73$, $k_{\mathrm{y}} = 16.66$ kJ mol^{-1} Å^{2}, $\alpha = 1.30$, $\beta = 0.56$ and $\gamma = 0.90$.

4.1.3 Elastic Network Model

Elastic network simulations were performed using in-house code, called Durham dynamic protein toolbox (DDPT), written by Dr Tom Rodgers in Fortran. The ENM calculations and analysis in this section were performed by Dr. Tom Rodgers and were first published in [59].

The nodes of the ENM were all C_{α} atoms, which were connected with springs if their inter-node distance was within a cutoff radius of 8 Å (see figure 4.3). CAMP was included in the model where appropriate by representing each heavy atom as an additional node in the model. For WT-CAP, all spring constants were set to a constant value of 1 kcal mol^{-1} Å^{-2} (4.19 kJ mol^{-1} Å^{-2}). Point mutations in the protein were modelled by adjusting the spring constants for all of the springs attached to the chosen mutated residue.

The allosteric free energy was calculated by summing the first n none zero normal modes, where n was determined by examining where the value of $K_{\mathrm{a}2}/K_{\mathrm{a}1}$ converged. To construct the ENM for CAP that is shown in this thesis, the PDB entry with the identification number 4HZF was used. The ENM calculations and analysis were performed by Tom Rodgers [59].

4.2 Results and Discussion

4.2.1 Super Coarse Graining CAP

Three models were discussed in section 4.1.1. They will all be evaluated and discussed in this section. Initially, the TMM will be investigated, by first working through the model to investigate which regions of parameter space exhibit cooperative binding

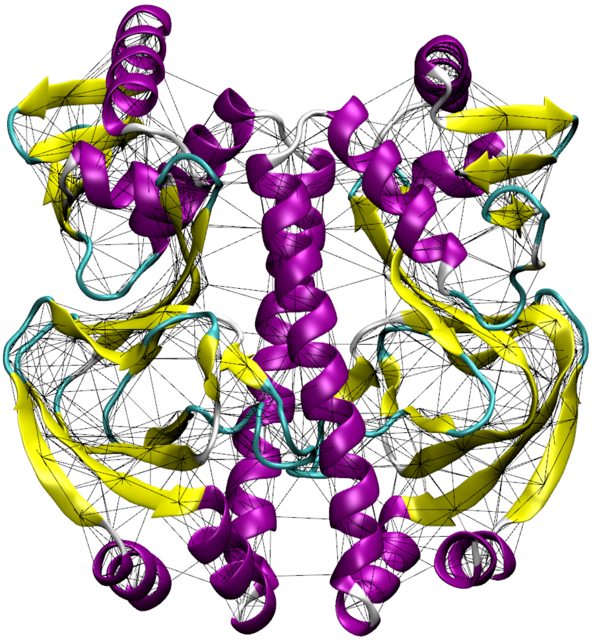


Figure 4.3: Network of springs for the elastic network of apo-CAP superimposed on the cartoon representation of CAP. All C_α atoms within 8 Å of each other were attached by springs, represented as black lines. The PDB code for the structure used is 1G6N [59, 107].

and explore what this could imply to allostery in protein homodimers. The region of parameter space that WT-CAP exists in (as fitted by the RTB model) will then be explored in greater detail. Next the SMM and MMM will be investigated. However, due to the added complication of added parameters, it becomes infeasible to investigate the entire parameter space as is done for the TMM, so only the parameter space surrounding the position of WT-CAP is studied in detail.

Truncated Monomer Model

The elasticity matrix for apo-CAP for the TMM is:

$$\mathbb{K}_0 = \begin{bmatrix} 2k_\mathrm{x} + k_\mathrm{xx} & k_\mathrm{x} & k_\mathrm{xx} \\ k_\mathrm{x} & k_\mathrm{x} + k_\mathrm{xx} & k_\mathrm{x} \\ k_\mathrm{xx} & k_\mathrm{x} & 2k_\mathrm{x} + k_\mathrm{xx} \end{bmatrix} \tag{4.2.6}$$

The free energy for apo-CAP is then:

$$G_\mathrm{apo} = \frac{1}{2}k_\mathrm{B}T\ln|\mathbb{K}_0| + C = \frac{1}{2}k_\mathrm{B}T\ln(8k_\mathrm{x}^2k_\mathrm{xx} + 4k_\mathrm{x}k_\mathrm{xx}^2) + C \tag{4.2.7}$$

Where C is a constant, which is the same for apo-CAP, holo$_1$-CAP and holo$_2$-CAP, assuming that there were no additional effects on binding. The free energy of holo$_1$-CAP and holo$_2$-CAP are:

$$G_{\text{holo}_1} = \frac{1}{2}k_{\text{B}}T \ln|\mathbb{K}_1| + C = \frac{1}{2}k_{\text{B}}T \ln(8\beta k_{\text{x}}^2\alpha k_{\text{xx}} + 2k_{\text{x}}\alpha^2 k_{\text{xx}}^2 + 2\beta k_{\text{x}}\alpha^2 k_{\text{xx}}^2) + C \quad (4.2.8)$$

$$G_{\text{holo}_2} = \frac{1}{2}k_{\text{B}}T \ln|\mathbb{K}_2| + C = \frac{1}{2}k_{\text{B}}T \ln(8\beta^2 k_{\text{x}}^2\alpha^2 k_{\text{xx}} + 4\beta k_{\text{x}}\alpha^4 k_{\text{xx}}^2) + C \quad (4.2.9)$$

This gives an allosteric free energy of:

$$\Delta\Delta G = \frac{1}{2}k_{\text{B}}T \ln\left(\frac{(8k_{\text{x}}^2 k_{\text{xx}} + 4k_{\text{x}}k_{\text{xx}}^2)(8\beta^2 k_{\text{x}}^2\alpha^2 k_{\text{xx}} + 4\beta k_{\text{x}}\alpha^4 k_{\text{xx}}^2)}{(8\beta k_{\text{x}}^2\alpha k_{\text{xx}} + 2k_{\text{x}}\alpha^2 k_{\text{xx}}^2 + 2\beta k_{\text{x}}\alpha^2 k_{\text{xx}}^2)^2}\right). \quad (4.2.10)$$

As there are only 4 parameters in this model, it becomes possible to explore what parameter combinations lead to positive or negative cooperativity. The model is easier to explore by looking at the term inside the logarithm for $\Delta\Delta G$, the cooperativity coefficient:

$$c = \frac{|\mathbb{K}_0||\mathbb{K}_2|}{|\mathbb{K}_1|^2} = 1 - \frac{|\mathbb{K}_1|^2 - |\mathbb{K}_0||\mathbb{K}_2|}{|\mathbb{K}_1|^2}, \quad (4.2.11)$$

where positive cooperativity is observed when $c > 1$, negative cooperativity is observed when $c < 1$ and no cooperativity is observed when $c = 1$. Of course $|\mathbb{K}_1|^2$ is always greater than zero for all real systems, so it can be seen that:

$$\begin{aligned} &|\mathbb{K}_1|^2 - |\mathbb{K}_0||\mathbb{K}_2| > 0 \quad \text{Returns positive cooperativity} \\ &|\mathbb{K}_1|^2 - |\mathbb{K}_0||\mathbb{K}_2| < 0 \quad \text{Returns negative cooperativity} \\ &|\mathbb{K}_1|^2 - |\mathbb{K}_0||\mathbb{K}_2| = 0 \quad \text{Returns no cooperativity} \end{aligned} \quad (4.2.12)$$

Therefore the term, $\Xi = |\mathbb{K}_1|^2 - |\mathbb{K}_0||\mathbb{K}_2|$, gives information about the sign of cooperative binding within parameter space; however information about the magnitude of the cooperativity is lost. For this model:

$$\begin{aligned}
\Xi &= (8\beta k_{\mathrm{x}}^2\alpha k_{\mathrm{xx}} + 2k_{\mathrm{x}}\alpha^2 k_{\mathrm{xx}}^2 + 2\beta k_{\mathrm{x}}\alpha^2 k_{\mathrm{xx}}^2)^2 \\
&\quad -(8k_{\mathrm{x}}^2 k_{\mathrm{xx}} + 4k_{\mathrm{x}}k_{\mathrm{xx}}^2)(8\beta^2 k_{\mathrm{x}}^2\alpha^2 k_{\mathrm{xx}} + 4\beta k_{\mathrm{x}}\alpha^4 k_{\mathrm{xx}}^2) && (4.2.13)\\
&= 4\alpha^2 k_{\mathrm{x}}^2 k_{\mathrm{xx}}^3(-2\alpha^2\beta k_{\mathrm{xx}} + \alpha^2\beta^2 k_{\mathrm{xx}} + \alpha^2 k_{\mathrm{xx}} \\
&\quad +8\alpha\beta k_{\mathrm{x}} + 8\alpha\beta^2 k_{\mathrm{x}} - 8\alpha^2\beta k_{\mathrm{x}} - 8\beta^2 k_{\mathrm{x}}) && (4.2.14)\\
&= 4\alpha^2 k_{\mathrm{x}}^2 k_{\mathrm{xx}}^3\left(\alpha^2 k_{\mathrm{xx}}(\beta-1)^2 - 8\beta k_{\mathrm{x}}(\alpha-1)(\alpha-\beta)\right). && (4.2.15)
\end{aligned}$$

Here, $4\alpha^2 k_{\mathrm{x}}^2 k_{\mathrm{xx}}^3$ is always positive, so removing this term from Ξ does not affect the sign of the cooperativity determined. It is therefore possible to look at:

$$\widetilde{\Xi} = \alpha^2 k_{\mathrm{xx}}(\beta-1)^2 - 8\beta k_{\mathrm{x}}(\alpha-1)(\alpha-\beta) \qquad (4.2.16)$$

to determine regions of cooperative binding. In this expression, $\alpha^2 k_{\mathrm{xx}}(\beta-1)^2 > 0$ for all α and β, therefore it is evident that positive cooperativity is observed when:

$$\alpha^2 k_{\mathrm{x}}(\beta-1)^2 > 8\beta k_{\mathrm{x}}(\alpha-1)(\alpha-\beta),$$

negative cooperativity is observed when:

$$\alpha^2 k_{\mathrm{xx}}(\beta-1)^2 < 8\beta k_{\mathrm{x}}(\alpha-1)(\alpha-\beta),$$

and of course non-cooperative binding is observed when:

$$\alpha^2 k_{\mathrm{xx}}(\beta-1)^2 = 8\beta k_{\mathrm{x}}(\alpha-1)(\alpha-\beta). \qquad (4.2.17)$$

By taking into account that all of the constants are greater than or equal to zero for all real systems, a few simple observations about the cooperativity of the system can be made. It is evident that when $\alpha = 1$, the right hand term from equation 4.2.17 disappears, so only positive cooperativity can be observed. Furthermore, when $\beta = 1$, the left hand term disappears, and the right hand term becomes $8\beta k_{\mathrm{x}}(\alpha-1)^2$, which is always positive. Hence negative allostery is always observed in this instance. If $\alpha = \beta$, the right hand term becomes zero, meaning that positive allostery is observed.

It is also evident that scaling both k_{xx} and k_{x} by the same amount cannot change the

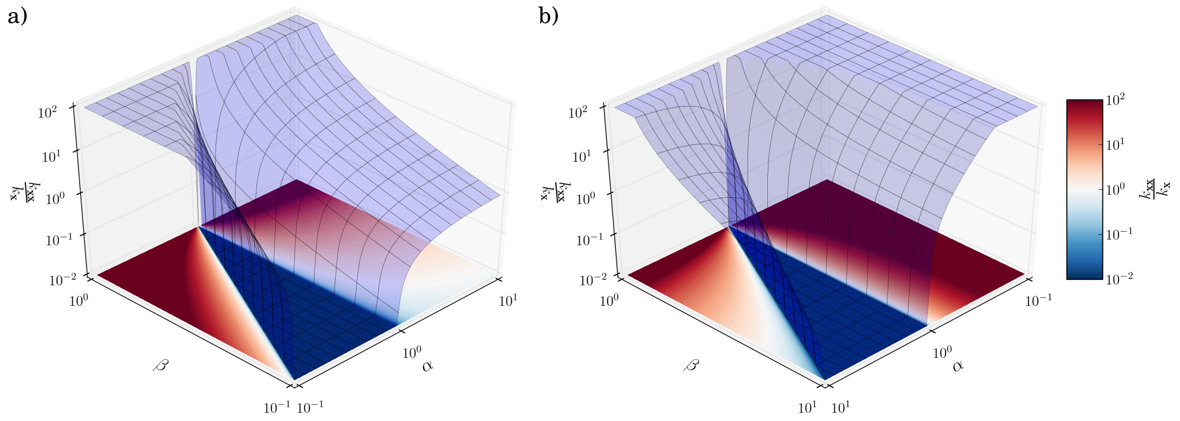


Figure 4.4: Plots of equation 4.2.18; the surface showing the values of $k_{\mathrm{xx}}/k_{\mathrm{x}}$ where no cooperative binding (positive or negative) occurs. The surface is split into two regions, a) $\beta < 1$, the interactions within a monomer decrease upon ligand binding. b) $\beta > 1$, the interactions within a monomer increase upon binding. a) is rotated 180° with respect to b) to improve the viewing angles of both surfaces. Regions of the surface below a value of 10^{-2} or above 10^2 are truncated to these values, as values this large are unlikely to be observed in proteins.

nature of cooperative binding, it likely only scales the magnitude of the cooperative binding. However, the denominator of equation 4.2.11 would have to be taken into account to see the true scaling of cooperativity upon altering k_{xx} and k_{x}. It is therefore possible to rearrange equation 4.2.17 to:

$$\frac{k_{\mathrm{xx}}}{k_{\mathrm{x}}} = \frac{8\,\beta(\alpha-1)(\alpha-\beta)}{\alpha^2(\beta-1)^2} \tag{4.2.18}$$

and plot a surface which shows the values of $k_{\mathrm{xx}}/k_{\mathrm{x}}$ that exhibit no cooperativity for given values of α and β (see figure 4.4). Any values of $k_{\mathrm{xx}}/k_{\mathrm{x}}$ that lie below the surface lead to negative cooperativity for the given values of α and β. The reverse is true for values of $k_{\mathrm{xx}}/k_{\mathrm{x}}$ below the surface. The ratio $k_{\mathrm{xx}}/k_{\mathrm{x}}$ is examined simply to avoid the singularity that appears when $\alpha = \beta$ in the case of $k_{\mathrm{x}}/k_{\mathrm{xx}}$.

A plot like figure 4.4 is useful as it theoretically allows us to see what changes in the interactions of a protein can be made to change its cooperativity. It is possible to examine which regions of the α and β space exhibit cooperative binding for given values of $k_{\mathrm{xx}}/k_{\mathrm{x}}$. For example there are regions of positive cooperativity for any values of k_{xx} and k_{x} when $\beta < \alpha < 1$ and when $1 < \alpha < \beta$.

Perhaps a clearer visual of regions in the global parameter space that exhibit positive or negative cooperativity can be created by taking slices through the plot of the surface

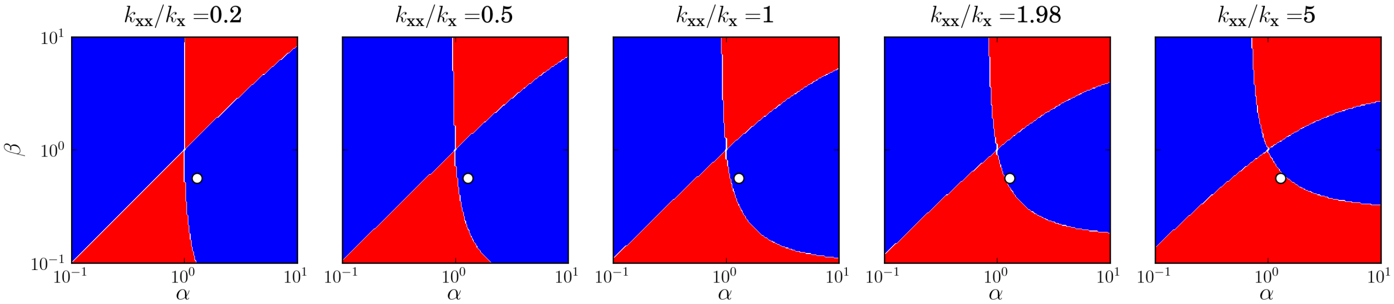


Figure 4.5: Regions of $\alpha - \beta$ space that exhibit cooperative binding for given values of k_{xx}/k_x. Red regions represent regions of positive cooperativity ($k_{xx}/k_x > (8\,\beta(\alpha-1)(\alpha-\beta))/(\alpha^2(\beta-1)^2)$ and blue regions represent regions of negative cooperativity. The position of WT-CAP in α, β space, as calculated for the 4 scissor model is plotted on each diagram as a white circle. The k_{xx}/k_x value of WT-CAP is plotted in the fourth axis ($k_{xx}/k_x = 1.98$), showing WT-CAP to be negatively cooperative.

of zero cooperativity at distinct values of k_{xx}/k_x and colouring all positively cooperative regions red and all negatively cooperative regions blue. This has been done in figure 4.5, while also showing the position of WT-CAP as determined by the RTB model (shown in section 4.1.2). Strictly speaking, the parameters determined for WT-CAP, place it on the slice of the plot where $k_{xx}/k_x = 1.98$, but is shown on all slices to demonstrate the effect of altering k_{xx}/k_x on allostery.

It is evident that in this model, WT-CAP sits in a region of negative cooperativity near the boundary where cAMP binding becomes positively cooperative. In the model, WT-CAP can be pushed into a positively cooperative regime either by decreasing the value of β (or strongly increasing β), decreasing the value of α or increasing k_{xx}/k_x (by increasing k_{xx} or decreasing k_x). As is discussed in 4.2.2 a V132A mutation would probably have the effect of reducing k_{xx}, pushing the system further into negative cooperativity and a V132A mutation would probably increase k_{xx}, which creates a positively cooperative system. This trend is seen in experiment (see 6.3.3).

Figure 4.6 shows the free energy surface in the $k_x - k_{xx}$ parameter space when $\alpha = 1.3$ and $\beta = 0.56$; the values fitted by the RTB model for WT-CAP. Again this shows WT-CAP to be in a position of high sensitivity, where the negative cooperativity of CAP is sensitive to small changes in k_x and k_{xx}.

Of course with all of these models, one should also take into account that in a system as complicated as a protein, adjusting one interaction strength may lead to changes in the strength of the other interactions. In the worst case, playing with these interactions could inadvertently remove a protein's ability to bind a ligand, so caution

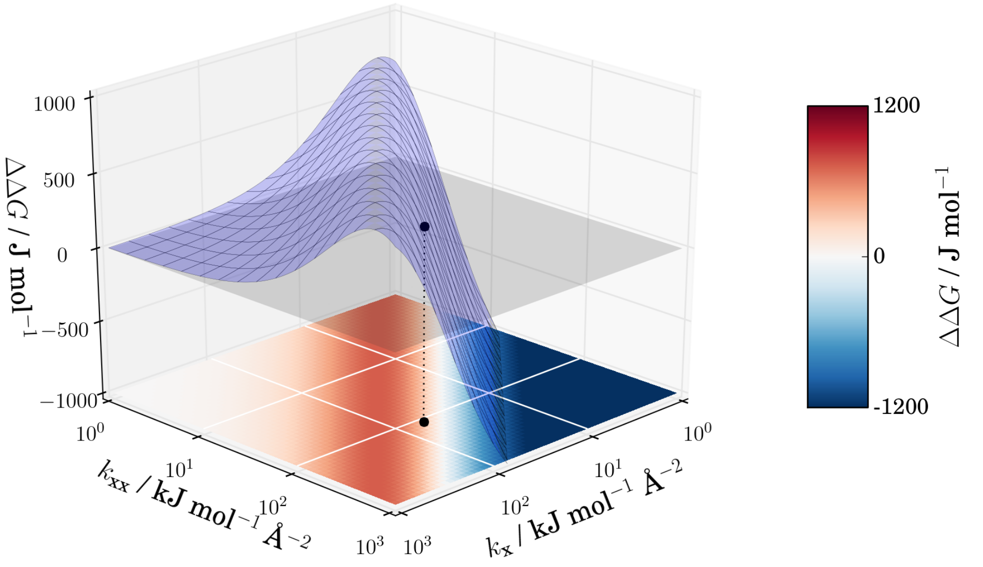


Figure 4.6: Plot of cooperative free energy as a function of the reduced two dimensional parameter space of the internal and coupling spring constants for the TMM. All other parameters for the model are set to WT values.

should be used. Also, the model investigated here is the truncated form of CAP, but as will be discussed later, the region of parameter space surrounding WT-CAP for this model is similar to the more complex full length CAP models.

Static Monomer Model

The elasticity matrix of apo-CAP for the SMM is:

$$\mathbb{K}_0 = \begin{bmatrix} 4k_x + k_{xx} + k_{xy} & k_{xx} & -k_{xy} & 0 \\ k_{xx} & 4k_x + k_{xx} + k_{xy} & 0 & -k_{xy} \\ -k_{xy} & 0 & 4k_y + k_{xy} & 0 \\ 0 & -k_{xy} & 0 & 4k_y + k_{xy} \end{bmatrix} \tag{4.2.19}$$

In this instance, the extra parameters k_{xy} and γ complicate the term for $\Delta\Delta G$, making it unfeasible to perform a similar analysis to that performed for the previous model. However, the free energy surface in the parameter space surrounding WT-CAP can be investigated again (figure 4.7).

In this model, WT-CAP is again shown to be at a negatively cooperative, sensitive point in the parameter space, where increasing k_{xx} or decreasing k_x can push the

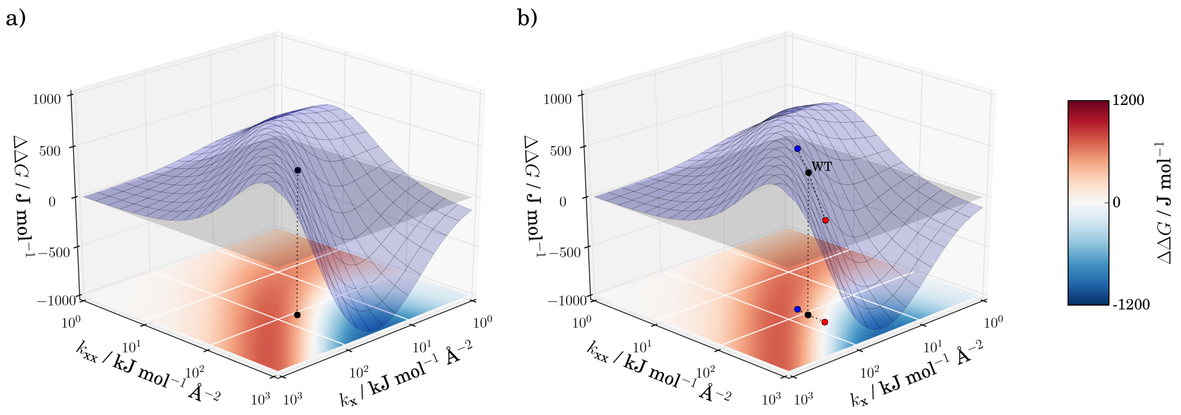


Figure 4.7: Plot of cooperative free energy as a function of the reduced two dimensional parameter space of the internal and coupling spring constants for a) the static monomer model and b) the mobile monomer model. The position of WT-CAP (●) is plotted on both plots and the positions of mutants that increase (●) or decrease (●) the value of k_{xx} are plotted on b). All other parameters for the models are set to WT values.

allostery into positive cooperativity. This free energy landscape and its similarities to the landscapes of the other models is discussed below.

Mobile Monomer Model

The elasticity matrix of apo-CAP for the MMM is:

$$\mathbb{K}_0 = \begin{bmatrix} 2\,k_x + k_{xx} + k_{xy} & k_x & k_{xx} & 0 & -k_{xy} & 0 & 0 \\ k_x & \frac{1}{2}k_x + \frac{1}{2}k_x + k_{xx} + \frac{1}{4}k_{xy} + \frac{1}{4}k_{xy} & k_x & -\frac{1}{2}k_{xy} & 0 & 0 & -\frac{1}{2}k_{xy} \\ k_{xx} & k_x & k_{xx} + 2\,k_x + k_{xy} & 0 & 0 & -k_{xy} & 0 \\ 0 & -\frac{1}{2}k_{xy} & 0 & k_{xy} + 2\,k_y & 2\,k_y & 0 & 0 \\ -k_{xy} & 0 & 0 & 2\,k_y & k_{xy} + 2\,k_y & 0 & 0 \\ 0 & 0 & -k_{xy} & 0 & 0 & k_{xy} + 2\,k_y & 2\,k_y \\ 0 & -\frac{1}{2}k_{xy} & 0 & 0 & 0 & 2\,k_y & k_{xy} + 2\,k_y \end{bmatrix} \qquad (4.2.20)$$

The reduced 2D ($k_x - k_{xx}$) parameter space for this model is plotted in figure 4.7. This shows that the allosteric free energy surface in this region of parameter space for this model is very similar to that of the SMM. This would imply that the simpler SMM is able to model this region of parameter space as well as the MMM. The plot of the MMM allosteric free energy surface shows the positions of WT-CAP, a mutant that increases k_{xx} (the interactions between the two monomers) and a mutant that decreases k_{xx}. A mutation that decreases k_{xx} makes the system more negatively cooperative,

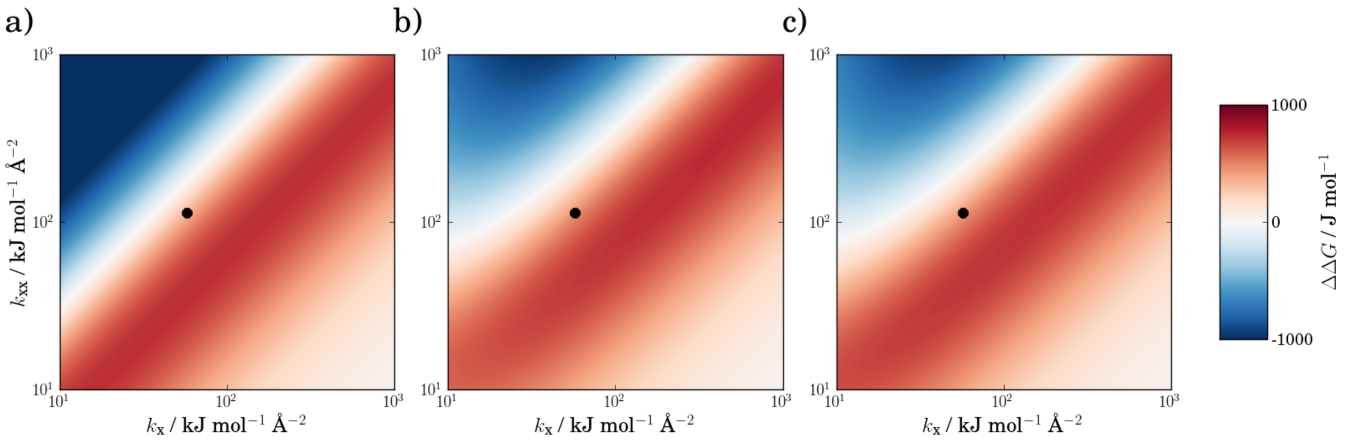


Figure 4.8: Comparison of the two dimensional parameter space for the three SCG models. The other parameters are set to WT values.

whereas a mutation that increases k_{xx} makes the system less negatively cooperative, or positively cooperative if k_{xx} is increased enough. This effect could be recreated experimentally by mutating a residue on the interface between the two monomers (and is done later in this thesis by mutating residue V132).

Super Coarse Graining Discussion

The shape of the free energy surfaces for SMM and MMM are nearly identical around the vicinity of WT-CAP. This would perhaps be expected due to the similarity between the models. What is perhaps surprising is that for the truncated model, the allosteric free energy is actually less (324 J mol^{-1}) than the full length SMM (443 J mol^{-1}) and MMM (418 J mol^{-1}). This is the opposite to what has been demonstrated by ITC experiments, where truncated CAP has been seen to exhibit much stronger negative cooperativity than full-length CAP [17, 59]. The disagreement of the model with experimental data could be due to the fitting of the WT parameters being performed on full-length CAP; it is unlikely that they would be the same for truncated CAP. As has previously been discussed, WT-CAP is in a position on the allosteric landscape that is very sensitive to small changes in some of the parameters. The magnitude of the negative cooperativity can be increased by raising the internal interaction strengths of the monomers (k_x), decreasing the intermonomer coupling strength (k_{xx}) or increasing the gain in interactions between monomers upon binding cAMP (α). With the removal of such a large domain, it is quite likely that there will be significant changes to these parameters.

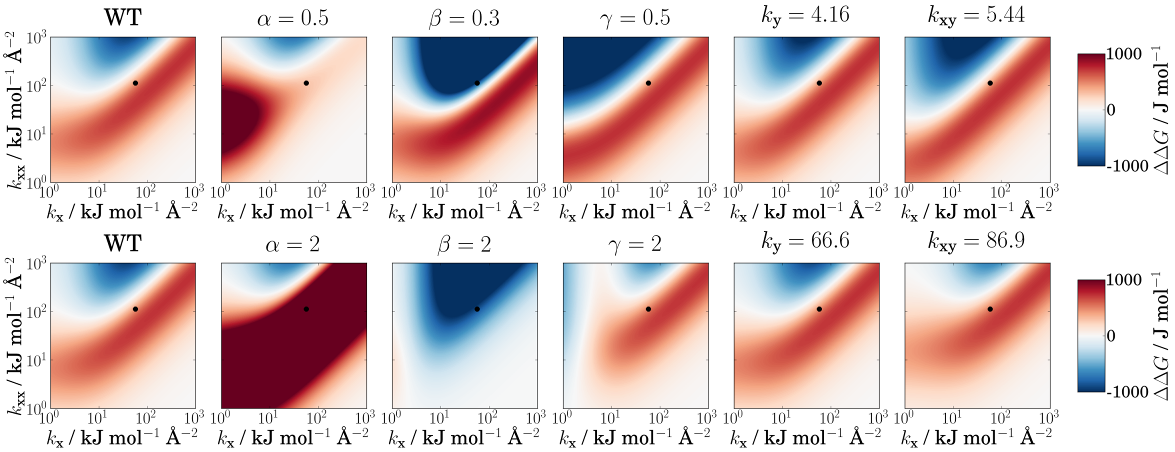


Figure 4.9: Comparison of changes in the parameter space upon changing one parameter. With all other parameters set to WT values, this figure explores the changes to the free energy landscape of the MMM, upon changing one of the parameters from the WT value. The top row shows how the free energy landscape is altered when decreasing the parameter indicated from WT value and the second row shows how it is altered when increasing the parameters.

What is also very interesting is that in the immediate vicinity of WT-CAP, the free energy landscape of the TMM is similar to those from the full-length CAP models. This is because CAP is nearing the region where k_x and k_{xx} are much greater than k_{xy} and k_y, meaning that these springs dominate the free energy of the system. In addition to this, looking at the parameters responsible for the interactions of the DNABD (γ, k_{xy} and k_y), it is evident that the effect that these parameters have on the allosteric free energy surface is almost insignificant when compared to the effect of α, β, k_x and k_{xx} (see figure 4.9).

Observing the allosteric free energy surfaces brings up the question as to why CAP would evolve to be on such a sensitive region of parameter space, where small changes to the interactions within the protein can make significant changes to the allosteric free energy. This would imply that mutations in the LBD can strongly affect the cooperative binding of cAMP. One argument for why CAP could have evolved in such a way is that in this region it is more sensitive to mutations, meaning that any changes in the environment that would change the cooperativity of cAMP to CAP would more likely be changed back to the desired cooperativity through a mutation. A mutation in the DNABD on the other hand may also have the undesired effect of removing the protein's ability to bind DNA, so would have no benefit of being mutated in a change of environment.

These SCG models do not take into account the fast modes of motion. However, as was shown by Cooper and Dryden [36] (discussed in section 1.4.3), the slow, global modes of motion are the significant contributions to the allosteric free energy. This hopefully means they are sufficient for modelling allosteric signalling transferred by thermal fluctuations. There have been previous models that include local fast modes [54], capturing the effect that the contributed to cooperative binding by coupling to the global slow motions. They contributed to the allosteric free energy by affecting the amplitude of the slow modes, but did not significantly alter the shape of the allosteric free energy landscape.

SCG models such as the ones discussed in this thesis are useful to investigate what global properties of the protein homodimer can cause cooperative ligand binding. They can also be used to investigate what changes to the interaction strengths within the system can be made to manipulate the cooperativity of ligand binding. However, the models cannot be used alone without advanced biochemical knowledge and thorough structural investigation to determine mutations that can be made to cause the changes to the models' parameters and hence manipulate allostery.

4.2.2 Elastic Network Model

The elastic network model on the other hand is able to explicitly mutate chosen residues in the protein by adjusting the spring constants for the springs connected to the mutated residue. This was performed for CAP by Dr. Tom Rodgers and the results in this section were first published in [59].

As previously explained, this model returns the normal modes of motion of the protein with a simple harmonic potential function. These normal modes were used to determine the free energy of the system and hence the allosteric free energy was determined by finding the differences between these free energies using equation 1.4.19.

To verify that the motion observed by the ENM was not an artefact of coarse graining the system, the B-factors determined from this method were compared to the B-factors from the original PDB and the B-factors determined with fully atomistic MD solutions (figure 4.10a). The experimental B-factors represent a combination of the static and dynamic disorder in the crystal, so represent a reasonable approximation of the local motions observed in the protein. The motions of the protein in a crystal

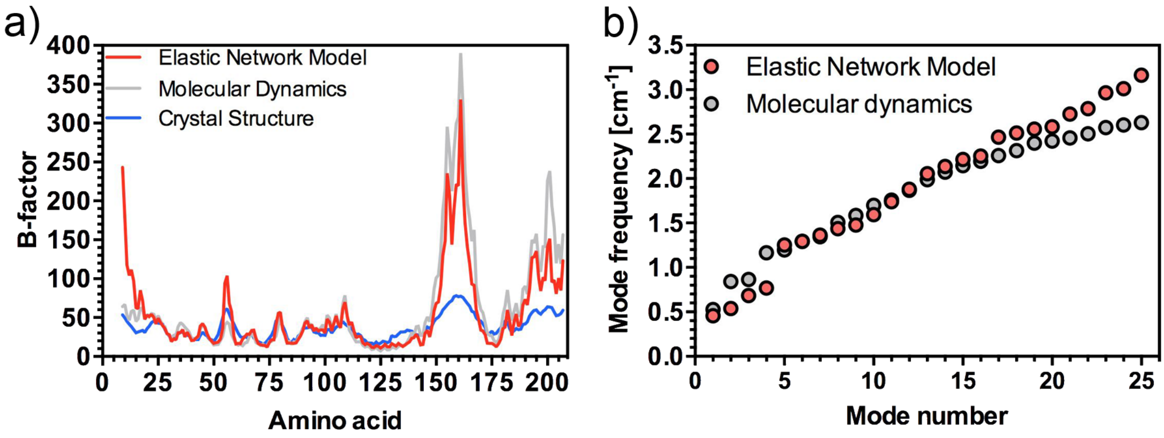


Figure 4.10: Validation of the ENM methodology. a) The plot of the B-factor against amino acid number for the ENM, molecular dynamics and the X-ray structure. The ENM and atomistic MD show good B-factor agreement with each other and qualitative agreement with the crystallographic B-factors. b) The plot of the normal mode frequency against the mode number as calculated by the ENM and by PCA show good agreement (from [59]).

are hindered due to the crystal packing, therefore the motion observed in crystallised proteins is typically less than the motion that would be observed in solution. Hence these B-factors are generally lower than B-factors that would be observed in solution. There was good agreement between the ENM B-factors, the atomistic B-factors from MD and the crystallographic B-factors when taking into account the reduced motion in a crystal. The ENM and MD, being unhindered show larger B-factor amplitudes in the chain termini and flexible loop regions such as residues 150-175 than the hindered crystallographic B-factors.

The frequency mode structure of the normal modes observed with ENM were verified by comparison to the structure of the normal mode frequencies observed by principal component analysis of a fully atomistic MD simulation (figure 4.10b). The distribution of mode frequencies from ENM agreed well with the vibrational frequencies derived from PCA, showing that the total predicted motion using the ENM is similar to other methods of analysis and is not a feature of the coarse graining in the model.

Free energies of each bound state of CAP were calculated using the frequencies of the normal modes and the full harmonic solution for the given system. The ENM was able to correctly identify the negative cooperativity of CAP with an allosteric free energy of $\Delta\Delta G = 749$ J mol^{-1}; a value consistent with experimental results (chapter 6) [59, 120–122].

The effect of mutating the side chains of amino acids at sites distant from the

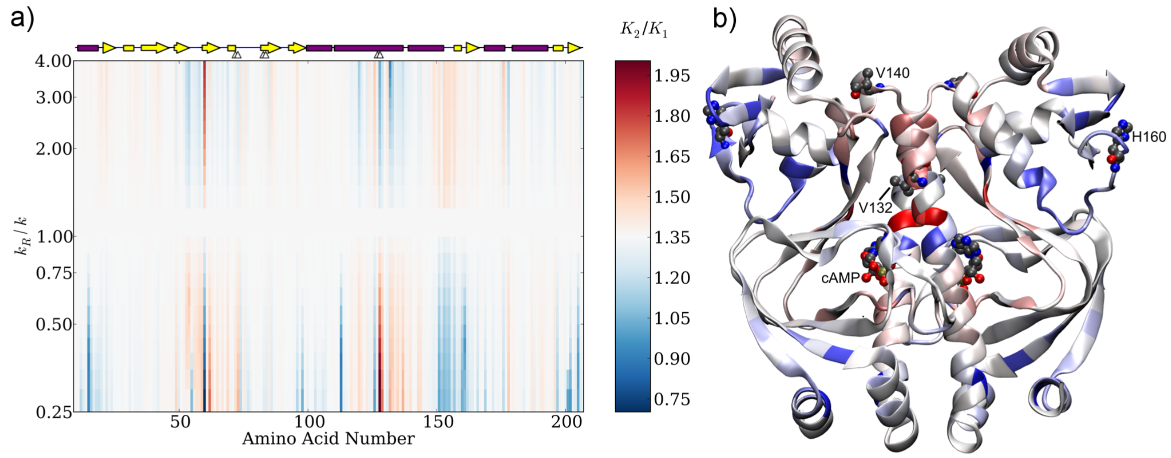


Figure 4.11: a) A map for the global control space of allostery in CAP calculated from the ENM. The map plots the change in cooperativity coefficient (K_{a2}/K_{a1}) due to the dimensionless change in the spring constant (k_R/k) for the mutated residue with the amino acid number shown. Red corresponds to an increase in K_{a2}/K_{a1} (stronger negative cooperativity), whereas blue corresponds to a decrease in K_{a2}/K_{a1} (weaker negative cooperativity or positive cooperativity). The secondary structural elements of CAP are represented above the plot, with purple bars representing α-helices and yellow arrows representing β-sheets. b) The global map demonstrating the effect of loosening mutations (k_R/k=0.25) mapped onto the real-space coordinates of WT-CAP. The mutation sites investigated both experimentally and through molecular dynamics simulations are indicated (From [59]).

binding site of cAMP alters the hydrophobic or electrostatic forces between the mutated residue and its neighbouring residues. As previously mentioned, this is modelled in the ENM by changing the relative spring constant of the mutated residue such that $k_\mathrm{R}/k \neq 1$, with $k_\mathrm{R}/k > 1$ corresponding to a tightening of local interactions and $k_\mathrm{R}/k < 1$ corresponding to a loosening of local interactions. It is assumed that a structural change would not be caused by this mutation. The ENM was then used to observe how these mutations affected the normal modes of the protein, which in turn would alter the cooperativity of cAMP binding.

The change in allosteric free energy ($\Delta\Delta G$) or the ratio of association constants (K_{a2}/K_{a1}) as a function of altering the amino acid sequence (one residue at a time) can therefore be viewed as a quantitative map of the contribution of the normal modes to cooperativity. A map such as this could theoretically be constructed experimentally by mutating each residue individually, however it is much more convenient to perform this computationally. A scanning mutagenesis was therefore performed using the ENM, changing the effective elastic potential of all springs extending from the mutated residue, one residue at a time. The increase and decrease in the spring constants simulated the strengthening and weakening of side chain interactions.

A colour coded map showing the effect that these mutations has on the cooperativity of cAMP binding in CAP is plotted by residue number (a) and in real space (b) in figure 4.11. The global map (a) exhibits large regions where the cooperative binding of cAMP is sensitive to changes in the strength of the side chain interactions. It is evident from (b) that many of these control regions are at sites distant from the cAMP binding site, such as residues 130 to 137 at the dimer interface and some are distant both from the binding sites and the dimer interface, such as residues 150 to 162 in the DNABD. These regions appear to control the cooperativity of cAMP binding without directly interacting with the ligand binding site or contributing to a spatially recognisable dynamic pathway. The study by Cann *et al.* also displayed evidence that these amino acids that make larger changes to the cooperativity of cAMP binding are more likely to be conserved in nature [59].

4.2.3 Overall Discussion

The ENM shows some consistency with the SCG models. The mutations in the ENM that would weaken the interactions between the monomers, for example the residues on the central alpha helix, appear in general to increase the negative cooperativity of the system. This even occurs in regions along the alpha helix that are distant from the cAMP binding site, so would not be directly affected by the presence of cAMP or lack thereof. These weakening mutations are equivalent to the reduction of the spring constant k_{xx} in the SCG models, which also increased the magnitude of the negative cooperativity observed.

Beyond this, it is more difficult to see similarities between this model and the SCG model. Mutations that would weaken the interactions within the monomer, which would lead to a reduction in k_{x} in the SCG models, appear to cause changes to the cooperativity in CAP in either direction. However, the majority do appear to push the protein more towards positive cooperativity, which is what was predicted by all of the SCG models. The regions where weakening mutations do not push the cooperativity towards positive cooperativity are generally near the cAMP binding site, which could cause changes to the interactions with cAMP. As the SCG models do not model the mutant as additional particles in the model in the way that the ENM does, there are less likely to be similarities between the two types of models for mutations in this region of the protein.

The consistency between the ENM and the SCG SMM and MMM also appears to break down when mutations of the ENM are made in the DNABD. The ENM suggests that a number of weakening mutations within this domain appear to push the system toward positive cooperativity, whereas the SMM and MMM would suggest that these mutations would make very little difference to the cooperativity of cAMP binding. Assuming the accuracy of the ENM (which is tested in chapter 6), this would suggest either that the DNABDs have been incorrectly modelled in the SCG model, or that the TMM, which would model the changes in the DNABD as changes to k_x is a better model for full length than the more complex models. One should also note that the position of the mutations that reduce the negative cooperativity are all distant from the DNA recognition helix, so if these mutations were made, they are unlikely to directly remove the protein's ability to bind DNA.

A number of residues were chosen to mutate experimentally and computationally using MD to test the ENM. The experiments aim to show that the changes to cooperativity observed by the ENM can also be seen without a change in the conformation of the protein backbone. The MD simulations aim to see if the changes in cooperativity can also be captured by atomistic detail simulations and also to investigate whether the ENM and SCG models are sufficient models for describing the allostery in CAP. For example, it will investigate if more subtle effects such as enthalpy-entropy compensation contribute to the allosteric signalling in CAP; the models in this chapter really only capture the contributions from the configurational entropy.

The mutants investigated experimentally were V132A, V132L, V140A, V140L, H160L and Q108A. V132A and V132L were chosen as loosening and tightening mutations respectively, acting by altering the strength of the local hydrophobic interactions. With the ENM V132A was predicted to make CAP more negatively cooperative and V132L predicted to push CAP toward positive cooperativity. The V140A, V140L and Q108A mutations were chosen as they were predicted by the ENM to not make a change to the cooperativity regardless as to whether mutations at these sites increased or decreased the local interactions. H160L was chosen as it was a surface mutation, distant from the interface between monomers that would remove local electrostatic interactions, and was predicted to push the system towards positive cooperativity. Chapter 6 contains the experimental investigation of these mutants.

The mutants investigated using MD were V132A, V132L, V140A, V140L and H160L; chosen for the reasons given above. The results from these MD investigations are included in chapters 5 and 6.

Chapter 5

Atomistic Studies of Dynamic Allostery in CAP

5.1 Methods

5.1.1 Molecular Dynamics Simulations

Molecular dynamics simulations were performed using the AMBER molecular dynamics package [130]. A number of different force fields were used in this thesis, which were ff99SB [215], ff99SB-ILDN [216], ff99SBnmr [217], ff03 [218] and ff12SB. The force field used for each simulation is outlined in the corresponding results sections of this chapter. A number of simulations were run for apo-CAP, holo$_1$-CAP and holo$_2$-CAP, for the WT, V132A, V132L, V140A, V140L and H160L variants of CAP. CAP variants were made by mutating the WT structure with Coot [219, 220].

Partial charges for cAMP were either AM1-BCC charges [221], determined using Antechamber [222] or B3LYP/6-311G** charges [223, 224], determined using Gaussian 09 [225]. The charge model used for each simulation is discussed in greater detail in 5.2.2. The general amber force field (GAFF) was used for the other force field parameters of cAMP.

The initial starting structures for each system were the vacuum energy minimised structures that were used for normal mode analysis as outlined in section 5.1.2. The force field was added to each system using LEaP from the AMBER tools package [130], using the default protonation for the side chains as determined by LEaP. Next, more water was added to the already present crystallographic water of the systems, bring the total to 10297 water molecules. The water molecules used the TIP3P force field. The net charge of CAP was neutral, so when the negatively charged cAMP molecule was present in the simulations, a neutralising Na+ ion was included for each cAMP molecule.

Periodic boundary conditions were employed using a periodic truncated octahedron. The energy was minimised in two stages; firstly the protein was held with restraints while the water was energy minimised, then the restraints were removed and the entire system energy minimised (using the configuration files in appendix A).

The temperature of each system was then raised to 300 K over a 20 ps period and density equilibration was performed over the next 20 ps by switching on pressure coupling using the thermostat and barostat described in section 2.1. The systems were then simulated for up to 300 ns, with up to 100 ns of the simulation being kept as the equilibration time. As the simulation times differ for each system, they are included in section 5.2, where the results are shown and discussed.

For each of the simulations, a time step of 2 fs was used. The lengths of all bonds containing hydrogen were constrained using the SHAKE algorithm. A short range-cutoff of 10 Å was used for non-bonded forces, with the long range portion of the electrostatic Coulomb interactions represented by an Ewald summation.

Simulations were performed using the AMBER 11 or AMBER 12 MD packages [130]. The AMBER 12 simulations used the GPU accelerated code for all simulations except energy minimisation and equilibration [226, 227]. All MD input files can be found in appendix A. The version of AMBER used for each simulation is explicitly stated in section 5.2 when the results for them are discussed.

5.1.2 Normal Mode Analysis

NMA was performed with the ff99SB-ILDN force field (5.1.1) using Groningen machine for chemical simulations (GROMACS) [155] (see 2.2). NMAs were performed both on the initial starting crystal structure (4HZF) and snapshots from the MD simulation. The more stringent L-BFGS was used to minimise the energy of the system until the maximum force acting on any atom in the system was less than the tolerance of 0.0005 kJ mol^{-1} nm^{-1}. A sodium ion was added for each cAMP molecule present in a system, to keep the charge of the system neutral. This was to ensure the structure is truly at a minimum of the free energy surface. No cut-off was used for non-bonded interactions for both the energy minimisation and the NMA.

The GROMACS configuration files used to perform the NMAs can be found in appendix B.

NMA on X-ray structure

Starting by minimising the energy of the WT holo$_2$-CAP crystal structure in vacuum; the WT holo$_1$-CAP structure was then created by deleting one cAMP molecule from the structure and re-minimising. This step was then repeated to obtain the structure for apo-CAP. These structures were used as the minimum energy structures in NMA and were mutated using crystallographic object-oriented toolkit (*Coot*) [219] to create starting structures for the variants, which were in turn energy minimised before NMA.

Normal mode analyses were then performed on the three bound states of each of the CAP variants.

NMA on MD snapshots

Snapshots from the MD simulations with the ff99SB-ILDN force field were taken every 4 ns for each of the CAP variants. These structures were then stripped of water and re-imaged to ensure the two monomers were not split across the periodic boundary. The systems were then energy minimised before performing a NMA.

5.1.3 Principal Component Analysis

PCA was performed on windows of the MD simulations described in 5.1.1 using command prompt process trajectory (CPPTRAJ) from the AMBER Tools package [228, 229] and analysed using the same program and in house software called DDPT [172] (see 2.3). Windows with a 20 ns duration were used after an initial equilibration period of 100 ns. 20 ns windows were chosen as a compromise of windows estimated to be long enough to capture global motions of the protein, while allowing a number of windows to be used to allow statistically significant averaging.

5.1.4 MM/PBSA Calculations

MM/PBSA calculations were performed on the MD-simulations of all of the systems using the ff99SB-ILDN force field using the MMPBSA.py scripts available in AMBER tools [130, 180] (see 2.7). The MM/PBSA contribution to the free energy was determined for apo, holo$_1$ and holo$_2$-CAP for each of the simulated variants, as well as cAMP. These were then used to help calculate ΔG_1, ΔG_2 and $\Delta\Delta G$. The calculations were performed on the systems after 100 ns equilibration time, using 2000 frames over the 200 ns of remaining simulation. A 20 ns simulation of cAMP in a box of water (with 2000 frames and a 4 ns equilibration time) was used for the calculation performed on the ligand alone.

For the PB portion of the calculation, an internal dielectric constant of 1 was used for the protein and a dielectric constant of 80 was used for the solvent. A salt concentration of 300 mM was used to try match the experimental conditions in chapter 6 as closely as possible. The radius of the solvent probe used was 1.4 Å. For the SASA portion of the calculation, when the entire nonpolar contribution was modelled with the SASA model, a surface tension coefficient, γ, of 0.00542 kcal mol^{-1} Å^2 and a cavity offset, c, of -1.008 kcal mol^{-1} were used. When only the repulsive part of the nonpolar contribution to the free energy was modelled by the SASA model, a surface tension coefficient, γ, of 0.0378 kcal mol^{-1} Å^2 and a cavity offset, c, of -0.5692 kcal mol^{-1} were used (see 2.1.12). In this case, the attractive portion was determined by an approximation to the van der Waals attractive forces [148].

5.2 Results and Discussion

5.2.1 Investigating the Protonation State of cAMP

The ligand cAMP has a number of protonation states that are all prospective states to model cAMP in this thesis. This section investigates which protonation state is the most sensible to study for further simulations.

Observing cAMP in the crystal structure of CAP (4HZF), it is apparent that it forms a number of hydrogen bonds and electrostatic interactions with the protein as shown in figure 5.1. The side chain carboxylate oxygens of Glu73 can both separately form hydrogen bonds as acceptors with a central oxygen in cAMP, however it is unlikely

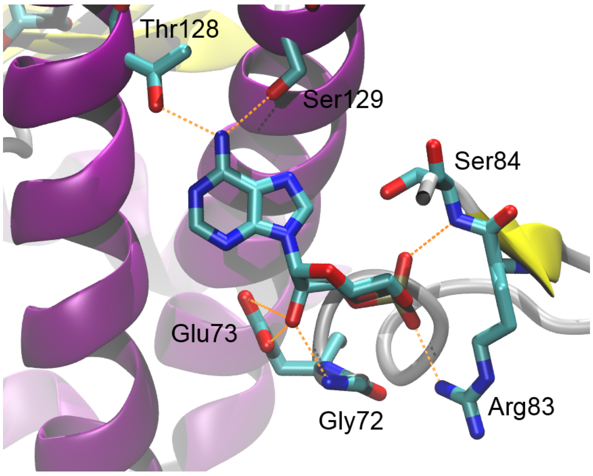


Figure 5.1: cAMP binding site, showing the hydrogen bonds present between cAMP and CAP (PDB ID 4HZF). Hydrogen atoms are not present in the PDB structure, so they are omitted for clarity.

that they would form hydrogen bonds simultaneously. The same oxygen in cAMP can also form a hydrogen bond with the donor nitrogen in the backbone of residue Gly72. The side chain oxygen atoms of Thr128 from the same chain and Ser129 from the opposite chain are both hydrogen bond acceptors from the amino-nitrogen in cAMP. The backbone nitrogen of Ser84 is a hydrogen bond donor to one of the oxygen atoms in the phosphate group of cAMP. A side chain nitrogen of residue Arg83 also either acts as a hydrogen bond donor for a hydrogen bond with oxygen of the phosphate group of cAMP or forms an electrostatic interaction with the same atom, depending on whether the oxygen is protonated or not.

A number of structural models for cAMP are proposed, to be included in the further simulations of this study. These models are shown in figure 5.2, and will be referred to as (a),(b),(c) and (d), their labels in the figure from this point. The first two of these models (figure 5.2a,b) can be distinguished from the others by an additional hydrogen bonded to one of the oxygen atoms in the phosphate group of cAMP, with the hydrogen in (b) bonded to the other oxygen to the one in (a). Model (c) removes this hydrogen, giving an overall charge of -1 to cAMP. Model (d) is a zwitterionic form of cAMP, with an additional hydrogen bonded to the amino-nitrogen at the opposite end of cAMP. The structures displayed in figure 5.2 are the electronic energy minimised structures of cAMP using B3LYP/6-311G** [223, 224] within the Gaussian09 package [225].

To investigate which model to use for the future simulations, firstly the structures were energy minimised using B3LYP/6-311G** to calculate their minimum electronic

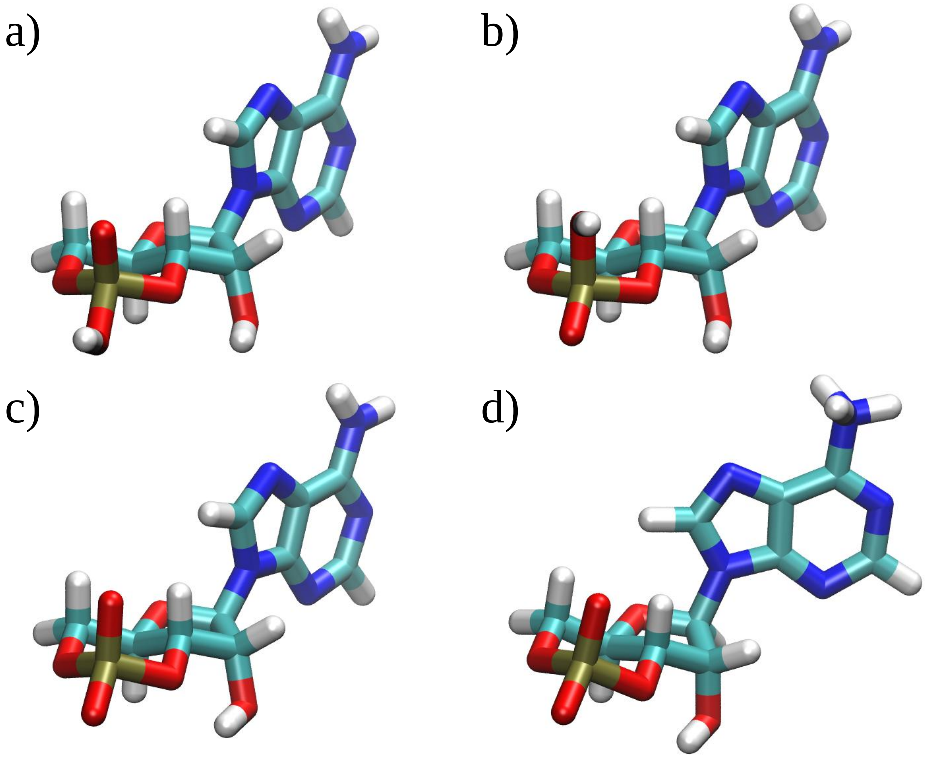


Figure 5.2: B3LYP/6-311G** cAMP energy minimised structures, determined using Gaussian [223–225]. The phosphate groups of (a) and (b) are protonated, each with the hydrogen on a different oxygen. For (c), the phosphate group is deprotonated, giving cAMP a charge of -1. (d) is a zwitterionic form of cAMP, with a deprotonated phosphate group and a protonated amino group at the opposite end of cAMP.

energies in vacuum and in a variety of solvents modelled implicitly with the polarisable continuum model [230]. These are shown in table 5.1. From this table it is evident that the electronic energy of form (a) and (b) is significantly lower than the electronic energy of form (d). For this reason, model (d) is not investigated further in this study. The electronic energy of these three forms cannot be directly compared to the electronic energy of the (c) form due to the loss of a proton in this form, so the (c) form is studied further alongside the (a) and (b) forms.

Next, the binding of cAMP models (a),(b) and (c) to the binding sites in WT-CAP was investigated by running 200 ns molecular dynamics simulations of holo$_2$-CAP with each of the three models of cAMP. The hydrogen bonding network in the binding site was then investigated for each model of cAMP. Table 5.2 shows the occupancy of the hydrogen bonds in the hydrogen bonding network identified in figure 5.1. Here the occupancy is calculated by defining the chosen hydrogen bond as being formed when the distance between the two heavy atoms in the hydrogen bond is less than 3.5 Å and

	a	b	c	d
Vacuum	0.00	-3.51	1345.93	264.99
Water	0.00	-7.62	1158.61	56.58
Methanol	0.00	-7.37	1161.95	61.46
Toluene	0.00	-4.44	1244.82	164.90

Table 5.1: The Electronic Energies of the different models of cAMP in vacuum, water, methanol and toluene. The energies are minimised in Gaussian 09 and are given in kJ mol^{-1} relative to model (a).

H-Bond Donor			**H-Bond Acceptor**		**Occupancy / %**		
Residue	**Atom**	**Hydrogen**	**Residue**	**Atom**	**a**	**b**	**c**
Gly72$_{\mathrm{A}}$	H	N	cAMP$_{\mathrm{A}}$	O$^{2'}$	99	97	99
Arg83$_{\mathrm{A}}$	H_{η}^{11}	N_{η}^{1}	cAMP$_{\mathrm{A}}$	O$^{\mathrm{2P}}$	0	0	100
Ser84$_{\mathrm{A}}$	H	N	cAMP$_{\mathrm{A}}$	O$^{\mathrm{1P}}$	83	49*	100
Ser84$_{\mathrm{A}}$	H_{γ}	O_{γ}	cAMP$_{\mathrm{A}}$	O$^{\mathrm{1P}}$	2	0	97
cAMP$_{\mathrm{A}}$	H^{6}	O$^{2'}$	Glu73$_{\mathrm{A}}$	$\mathrm{O}_{\epsilon}^{1}$	61	80	63
cAMP$_{\mathrm{A}}$	H^{6}	O$^{2'}$	Glu73$_{\mathrm{A}}$	$\mathrm{O}_{\epsilon}^{2}$	78	76	67
cAMP$_{\mathrm{A}}$	H^{10}	N^{6}	Thr128$_{\mathrm{A}}$	O_{γ}^{1}	96	98	84
cAMP$_{\mathrm{A}}$	H^{11}	N^{6}	Ser129$_{\mathrm{B}}$	O_{γ}	92	94	90

Table 5.2: Occupancy of the hydrogen bonds for the three models of cAMP identified in figure 5.2. The occupancy values shown are the average values across the two chains. *The (b) form of cAMP forms a partially occupied hydrogen bond (49%) between O$^{\mathrm{2P}}$ in cAMP (rather than O$^{\mathrm{1P}}$).

the heavy atom-hydrogen-heavy atom angle is greater than 120°.

From this table it is evident that certain hydrogen bonds in the cAMP hydrogen bond network with CAP do not form properly for cAMP models (a) and (b). Model (b) also manages to rearrange within the binding site so that the other oxygen from cAMP's phosphate group forms a hydrogen bond with residue Ser84. This leaves cAMP in a rather constrained orientation that does not agree with the crystal structures. A number of other hydrogen bonds are lost entirely for models (a) and (b). Model (c) is the only model that manages to correctly form the hydrogen bonding network seen in the crystal structures for the duration of the MD simulation. Images of cAMP binding using the three different models can be seen in figure 5.3. As can be seen in this figure, when either model (a) or model (b) are used to represent cAMP, the residue Arg83, rearranges, so that it no longer forms the correct hydrogen bonds with cAMP, instead

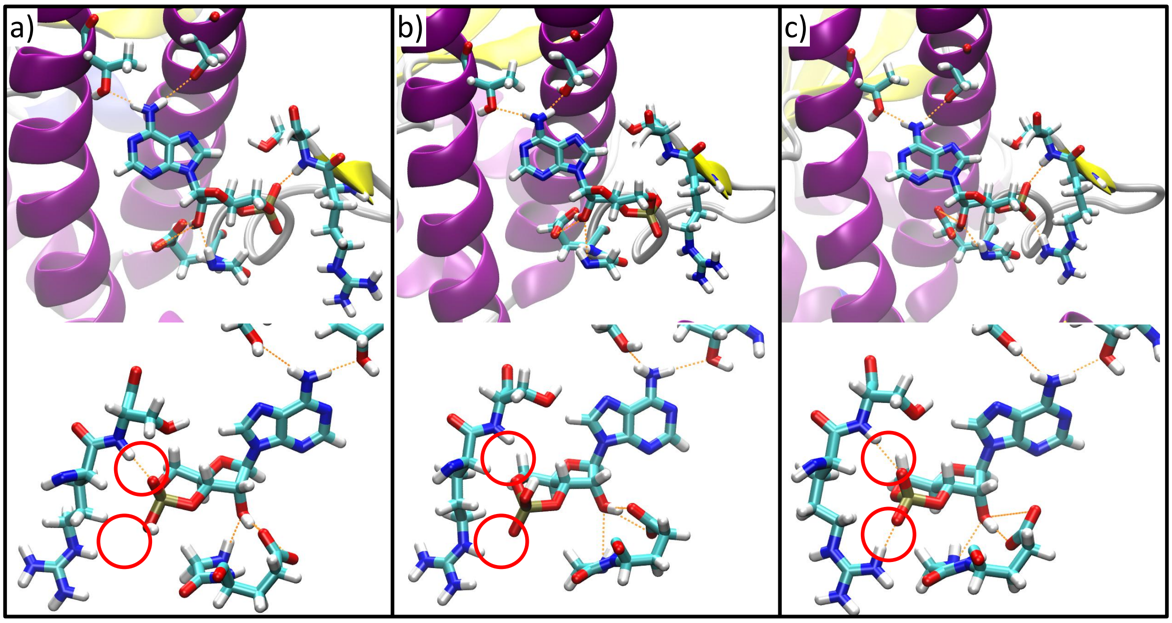


Figure 5.3: Snapshots of the binding of cAMP to CAP using three different models of cAMP. For models (a) and (b), the interaction between cAMP and Arg83 is incorrect, and for model (b), the hydrogen bond between Ser84 and cAMP does not form either. The location of these interactions, or the location of where the interactions are missing from are highlighted with red circles. It is only for model (c) that all of these interactions are present.

creating interactions with other parts of CAP because the hydrogen on the phosphate group of cAMP creates too many steric clashes in this region of the binding site. From the crystal structures available it is obvious that the Arginine does not rearrange in the binding site. Therefore, further simulations in this thesis use model (c) of cAMP.

5.2.2 AMBER Force Field

A number of AMBER force fields are available for running simulations in the AMBER MD package; this section investigates which one is the most suitable for simulating CAP. Some of the more recent are the ff99SB, ff99SB-ILDN, ff99SBnmr, ff03 and the ff12SB force fields. For the reason that ff99SBnmr is not easily ported into GROMACS and some of AMBER's tools are not optimised for use with this force field, it was not used for the production simulations in this study. Furthermore, the ff03 force field was not used, as although it was shown in a study to perform similarly well to ff99SB, it was shown to over stabilise alpha helices [215]. A later study has also shown that the ff99SB based force fields also perform better when calculating binding energies, using the MM/PBSA method on longer simulations, as is performed in this thesis [231] (section 5.2.8). The recent ff12SB force field was not used for the variant simulations

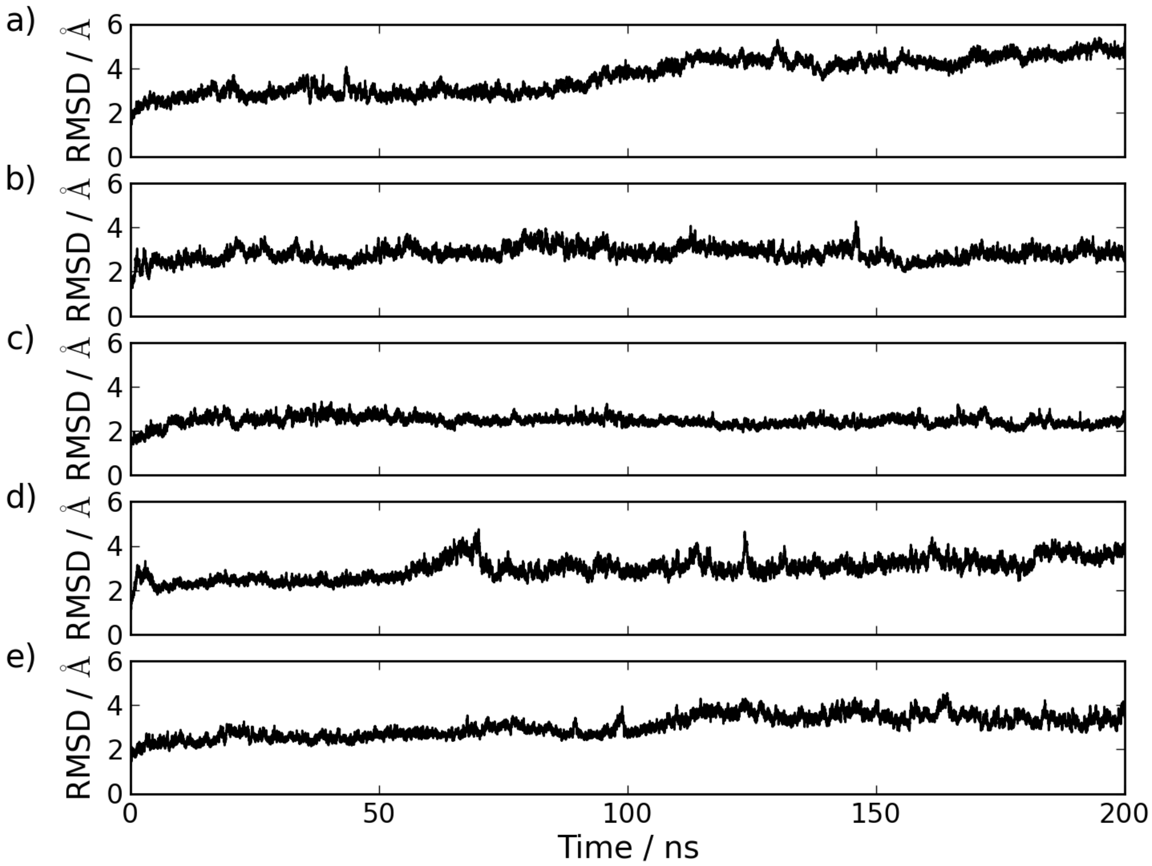


Figure 5.4: RMSD of simulations of apo-CAP against the first frame of the simulation using different AMBER force fields: a) ff99SB b) ff99SB-ILDN c) ff99SBnmr d) ff03 e) ff12SB

as this force field was added to the available AMBER force fields when most of this study had been performed.

The force fields ff99SB-ILDN was chosen to use for the majority of the other simulations shown in later sections of this chapter based on these conclusions from the literature. ff99SB-ILDN was chosen rather than ff99SB as it contains a few improvements to side chain torsional parameters over ff99SB for a small number of residues.

Initially, a number of 200 ns simulations were performed to look at the effect of the force field on the stability of the simulation. These simulations were performed using the PDB with ID code 1G6N, removing the cAMP molecules to make an apo-CAP system. The AMBER 11 CPU code was used for all simulations except the ff12SB force field, which used AMBER 12 GPU code, the release of AMBER that this force field became available.

Figure 5.4 shows the root-mean-square deviation (RMSD) of simulations of apo-CAP using the different AMBER force fields. This figure suggests that all of the simulations tend to converge within the first 100 ns of the simulation, with the structure in the simulations settling at around an RMSD of 3-5 Å from the starting structure. The simulations using ff99SB seems to have a change in the RMSD after 100 ns into the simulation. This could indicate a move to another minimum in the global conformation landscape, however as the change is only small it is only a minor change in the conformation. From visual inspection of the simulation, this can be confirmed to be the case. The other simulations all appear to converge in terms of the RMSD.

Of these force fields, the ff99SB and ff99SB-ILDN force fields were used to check the effect of the force field on the binding site of cAMP by simulating holo$_2$-CAP for 200 ns with the same conditions described above for apo-CAP. Later in the study, the ff12SB force field was added to this analysis and is included in this section

The first check to see the efficacy of each force field to model the cAMP binding sites was to study the hydrogen bonds between cAMP and the protein. The hydrogen bonds studied are the same as in section 5.2.1 and using the same method. The occupancy of each of the hydrogen bonds are displayed in table 5.3.

It is evident from this table that both the ff99SB and ff99SB-ILDN force fields perform well when recreating the hydrogen bonds between cAMP and CAP, with the occupancy of most of the hydrogen bonds being close to 100% occupied for most of the hydrogen bonds observed in the crystal structures. They both seem to represent the cAMP hydrogen bonding network equally well, which is not surprising as the two force fields are closely related.

The third force field, ff12SB does not form all of the hydrogen bonds to sufficient occupancy, with a number of hydrogen bonds having a lower occupancy. In this case, it is due to one chain having almost fully formed hydrogen bonds at all times and the other chain losing a few hydrogen bonds for parts of the simulation. For example, the hydrogen bond between Arg83 and cAMP has 99% occupancy in the first monomer and 70% occupancy in the other and the hydrogen bonds from Ser84 are 100% and 96% in the first monomer and 64% and 63% in the other. Studying the simulation of this system, it is evident that the DNABD rotates during the simulation by a significant amount for the monomer with lower hydrogen bond occupancy. This rotation allows the binding site of cAMP to open up slightly, allowing these hydrogen bonds to break.

To quantify the extent of the DNABD rotation, the LBDs were aligned and the

H-Bond Donor			H-Bond Acceptor		Occupancy / %		
Residue	**Atom**	**Hydrogen**	**Residue**	**Atom**	**ff99SB-ILDN**	**ff99SB**	**ff12SB**
Gly72$_A$	H	N	cAMP$_A$	$O^{2'}$	99	100	99
Arg83$_A$	H_η^{11}	N_η^1	cAMP$_A$	O^{2P}	100	100	85
Ser84$_A$	H	N	cAMP$_A$	O^{1P}	100	100	82
Ser84$_A$	H_γ	O_γ	cAMP$_A$	O^{1P}	97	97	79
cAMP$_A$	H^6	$O^{2'}$	Glu73$_A$	O_ϵ^1	63	63	70
cAMP$_A$	H^6	$O^{2'}$	Glu73$_A$	O_ϵ^2	67	68	63
cAMP$_A$	H^{10}	N^6	Thr128$_A$	O_γ^1	84	95	79
cAMP$_A$	H^{11}	N^6	Ser129$_B$	O_γ	90	96	88

Table 5.3: Occupancy of the hydrogen bonds for three AMBER force fields; ff99SB-ILDN, ff99SB and ff12SB. The occupancy values shown are the average values across the two chains.

Force Field	RMSD (LBD) / Å	RMSD (Full) / Å
ff99SB-ILDN	0.0	0.0
ff99SB	1.2	2.7
ff12SB	1.5	9.3

Table 5.4: RMSDs between final structures of simulations using the force fields ff99SB-ILDN, ff99SB and ff12SB. The structures were aligned using the backbone of the LBDs and RMSDs calculated for the LBDs and the full protein.

RMSD calculated, using the Kabsch algorithm [232] in VMD [214], for the LBDs alone and the full length protein using the final structures of each simulation. This is shown in table 5.4. This shows that there are not many structural differences between the backbones between the LBDs of each force field, however the LBD from the ff12SB simulation shows greater deviation from the ff99SB-ILDN structure than the ff99SB structure does. This is most likely due to the loosening of the LBD caused by the rotation of the DNABD.

The extent of the rotation of the DNABD can be seen in table 5.4, as it increases the RMSD of the full length CAP to 9.26 Å from the ff99SB-ILDN structure. It can also be seen in figure 5.5. This figure shows an analysis to find the axis and extent of rotation needed to rotate the DNABDs from the final structure of the ff99SB-ILDN to the DNABDs from the final structure of the ff12SB simulation performed using the hingefind plugin of VMD [214, 233]. The domains of the protein were determined by finding rigid regions of the two structures that had an RMSD below the tolerance of

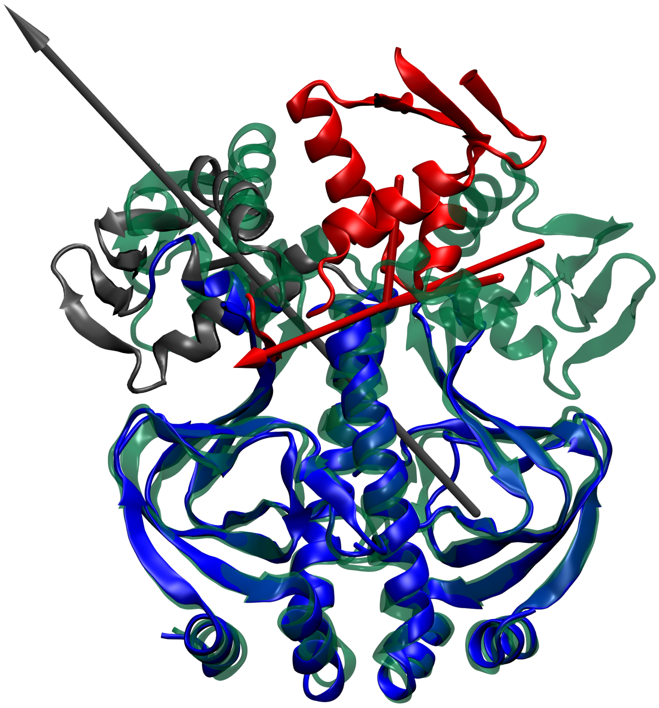


Figure 5.5: Rotation of DNABD in simulation of holo$_2$-CAP using the ff12SB force field, shown relative to the final structure from the ff99SB-ILDN simulation. The axis of rotation between a DNABD and the LBDs is shown as an arrow in the same colour as the DNABD.

4.5 Å.

The most perturbed DNABD of the ff12SB structure has a rotation angle of 66.6°, while the less perturbed domain had a rotation angle of 27.5°. With both DNABDs, the axis of rotation is near the region between the DNABD and the LBD.

This large rotation of the DNABD is also accompanied by a rearrangement of the cAMP binding site in the monomer which undergoes the DNABD rearrangement. Figure 5.6 shows that the ligand binding site undergoes a structural rearrangement just over 100 ns into the simulation for the ff12SB simulation of holo$_2$-CAP. The other two force fields used however, do not show a structural change in the binding site.

The experimental crystal structures of a number of variants of holo$_2$-CAP do not show the rotation of the DNABD seen in the ff12SB simulation. The simulations using ff99SB and ff99SB-ILDN on the other hand do not exhibit structural rearrangements to the extent seen in the ff12SB simulation. Furthermore, as seen in figure 5.1, the structure of the binding site of holo$_2$-CAP seen experimentally (4HZF) shows the hydrogen bonding network that is investigated in table 5.3. The ff99SB and ff99SB-ILDN force fields represent this hydrogen bonding network better than the ff12SB force field. This would perhaps suggest that ff99SB and ff99SB-ILDN are better force fields for

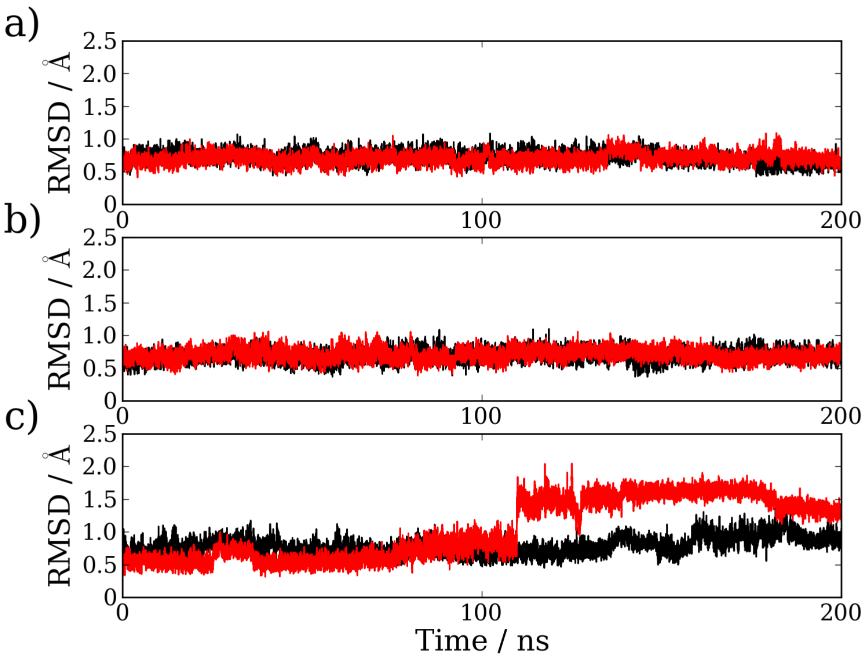


Figure 5.6: RMSD of cAMP and the residues it forms hydrogen bonds with for a) ff99SB-ILDN, b) ff99SB and c) ff12SB. The RMSDs were calculated with the residues in table 5.3, with the black and red plots representing the two different monomers of CAP. The RMSDs were determined using cpptraj, part of the AMBER tools package [228].

representing holo$_2$-CAP than ff12SB. As mentioned earlier, ff99SB-ILDN was chosen over ff99SB for the remaining simulations due to the small improvements to torsional parameters that it included.

5.2.3 Normal Mode Analysis

Normal mode analyses were performed using the ff99SB-ILDN force field for a number of variants of CAP. The variants investigated were WT, V132A, V132L, V140A, V140L and H160L. They were chosen as these were identified by the elastic network model as variants that could either affect the allostery in CAP (V132A, V132L and H160L) or cause no change to the allostery in CAP (V140A and V140L). The latter two variants were contained as negative controls.

The structure used to create the starting structures for the normal mode analyses was the X-ray structure with a PDB ID number of 4HZF. The structure of holo$_2$-CAP was energy minimised in GROMACS using L-BFGS. One molecule of cAMP was then stripped to create the structure of holo$_1$-CAP, which was again energy minimised using the L-BFGS method. The second cAMP molecule was then stripped before

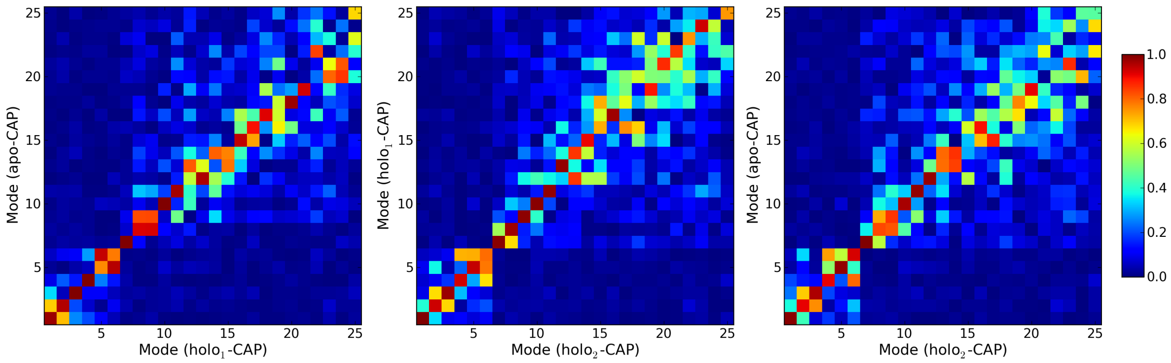


Figure 5.7: Overlaps of the first 25 normal modes from the three bound states of WT-CAP. The overlaps of the normal modes from apo-CAP vs. holo$_1$-CAP, holo$_1$-CAP vs. holo$_2$-CAP and apo-CAP vs. holo$_2$-CAP. The overlaps were calculated using only the C$_\alpha$ atoms for each system. The first six normal modes are translational and rotational modes for each system.

minimisation, giving the energy minimised structure of WT apo-CAP.

The WT apo-CAP, holo$_1$-CAP and holo$_2$-CAP systems were each mutated to create the corresponding systems for each of the variants, which were in turn energy minimised using the L-BFGS method. All of the energy minimised structures of the CAP variants were used for the normal mode analyses. All crystallographic water molecules were retained for all of these NMAs.

From the NMAs, information regarding the frequency and directionality of the global modes of motion can be obtained. Only the directionality of the motion is investigated in this section. The frequency of the low frequency modes are investigated in section 5.2.7 to estimate binding entropies.

Figure 5.7 shows the overlaps (see section 2.5.3) of the normal modes calculated for the three bound states of WT-CAP. It is evident from this figure that the slow modes for CAP as calculated by NMA only have subtle changes, with the majority of the modes changing only slightly. The first five or so (none zero) low frequency modes change the least between the three bound states, but the higher frequency modes have a larger change upon binding.

One interesting observation in the overlap of the modes from apo-CAP and holo$_1$-CAP is that the line of greatest overlap does not perfectly follow the diagonal, instead it appears that the overlap of the first 25 modes for holo$_1$-CAP correspond to the first 22 or so normal modes for apo-CAP. This could imply that more low frequency modes are activated for holo$_1$-CAP. This could be equivalent to a few of the low frequency modes in holo$_1$-CAP disappearing in apo-CAP by being decomposed across a number of the low frequency modes in apo-CAP. This appears to happen to some extent for mode numbers 12, 14, 19 and 25 of holo$_1$-CAP as well as a few others to a lesser extent. This activation of low frequency motions in the protein upon binding the first cAMP is consistent with the observation by Kalodimos *et. al* [17] when studying the change in dynamics upon binding cAMP to the truncated form of CAP (see section 1.5.5).

The picture upon binding the second cAMP is less clear, as the line of maximum overlap between the modes of holo$_1$-CAP and holo$_2$-CAP appear to follow the diagonal better, however the mode structure between these two bound states appears to deviate more than between apo and holo$_1$-CAP. It appears that a number of holo$_1$-CAP modes are decomposed over a number of holo$_2$-CAP modes upon binding and a number of holo$_2$-CAP modes are composed of parts of a number of holo$_1$-CAP modes. This cannot be said to show agreement with what was observed by Kalodimos *et. al* [17].

Figure 5.8 shows the cross correlations (see section 2.5.2) between the residue motion for apo-CAP, holo$_1$-CAP and holo$_2$-CAP. This shows the extent to which the motion exhibited by a certain residue is similar to the motion exhibited by the other residues.

It is evident that the residues that have correlated motion with the opposite chain tend to be either, near the interface between the two monomers, or near the cAMP binding site. These residues also seem to correlate well with the motion of cAMP in the same chain when it is present. The motion of the residues surrounding Thr128 and Ser129 also seems to correlate well with the motion of cAMP in the second chain. The proximity of these residues to the binding sites and the correlated motion across the interface could contribute to the dynamic allostery observed in CAP.

When viewing the cross correlation as a function of residue space, as seen in figure 5.8, it is difficult to pick out subtle changes to the protein motion upon cAMP binding. When looking at how the motion in CAP correlates to the motion of a single chosen residue, it is possible to study this correlation in the real space. This is shown in figure 5.9, where the residue Thr128 is chosen. This residue is chosen as it is one of the residues that forms a hydrogen bond with cAMP, and is on the dimer interface. This figure suggests that upon binding one cAMP, the correlated motion across the dimer interface seems to increase slightly, and binding the second cAMP causes another increase in the correlated motion across the interface. The motion of cAMP also correlates well with the motion of Thr128, even if cAMP binds to the opposite monomer. This hints at the importance of the residues within the vicinity of this one being important for allosteric signalling.

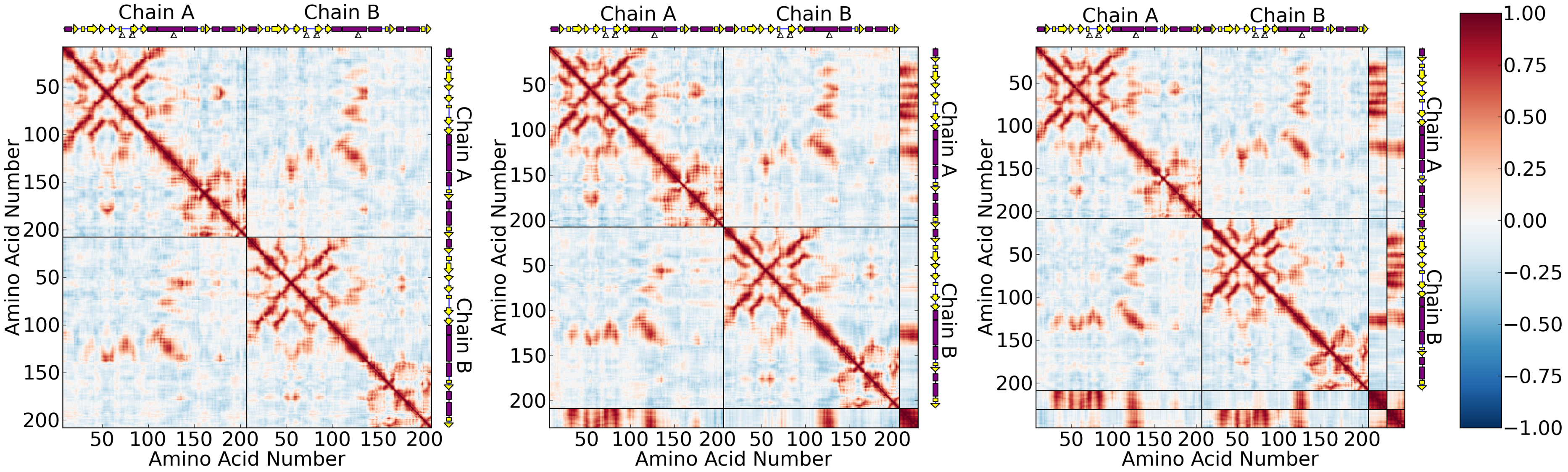


Figure 5.8: The cross correlation of the motion between residues as shown in 2-Dimensional residue space for the three bound states of CAP. The colour scale show the extent of cross correlation, with a cross correlation of 1 (red) indicating perfectly correlated motion and -1 (blue) showing perfectly anticorrelated motion. The secondary structure of CAP is indicated along the residue axes, with purple rectangles indicating an alpha helix and yellow rectangles and arrows indicating beta strands. The white triangles indicate the location of residues that form hydrogen bonds with cAMP. The cross correlation matrix was calculated using only the C_α atoms for the protein and all heavy atoms for cAMP.

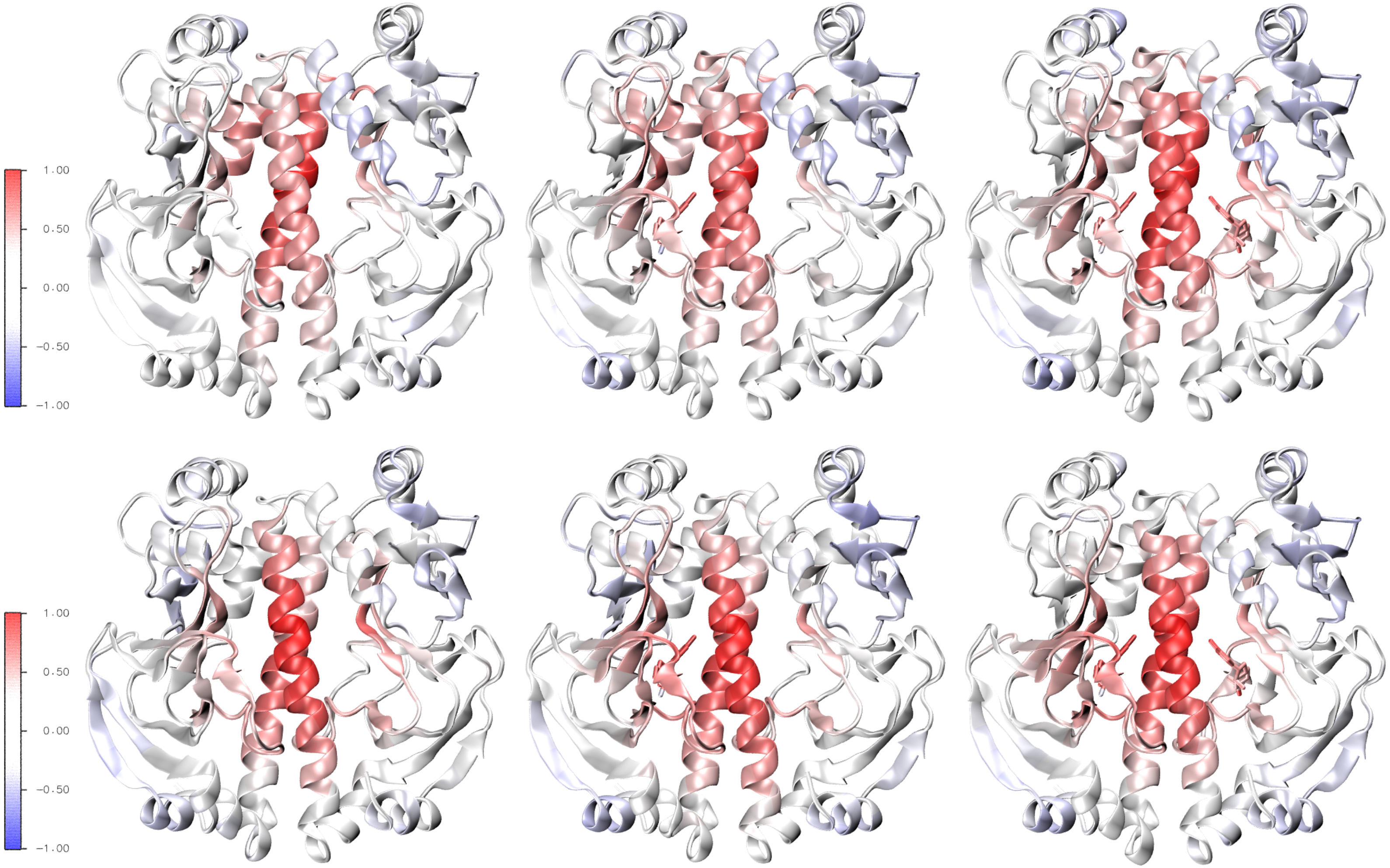


Figure 5.9: The cross correlation of the motion between residue Thr128 and the rest of the protein plotted in real space. The top row shows the correlation of the residues in the protein with Thr128 from the monomer which cAMP binds to first and the bottom row shows the correlation of the other Thr128 with the other residues in the protein. For both rows, apo-CAP is on the left and holo$_2$-CAP is on the right.

Figure 5.10 shows the overlaps of the normal modes between WT-CAP and the variants. This shows that the shape of the normal modes for the CAP variants are not very different to those for WT-CAP. Implying that making these variations to the secondary structure of CAP does not make large changes to the shapes of the normal modes. This is especially true for the variants V132A, V132L and V140A and appears to be the case for all three bound states of CAP. The variants V140L and H160L appear to change the shape of the normal modes to a greater extent.

The lack of change in the shape of the normal modes for some variants compared to WT-CAP would perhaps suggest that these vibrations are not the carriers of the allosteric signal. However, these NMAs are all performed using the same crystal structure, whereas in reality, the protein would explore a number of local minima, not just the one observed in the crystal structure.

What is quite surprising is that the negative control variant, V140L, appears to change the shape of the modes more than V132L does. It is hard to explain the reason for this, but could be due to the mutation in the V140L causing a larger perturbation in local structure.

It is perhaps less surprising that H160L makes larger changes to the structure of the normal modes, as this mutation is converting a large hydrophilic amino acid to a large hydrophobic one. Therefore it is more likely to make larger changes to the interactions within the local region of the protein.

However, as has already been discussed, the normal modes that were studied in this section were all calculated from very similar static structures that are simple mutations of the WT crystal structure. The method used does not allow the structure of variant to equilibrate once the mutation is made. The NMA investigation in section 5.2.7 attempts to tackle this problem by performing NMAs on snapshots of long molecular dynamics simulations of the CAP variants.

5.2.4 Molecular Dynamics Simulations of CAP Variants

A number of 300 ns MD simulations were performed on the three bound states of the CAP variants; WT, V132A, V132L, V140A, V140L and H160L as described in 5.1.1 and using the AMBER 12 GPU code [130]. B-factors were calculated by looking at the RMSF of the C_α atoms averaged over time. These B-factors are shown for the case

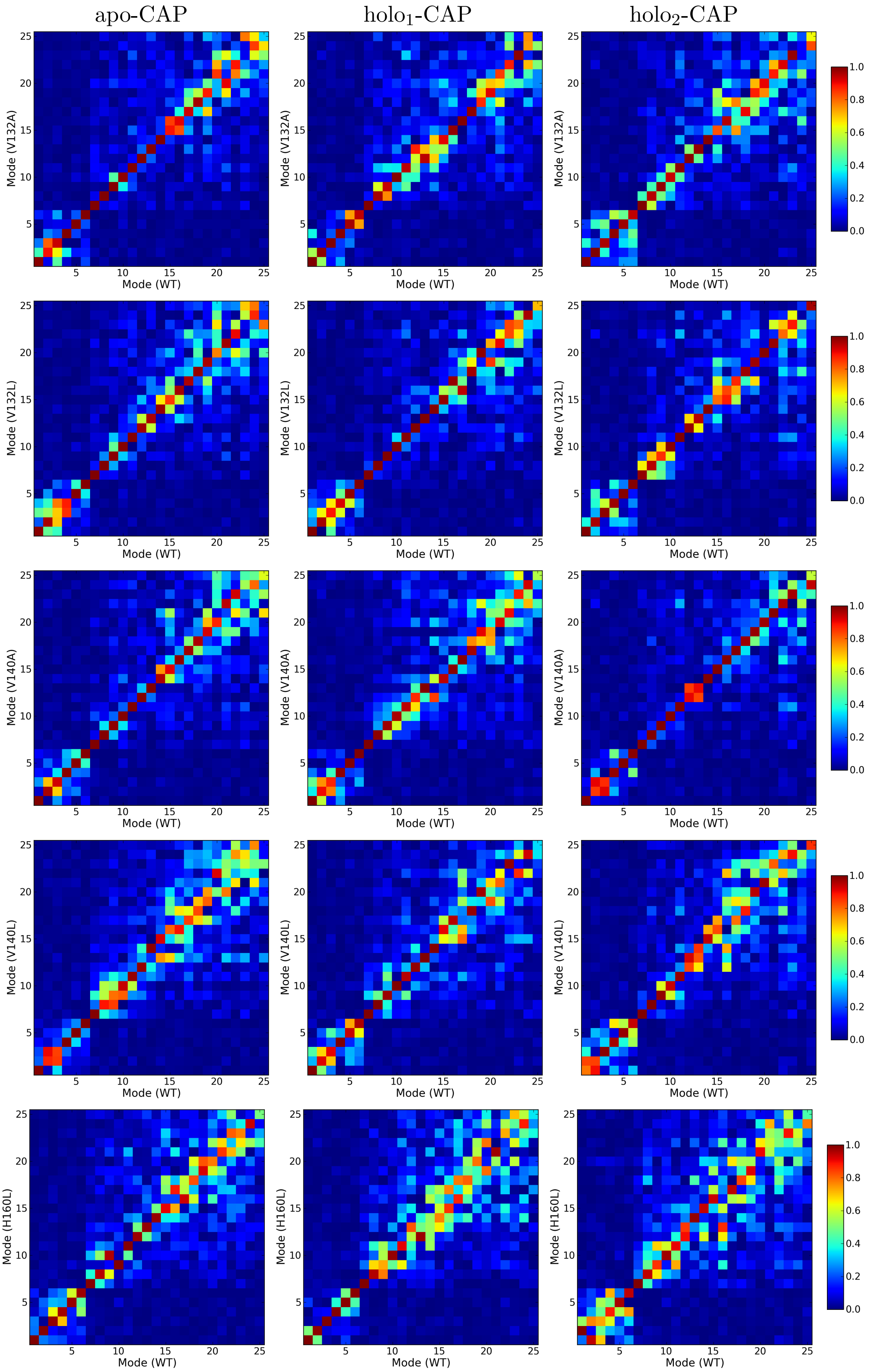


Figure 5.10: Overlaps of normal modes from WT-CAP vs. normal modes from the other variants for apo, holo$_1$ and holo$_2$-CAP. The order of the variants from top to bottom is: V132A, V132L, V140A, V140L and H160L.

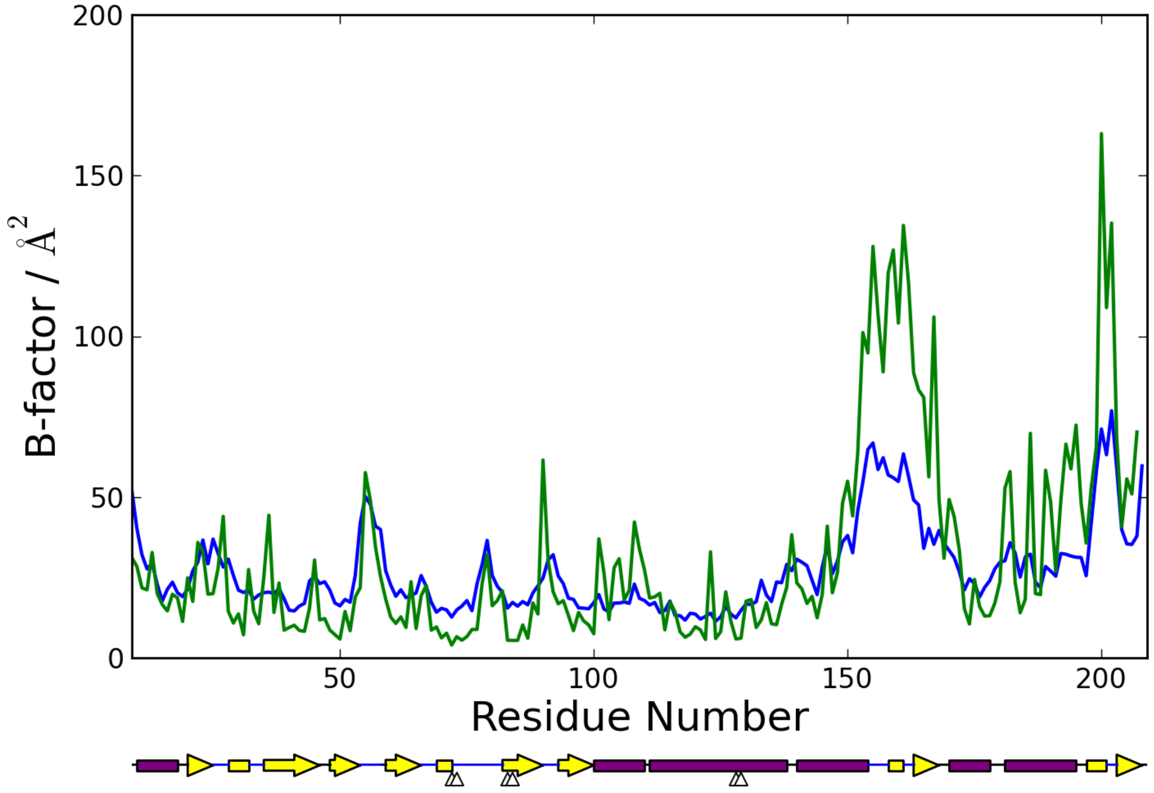


Figure 5.11: Comparison of experimental B-factors to the B-factors calculated from the MD simulation (for chain A of WT holo$_2$-CAP). The B-factors calculated by MD are scaled by a factor of $\frac{1}{3}$ for the plot for clarity. The secondary structural elements are displayed below the plot, with α-helices displayed as purple rectangles and β sheets displayed as yellow arrows. The positions of the residues that form hydrogen bonds with cAMP are indicated with triangles (△).

of WT holo$_2$-CAP in figure 5.11, with the experimental B-factors for comparison. The calculations show a fairly good agreement with the experimental B-factors. Evidently, the B-factors from the simulations have a much greater magnitude than those from experiment, however this is to be expected as the experimental B-factors are from the confined crystal structure where the molecule is not as free to move as it is in solution.

The effect of the increased motion when the protein is released from the crystal confinement can be seen when observing how the RMSF varies as the simulation progresses (see figure 5.12). At the start of the simulations the RMSF values are at their maximum, and as the simulation progresses, the RMSF values settle down and become more consistent between 20 ns windows towards the end of the simulation.

It can also be observed that for WT holo$_2$-CAP, one of the DNABDs undergoes larger fluctuations that the other, especially at the start of the simulation. This eventu-

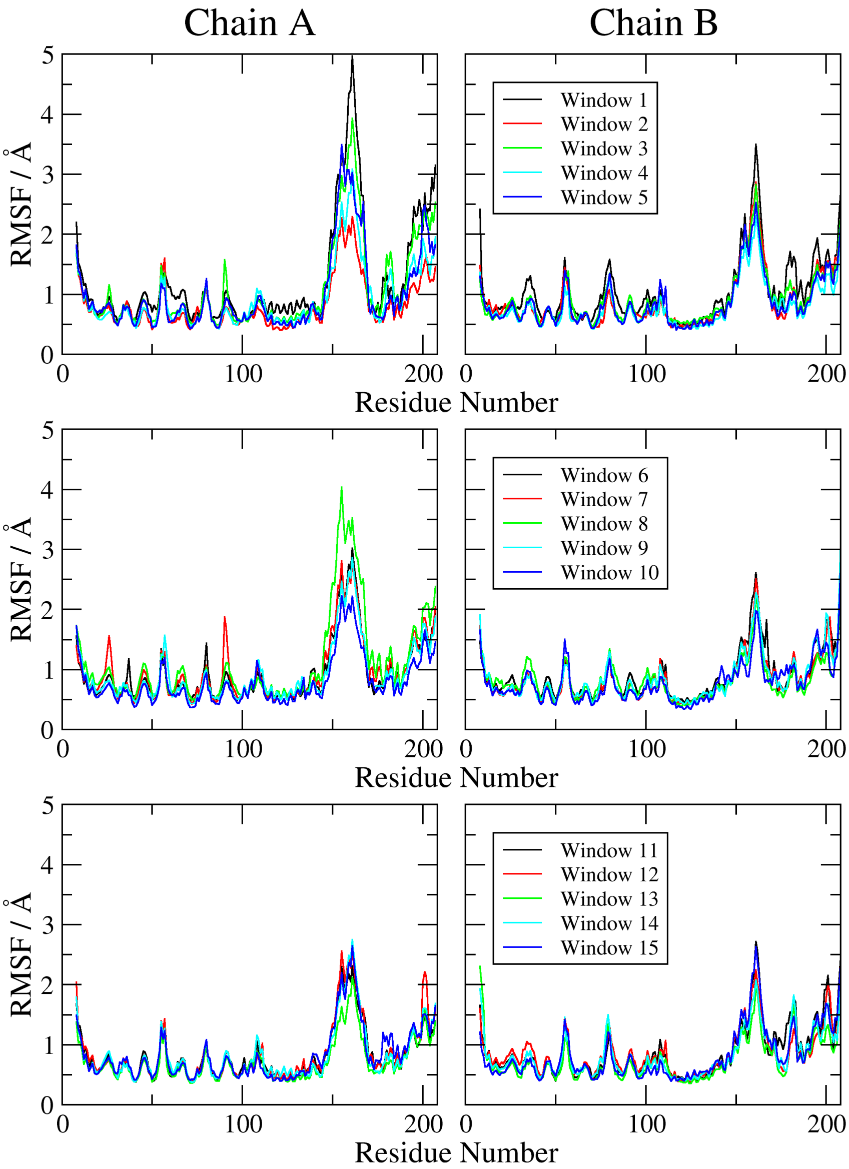
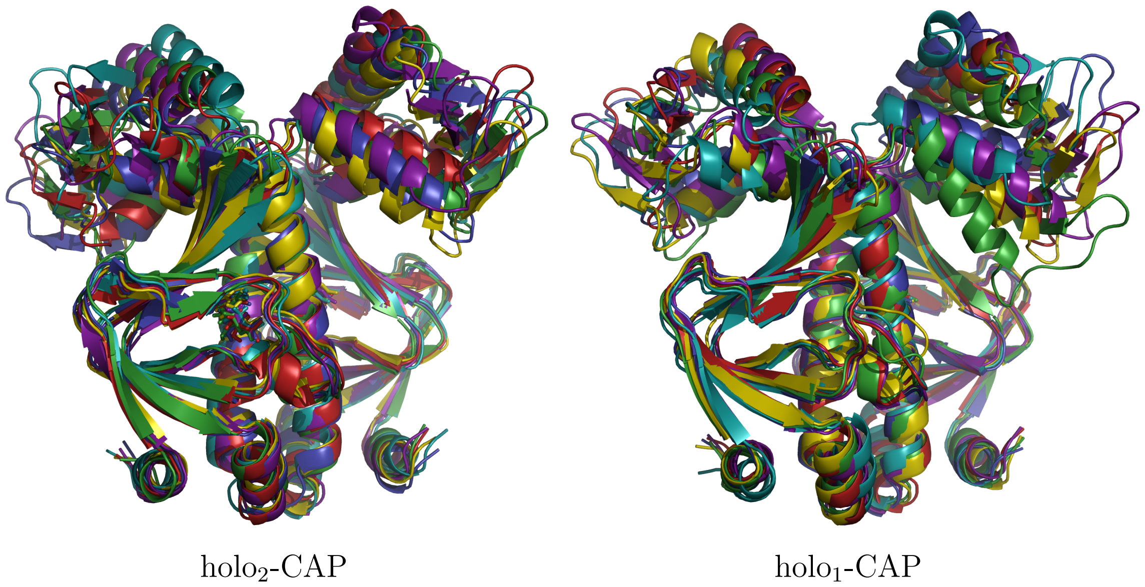


Figure 5.12: The progression of the per-residue RMSF for every 20 ns window throughout the simulation. The different chains are plotted separately and the windows are plotted on three sub plots of five windows for clarity.



holo$_2$-CAP

holo$_1$-CAP

Figure 5.13: Final structure of holo$_2$- and holo$_1$-CAP after 300 ns of simulation for: 🟥 WT, 🟦 V132A, 🟨 V132L, 🟦 V140A, 🟩 V140L and 🟪 H160L. The structures are RMS fit to the corresponding final structure for WT-CAP, using the PyMOL [234] implementation of the Kabsch Algorithm, using only the ligand binding domains for the fit.

ally diminishes; with the RMSFs of the two DNABDs becoming more similar towards the end of the simulation. The differences between the RMSFs of the LBDs is much more subtle, if at all. The differences between the DNABDs of each monomer is most likely due to the crystal packing, leaving the two DNABDs in slightly different orientations before the simulation is even started.

Throughout the simulation, there is very little change to the magnitude of the fluctuations in the LBD. This is because the structure of the LBD does not change much throughout the simulations. This can be seen to be the case when comparing the final structures after 300 ns of simulation of the CAP variants studied. For holo$_2$-CAP and holo$_1$-CAP, these are shown in figure 5.13, where it can be seen that there is very little change to the structure of the LBD between each of the variants. The largest change to the structure appears to be a rigid body rearrangement of the DNABDs with respect to the LBDs. The changes in structure upon binding will be investigated further, later in this section.

For apo-CAP there is also very little change to the structure of the LBD upon the mutations being made. Also, for most of the systems the structural differences between the DNABDs appear to be rigid body movements about the hinge region between the LBD and the DNABD. Only WT undergoes a large structural rearrangement of

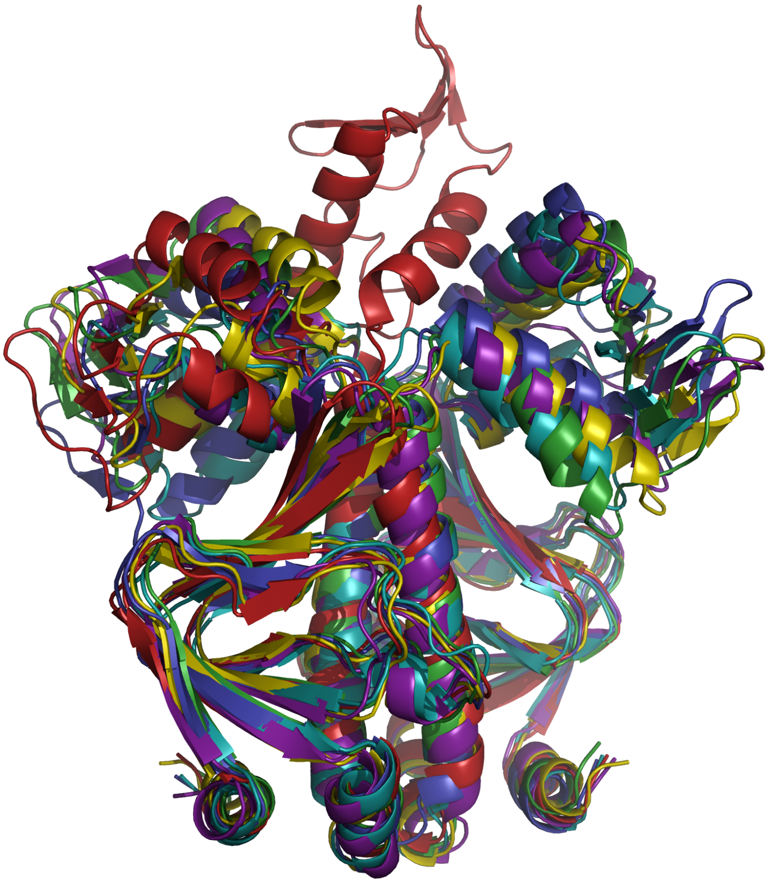


Figure 5.14: Final structure of apo-CAP after 300 ns of simulation for: ■ WT, ■ V132A, ■ V132L, ■ V140A, ■ V140L and ■ H160L. The structures are RMS fit to the corresponding final structure for WT-CAP, using the Kabsch algorithm with the backbone atoms of the ligand binding domains for the fit [232]. This was performed in PyMOL [234].

one DNABD when compared to the other variants (and bound states of WT-CAP). This structural rearrangement is discussed in greater detail later in this section. The structure of the LBD, however, even in this WT system does not change much in comparison to the other variants.

Table 5.6 shows the RMSDs of the LBDs for the final structures of the variants against the final structure of WT-CAP as well as the RMSDs including the DNABD in the calculation. This table confirms that there is very little structural change to the LBDs when the mutations are made. This is the case for apo, holo$_1$ and holo$_2$-CAP, as can be seen from the low RMSD values between the LBDs of the variant systems and equivalent WT systems. This can especially be seen for holo$_1$ and holo$_2$-CAP, where the RMSD between the LBDs does not exceed 1.53 Å. For apo-CAP, the RMSD between the variants and WT is slightly larger. This is most likely due to changes in the shape of the large central alpha helices near the hinge joining the LBD to the DNABD.

	RMSD LBD / Å			RMSD CAP / Å		
	apo	**holo$_1$**	**holo$_2$**	**apo**	**holo$_1$**	**holo$_2$**
WT	0.0	0.0	0.0	0.0	0.0	0.0
V132A	2.31	1.53	1.28	12.24	3.58	2.65
V132L	1.76	1.17	1.21	12.35	2.28	4.03
V140A	1.82	0.98	0.96	11.24	2.46	3.69
V140L	1.96	0.95	1.35	12.68	3.77	2.36
H160L	2.16	0.92	0.94	11.83	1.86	3.51

Table 5.5: RMSDs between the variants and WT-CAP for the three bound states of CAP. The RMSD alignment was performed on just the backbone atoms of the LBDs of CAP (up to residue 138), using the PyMOL [234] implementation of the Kabsch algorithm [232]. RMSD values are shown for the LBD alone and the full length CAP.

The differences between the RMSDs, however are still fairly small and definitely do not support a structural change of this domain.

The RMSD between the variants and WT-CAP for holo$_1$ and holo$_2$-CAP, when taking the entire CAP protein into consideration is still fairly small, not exceeding 4 Å by much in any instance. The RMSD for full length CAP is greater than the RMSD for just the LBD in all instances. This is again consistent with the view that the change in structure is a small rigid body movement of the DNABDs.

However, for apo-CAP, the RMSD between the variants and WT is much larger. This is not due to large differences between the variant structures, but due to the large rearrangement of the DNABD in WT-CAP. To tackle the problem that arises when comparing these variant structures to WT apo-CAP and also to see how the structures of each variant differs between their three bound states, comparison of the structures of the three bound states for each variant was then performed separately.

Again, as can be seen in figure 5.15, there is very little structural change in the LBD between the three bound states for each variant. This can be confirmed by observing the low RMSD values for this domain in table 5.5. The only structural differences between the three bound states are again only small rigid body rearrangements of the DNABD for all variants except WT-CAP. For WT-CAP, only small differences are seen between holo$_1$ and holo$_2$-CAP, however, apo-CAP exhibits a much larger rearrangement of one of the DNABDs. This is exemplified in the RMSD of apo to holo$_2$-CAP, which is 10.62 Å for full length WT-CAP.

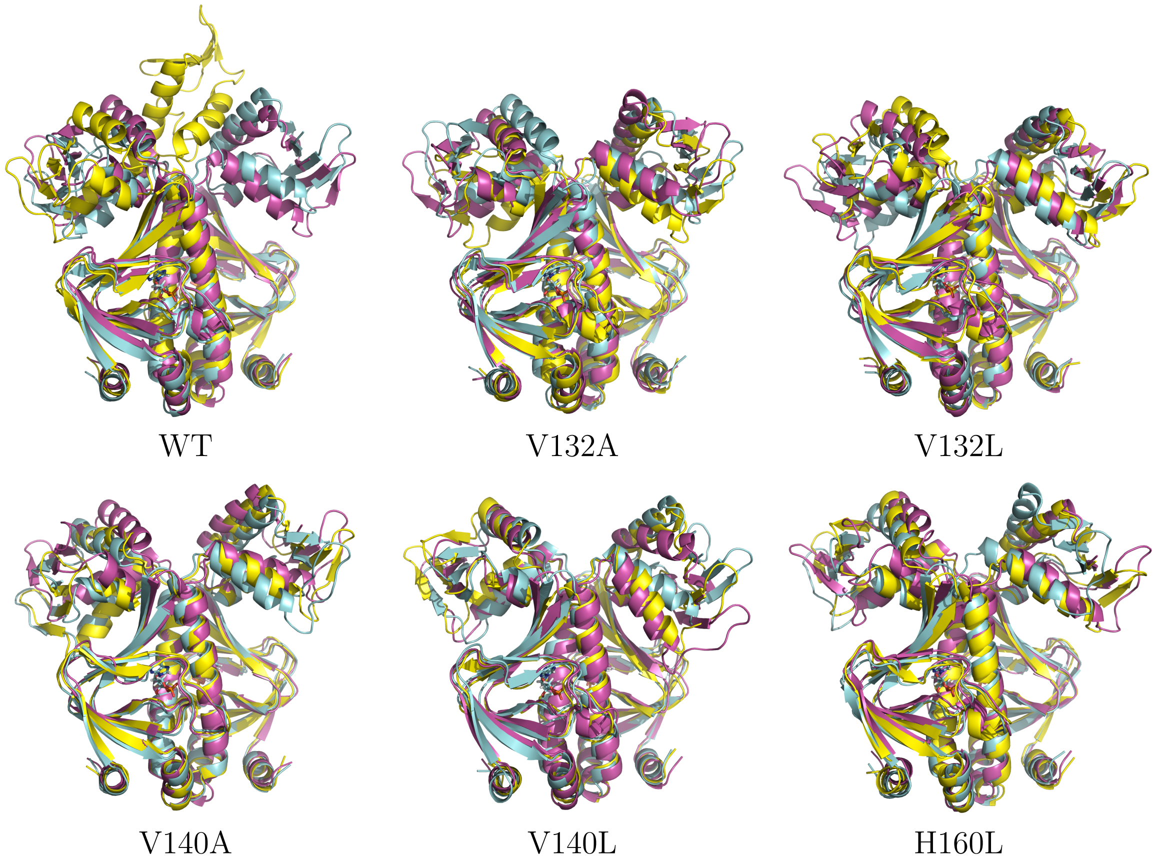


Figure 5.15: Structural changes between apo, $holo_1$ and $holo_2$-CAP for each variant: 🟨 apo-CAP, 🟪 $holo_1$-CAP and 🟦 $holo_2$-CAP. Structures were fit in PyMOL using the Kabsch algorithm [232] to fit the LBDs.

	RMSD LBD / Å		**RMSD CAP / Å**	
	apo	**$holo_1$**	**apo**	**$holo_1$**
WT	2.01	1.49	10.62	2.66
V132A	1.45	1.40	3.96	3.64
V132A	1.41	1.45	4.95	3.77
V132A	0.92	1.19	2.58	4.46
V132A	1.93	1.61	2.89	3.60
V132A	1.52	1.57	3.63	4.37

Table 5.6: RMSDs of apo and $holo_1$ cap against $holo_2$-CAP for all mutations of CAP. The RMSD alignment was performed on the backbone atoms of the LBDs of CAP, and RMSD values determined for the LBD alone and the full length CAP. The calculations were performed in PyMOL [234] using the Kabsch algorithm [232].

Looking at the rotation and translation that is undergone by the DNABD, by fitting the LBD of the protein to the same for holo$_2$-CAP, the centre of mass of the DNABD translates by 22.93 Å with the domain rotating 83.59° with respect to the LBD. The hinge is effectively at the very top of the large central α helix. The location of this rotation axis and the extent of the rotation is shown in figure 5.16.

When compared to the movement of the other DNABD, the centre of mass only moves 5.31 Å, corresponding to a rotation angle of only 5.52°. However the rotation hinge is much further away near the base of the long α helix in this case, as can be seen in figure 5.16. These hinges and the extent of rotation were calculated by treating both LBDs as one large rigid domain and each DNABD as separate rigid domains, and then calculating the rotation relative to the final structure of WT holo$_2$-CAP. These domains were defined by finding rigid regions of the two structures that had an RMSD below the tolerance of 4.5 Å. The rotation axis was then calculated by finding the angle difference between the LBDs of holo$_2$-CAP and the DNABD investigated for apo and holo$_2$-CAP. This analysis was performed using the hingefind plugin of VMD [214, 233].

The MD simulations were analysed further, using a number of different techniques. Firstly, the role of residue Thr128 in the allosteric signalling was investigated (see section 5.2.5). Then, PCA was performed on the entire simulation as well as smaller windows of the simulation (see section 5.2.6). Free energy calculations were performed using MM/PBSA [130, 180] and NMA for the conformational entropy to try to determine the free energy of binding for each binding event (see sections 5.2.7 and 5.2.8).

5.2.5 Investigating the Role of Thr128 in the Allosteric Signalling of CAP

While investigating the hydrogen bonds between Thr128 and cAMP, it became apparent that the hydrogen bond occasionally got broken in holo$_2$-CAP (see table 5.2). Upon further investigation, it was apparent that there were two conformations that each Thr128 residue could be found in, which can be described by the dihedral shown in figure 5.17). This dihedral will be labelled χ_a for chain A and χ_b for chain B, where chain A is the chain that binds cAMP first (in holo$_1$-CAP). When the dihedral is referred to for either chain, the notation χ_x will be used.

The most prominent dihedral angle that these atoms are found in for holo$_2$-CAP

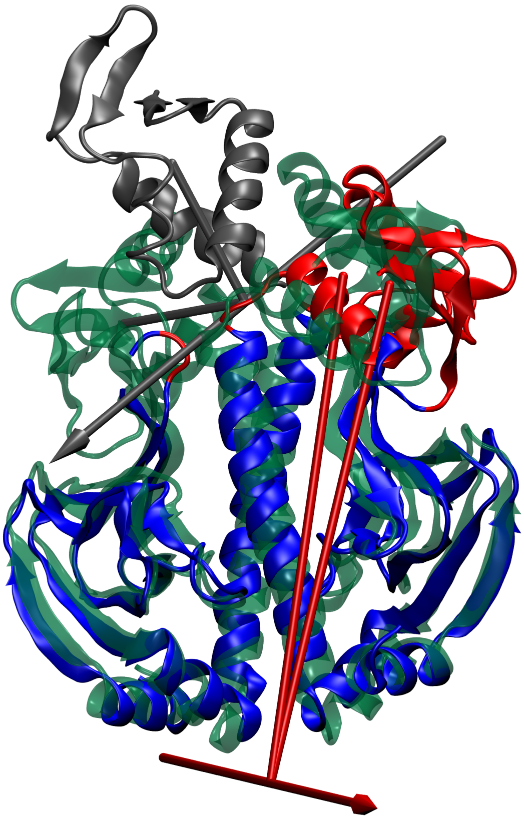


Figure 5.16: Rotation axis (arrows) and the angle of rotation observed in the apo-CAP DNABDs, when compared to the structure of holo$_2$-CAP (translucent green structure). The structure apo-CAP is split into three domains, one made of the LBDs (blue), and two comprising of each DNABD (grey and red). The arrows indicate the axis of rotation and the rods indicate the extent of the rotation.

is around -60°, but they can also be found occasionally with a dihedral around 60° in some of the simulations. The dihedral distributions for each conformation actually average to -50.25° and 50.09°, for WT holo$_2$-CAP, but they will be referred to as -60 and 60 for clarity. In all of the available crystal structures of holo$_2$-CAP in the PDB, only the χ_x conformation is found. With the dihedral in Thr128 from each chain, being able to have two different conformations, the system has four possible states.

Figure 5.18 shows these four states in their local environment within CAP. As it can be seen, when $\chi_x = -60°$ and cAMP is present, Thr128 is able to form a hydrogen bond with cAMP and it appears as if both chains are in this orientation, the C_γ methyl groups of Thr128 may be oriented in a way where they form favourable hydrophobic interactions with each other and the Val132 residue. However, in the $\chi_x = 60°$ orientation, the hydrogen bond to cAMP cannot be formed and it appears as if fewer hydrophobic interactions will occur across the interface. When $\chi_x = 60°$

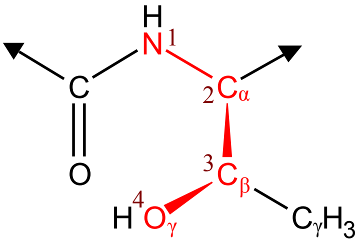


Figure 5.17: The dihedral used to determine the conformation of Thr128, highlighted in red, with the order used in the dihedral calculation indicated.

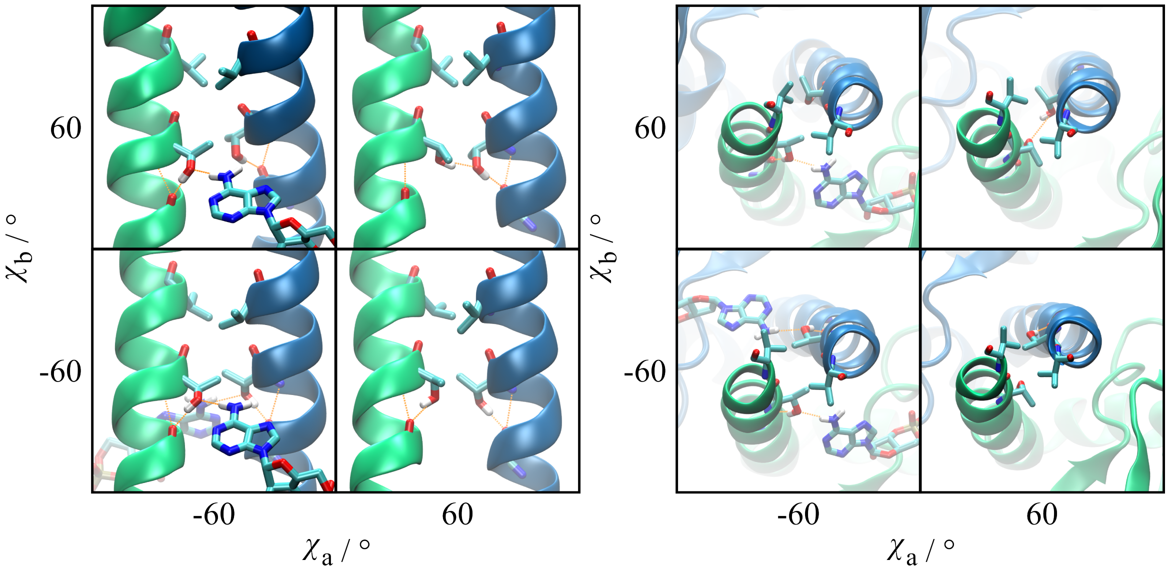


Figure 5.18: The two different conformations of Thr128 leads to four different states when both chains are taken into account. Shown are images of each of these four states, firstly from the side of the two α helices on the dimer interface, and secondly from the top, looking down the length of the α helices.

for both chain A and B, the Thr128 sidechain can be rearranged in such a way that a new hydrogen bond can form across the interface. This however occurs by breaking a hydrogen bond internally in the monomer with residue Arg124.

Histograms of Thr128

To observe which orientation is most prominent for each Thr128 in each bound state, histograms of χ_a and χ_b were plotted (see figure 5.19).

For WT holo$_2$-CAP (figure 5.19 (a)), χ_x for both chains is almost exclusively -60°. For the simulation of WT holo$_1$-CAP, it is apparent that the $\chi_b = 60°$ occurs more frequently. However, χ_a remains in the -60° orientation. As the cAMP is still bound to chain A, but not chain B in this orientation, this implies that Thr128 for chain B is more free to explore the other orientation. For apo-CAP, the cAMP from chain A is

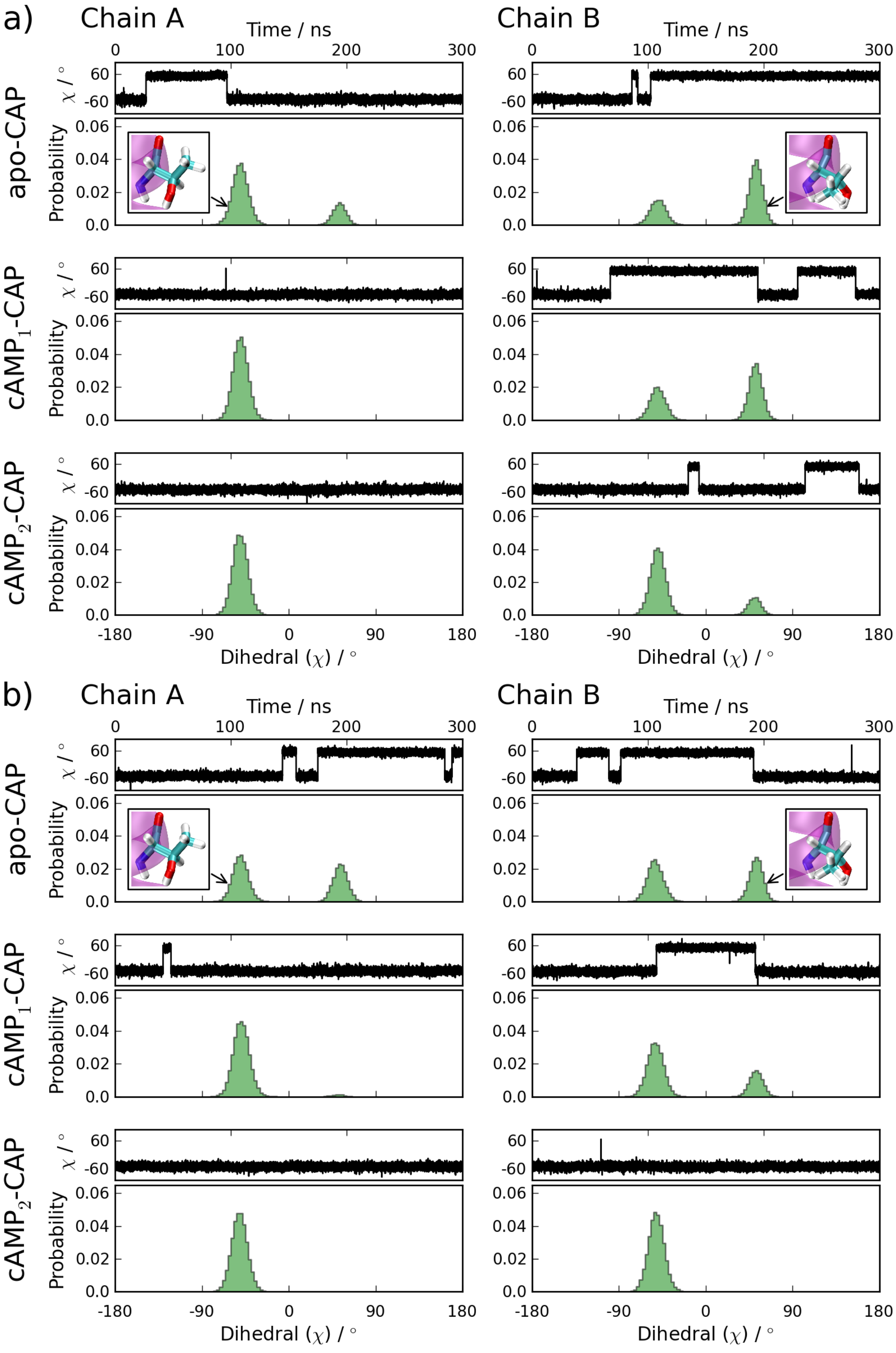
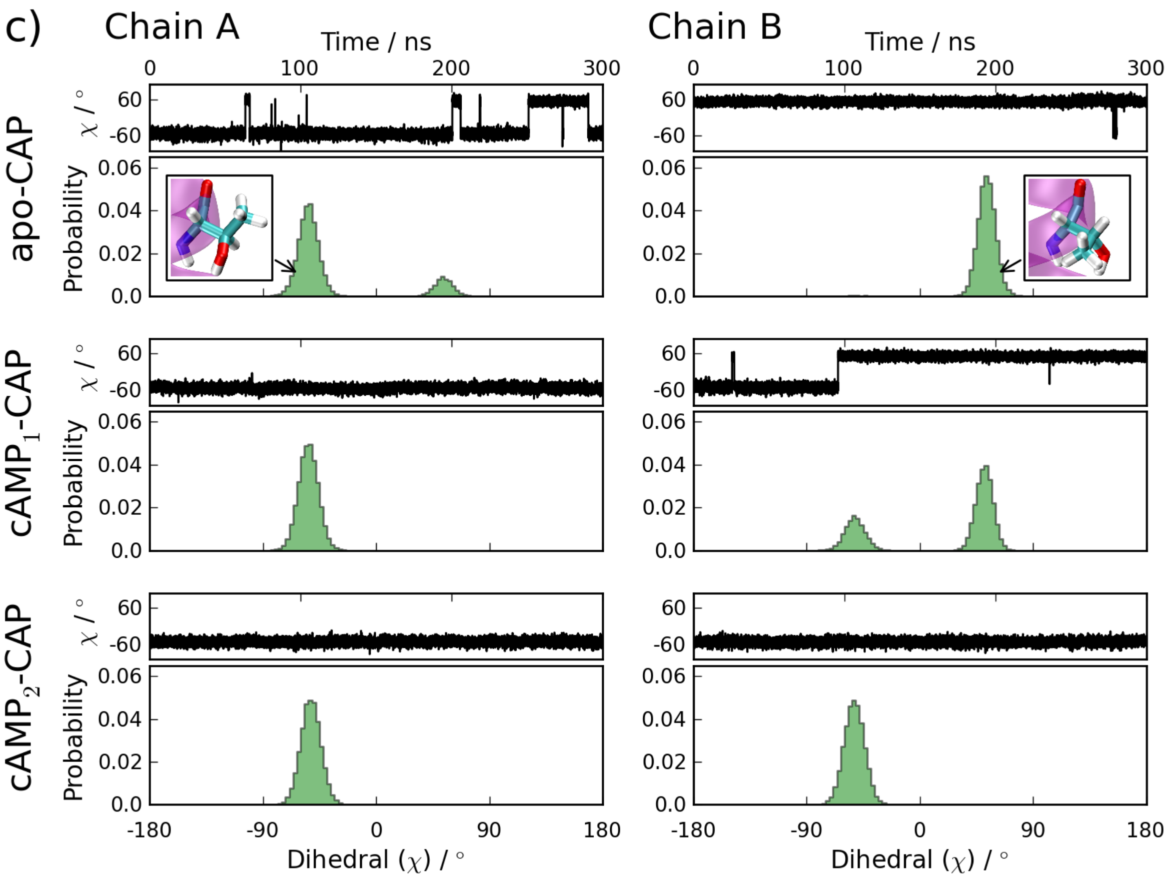


Figure 5.19: Histograms of the dihedral space for residue Thr128. The histograms for the individual chains are shown separately. Variants shown are a) WT and b) V132A.



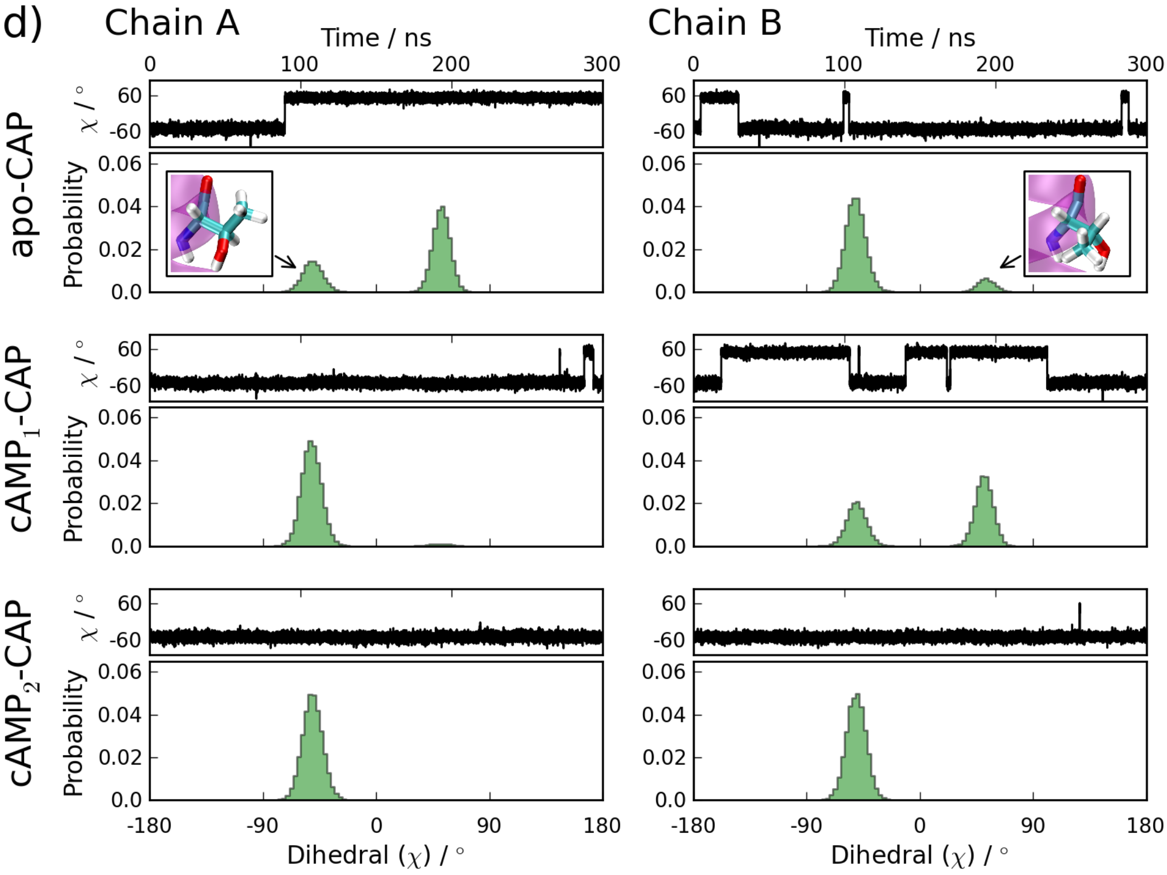
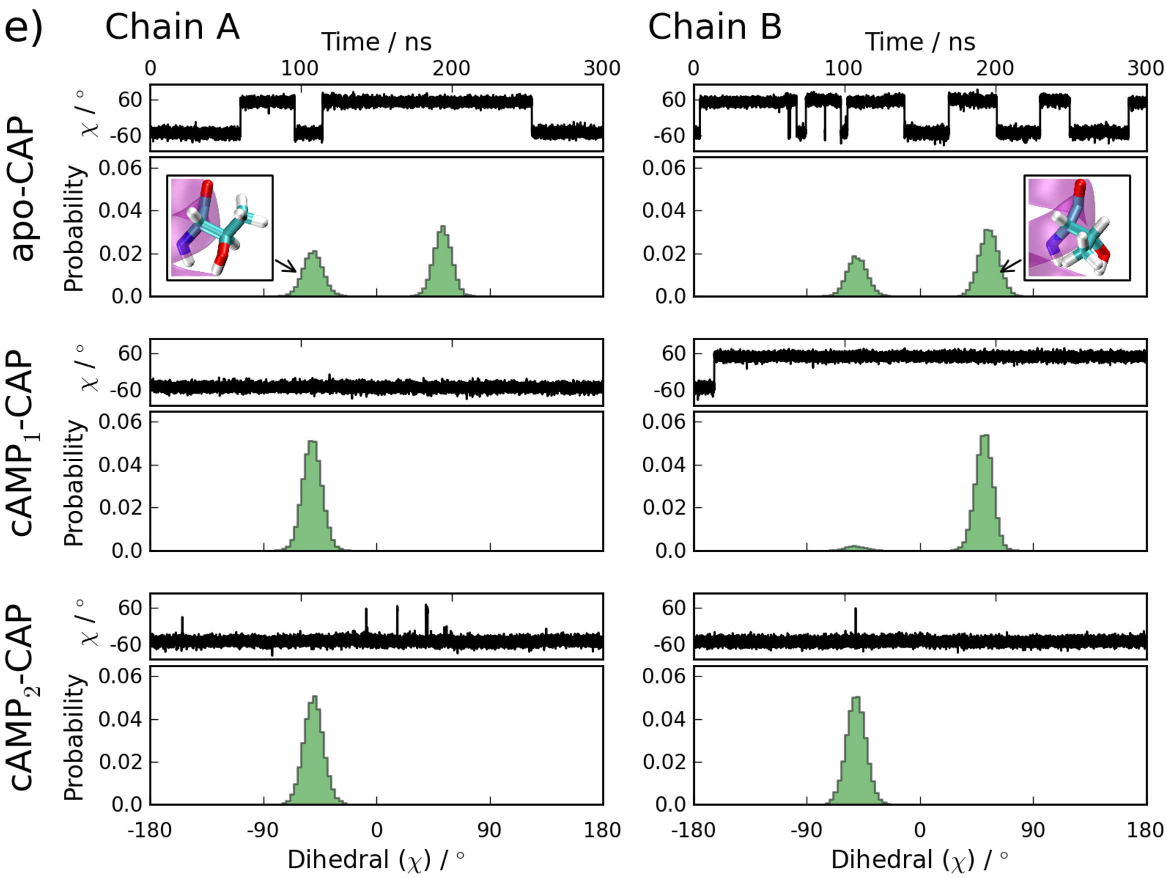


Figure 5.19: Histograms of the dihedral space for residue Thr128. The histograms for the individual chains are shown separately. Variants shown are c) V132L and d) V140A.



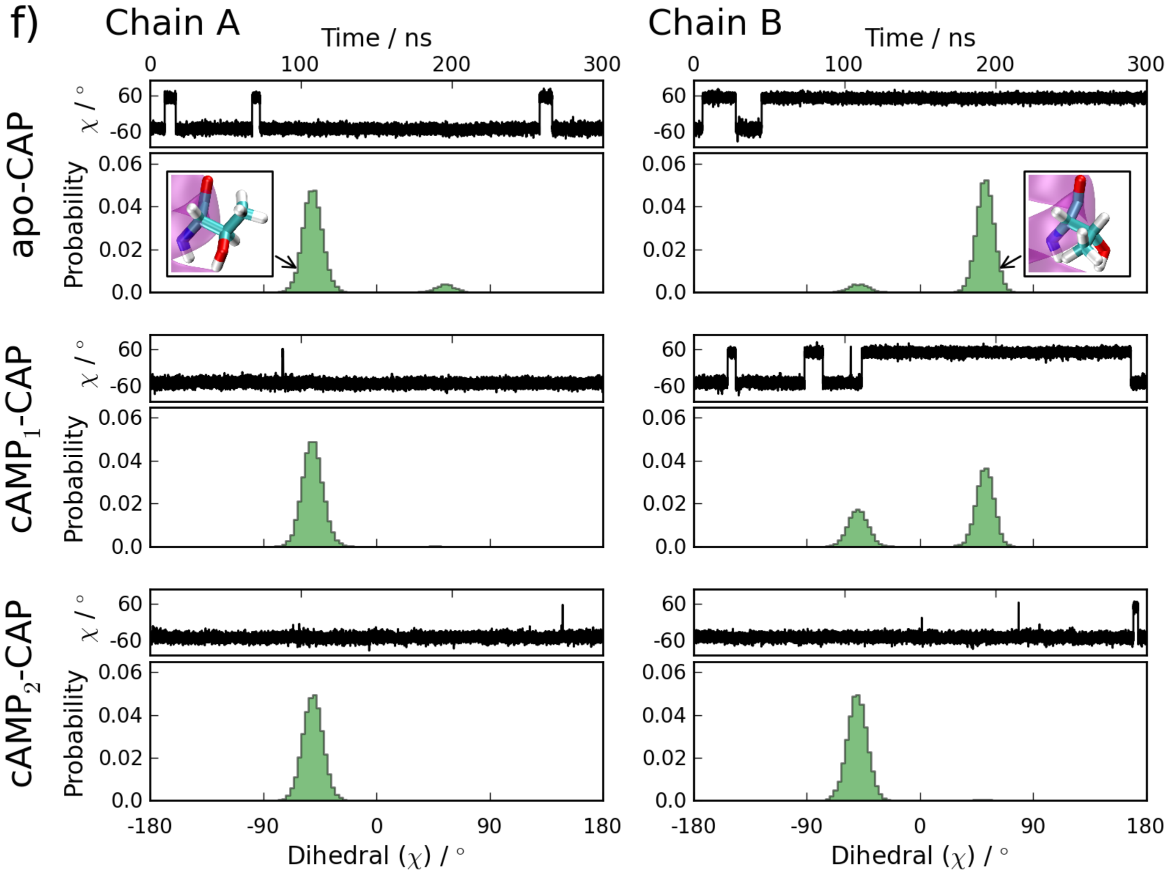


Figure 5.19: Histograms of the dihedral space for residue Thr128. The histograms for the individual chains are shown separately. Variants shown are e) V140L and f) H160L.

removed, and in this simulation, both $\chi_\mathrm{x} = -60°$ and $\chi_\mathrm{x} = 60°$ are sampled for both chains. This would imply that in the absence of cAMP, the Thr128 residue side chain is more likely to rotate along the C_α-C_β axis to form different interactions across the dimer interface in these simulations of WT-CAP.

When observing the preferred values of χ_a and χ_b for the holo$_2$ simulations of the other variants (figure 5.19 (b-f)), it is evident that when both cAMP are present, the strongly preferred orientation for Thr128 of both chains is when $\chi_\mathrm{x} = -60°$. For holo$_1$-CAP, the Thr128 residue in chain A, the chain to which cAMP is bound, also appears to be firmly held in the $\chi_\mathrm{x} = -60°$ orientation. This would imply that when cAMP is present, the hydrogen bonds between cAMP and Thr128 are strong enough to hold Thr128 in the $\chi_\mathrm{x} = -60°$ orientation.

Also, the population of Thr128 in chain B (the chain without cAMP) of holo$_1$-CAP in the $\chi_\mathrm{b} = 60°$ orientation increases for all variants when compared to holo$_2$-CAP. The distribution of $\chi_\mathrm{b} = 60°$ and $\chi_\mathrm{b} = -60°$ is fairly even for all variants except V140L, suggesting that for these variants, both configurations have similar energies. For V140L (figure 5.19 (e)), it could either be the case that the $\chi_\mathrm{b} = -60°$ configuration is unfavourable, or the barrier between the two configurations is large enough that the side chain cannot get out of the $\chi_\mathrm{b} = 60°$ orientation. As the system does not jump between the two configurations except once into $\chi_\mathrm{b} = 60°$, the second case is the most likely.

For apo-CAP, the variants V132A, V140A and V140L appear to sample both $\chi_\mathrm{x} = 60°$ and $\chi_\mathrm{x} = -60°$ in both chains. This again shows that the presence of cAMP stabilises the $\chi_\mathrm{x} = -60°$ orientation and implies that both orientations have not too dissimilar energies when cAMP is absent. Only two variants; V132L and H160L appear to not follow this rule, which is explored when the four different states are investigated.

Free Energy Surface of Four States of Thr128

Rather than just plotting a one dimensional histogram, it is also possible to investigate which of the four states (in figure 5.18) are most favourable for each variant and bound state by plotting a two dimensional histogram, with χ_a and χ_b as the axes. It is also possible to plot the free energy surface of this dihedral space by performing a Boltzmann inversion on the histogram. This plot is shown for WT apo-CAP in figure

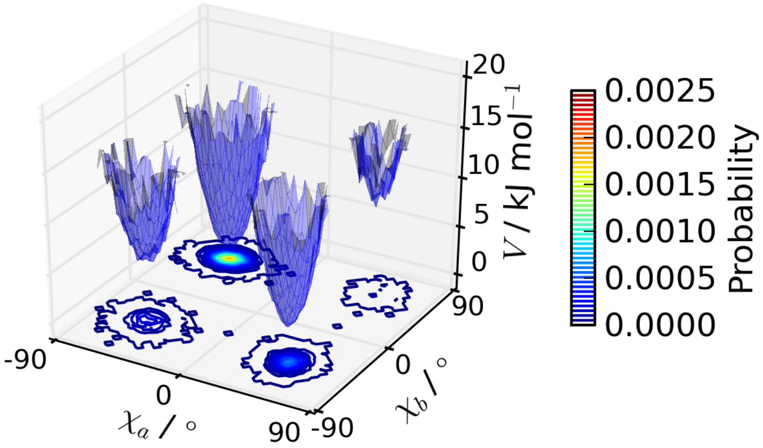


Figure 5.20: Histogram and free energy surface of the 2D dihedral space for Thr128 of both chains. The histogram is shown on the xy plane and has the scale shown in the the colour bar. The free energy surface is generated by performing a Boltzmann inversion on the histogram and normalising so that the deepest well has an energy of 0 kJ mol^{-1}.

5.20. The histogram is displayed as a flat contour plot on the base of the figure, and the calculated free energy surface shown above it. This system, for example samples all four possible states, with the most populated state being the $\chi_{\mathrm{a}} = -60|\chi_{\mathrm{b}} = 60$ state. This state therefore has the lowest energy on the energy surface. Figure 5.21 shows the histograms and energy surfaces for the 2 dimensional dihedral space for all of the variants and cAMP binding states.

These free energy surfaces again confirm that in the presence of cAMP, the preferred value of χ_{x} is -60° for all variants. Thr128 is also again shown to be more flexible in apo-CAP, with all four states on the dihedral being sampled for most variants. Only mutations V132L and H160L do not sample all four states on the dihedral space of apo-CAP. This may suggest a tightening in the region around this residue. These variants are the only two identified by the ENM to push the protein towards positive cooperativity of the variants studies. In the next chapter it is shown with ITC that these two mutations do in fact push the system towards positive cooperativity. This helps strengthen the argument that Thr128 is an important residue in the signalling pathway for the cooperative binding of cAMP.

It is easy to see why making the V132L mutation may hinder the rotation of the χ_{x} dihedral. Figure 5.22 shows that due to the location of the mutation being near the location of residue Thr128, the added bulk of the leucine most likely hinders rotation of the threonine compared to WT. Furthermore, mutating the valine to alanine as

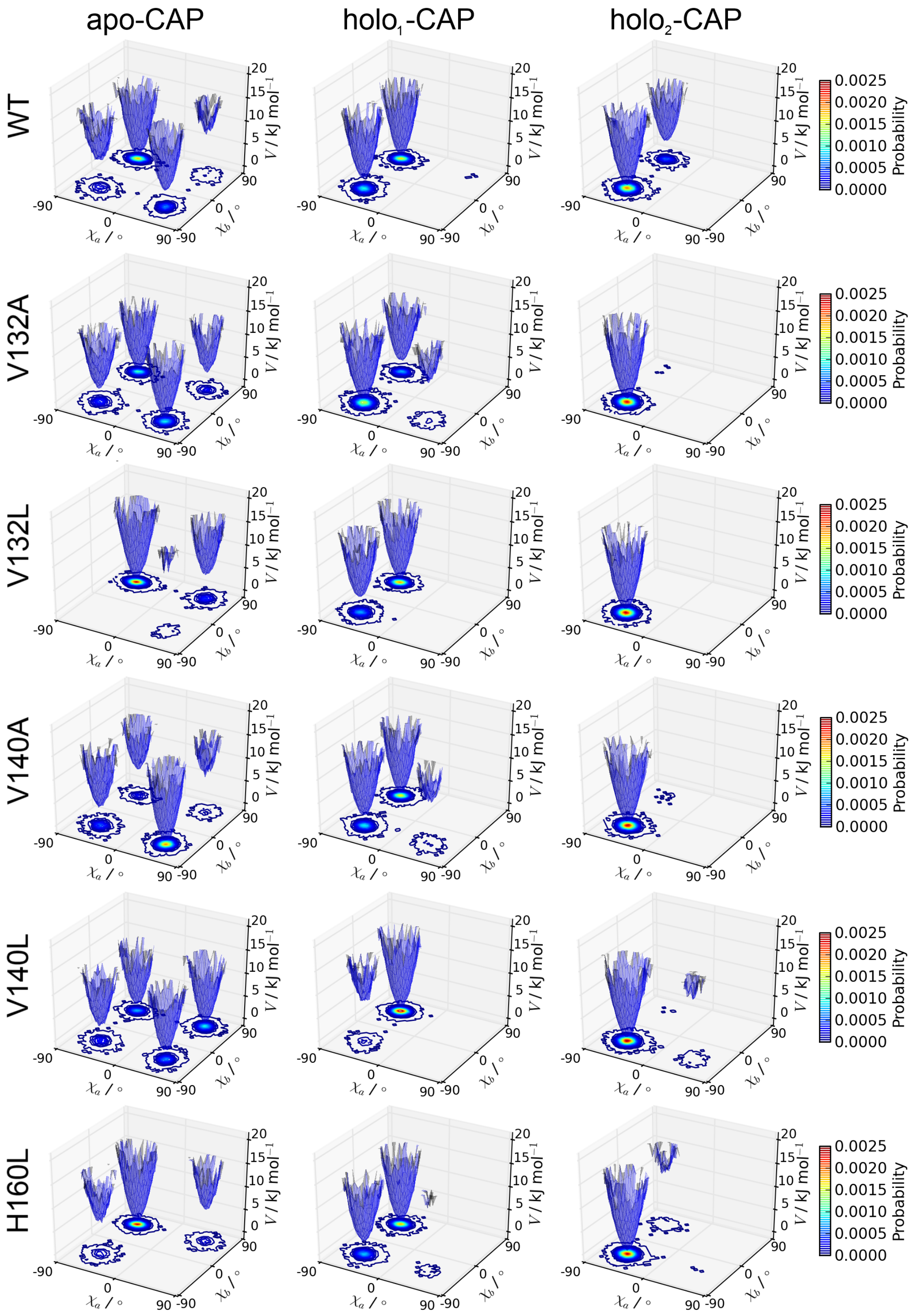


Figure 5.21: Histograms and free energy surfaces of the 2D dihedral space of Thr128 for all variants; WT, V132A, V132L, V140A, V140L and H160L. The three bound states: apo-CAP, holo$_1$-CAP and holo$_2$-CAP, were investigated for each variant.

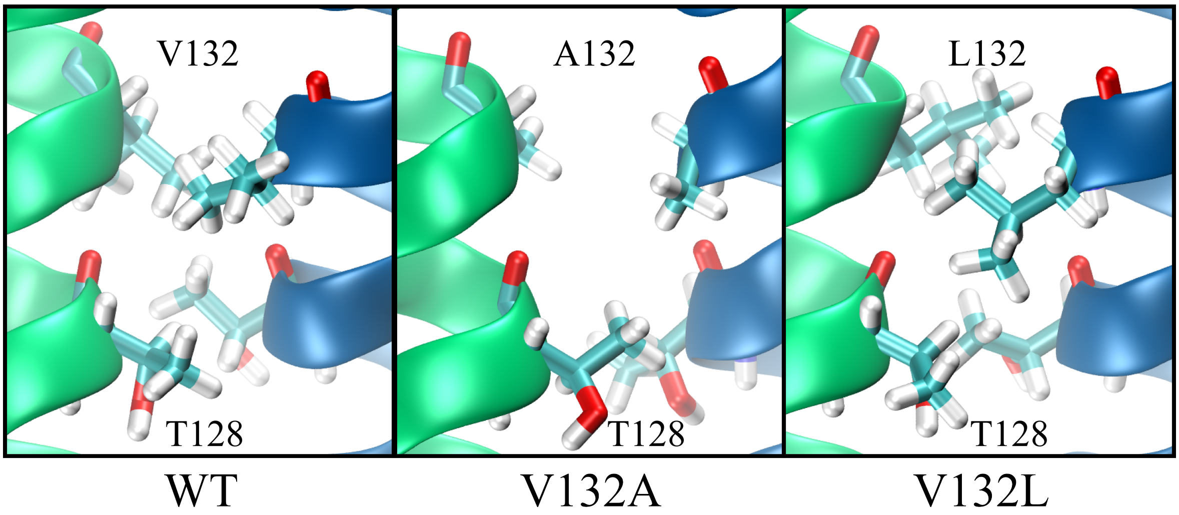


Figure 5.22: Mutating Val132 adjusts the hydrophobic interactions with Thr128. The figure shows the Thr128 and Xaa132 residues for WT, V132A and V132L. As well as altering the hydrophobic interactions V132L also gives residue 132 a lot more bulk, which could hinder rotation of residue Thr128.

in V132A, creates more space for rotation of the threonine. The effect of the H160L mutation is not as clear, as the site of the mutation is in the DNABD very distant from Thr128.

ff12SB Investigations

To ensure that the effect seen here was not a force field dependent effect, 200 ns simulations of WT-CAP were performed for the apo, holo$_1$ and holo$_2$ systems using the newer AMBER ff12SB force field. These simulations found that for all three bound states, the Thr128 residue was much less flexible, being constantly held with a dihedral of -60°, as can be seen in figure 5.23.

To investigate if this occurred because the energy barrier between $\chi_x = -60°$ and $\chi_x = 60°$ was larger, thus preventing the flip between the two dihedral angles, another 200 ns simulation of WT apo-CAP was performed, with a starting structure where $\chi_x = 60°$ for both chains. In the simulation, the dihedrals χ_a and χ_b, soon flipped and stayed in the $\chi_x = -60°$ orientation, as can be seen in figure 5.24. This suggests that it is not due to a large energy barrier that Thr128 stays in the -60° orientation, but that the energy of the 60° orientation is unfavourable with this force field.

Plotting the force field dihedral terms for this force field, it is evident that their is a large change in these terms, which destabilises the $\chi_x = 60°$ orientation. The barrier

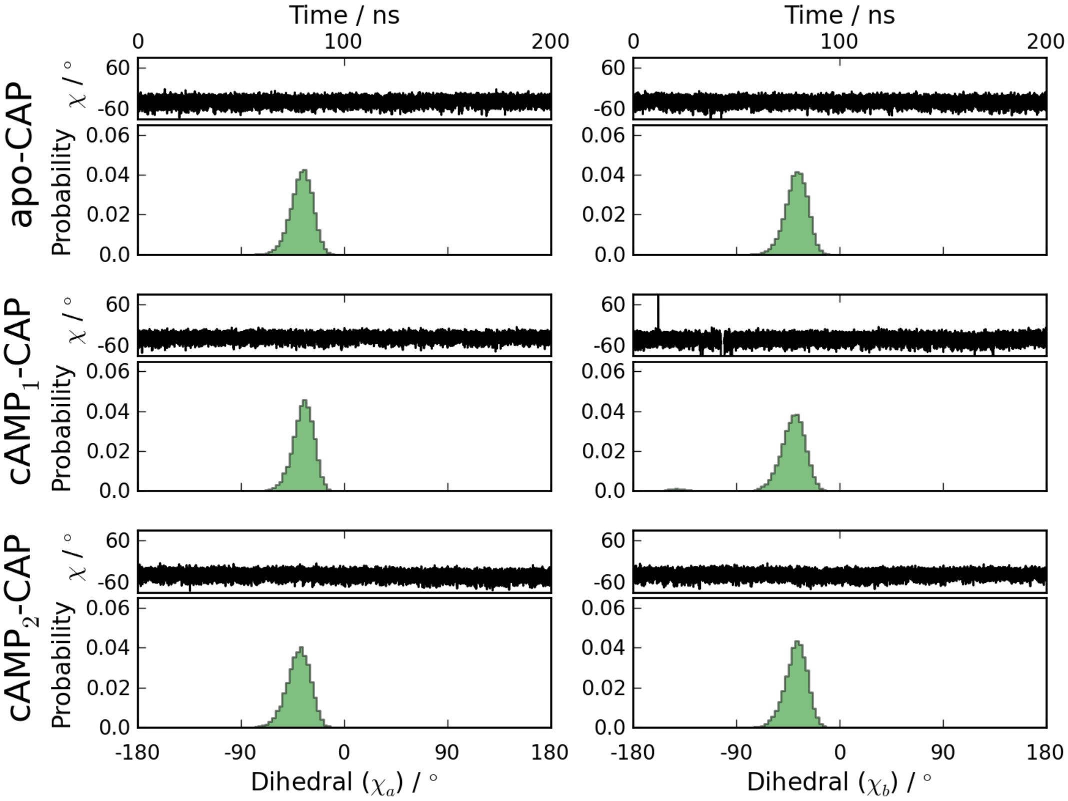


Figure 5.23: Histograms of the dihedral angle of the sidechain of Thr128 using ff12SB. The three bound states: apo, holo$_1$ and holo$_2$ are shown for WT-CAP.

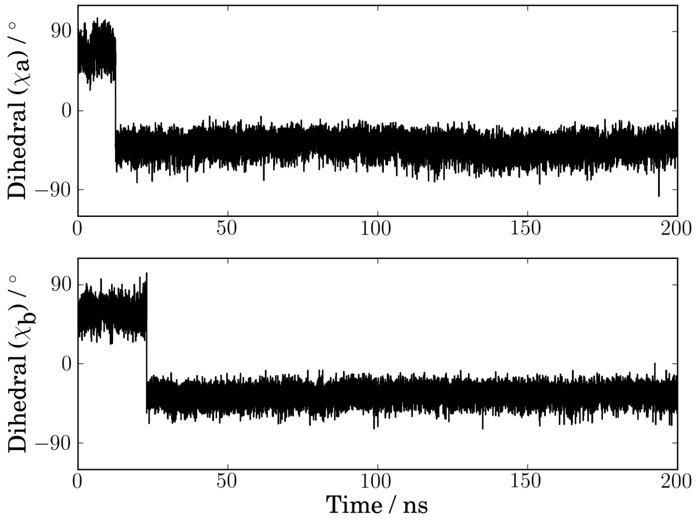


Figure 5.24: The dihedral of Thr128 over time with $\chi_x = 60$ as the starting structure.

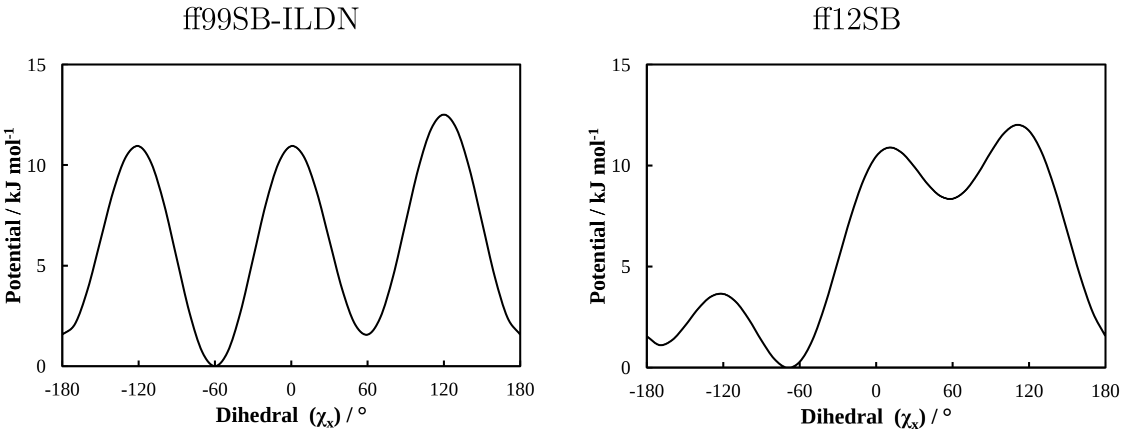


Figure 5.25: Dihedral functional forms for the N-C$_\alpha$-C$_\beta$-O$_\gamma$ dihedral in Thr128 with ff99SB-ILDN and ff12SB.

between the two energy wells, however, does not change much between the two force fields. Of course, this is only part of the picture, as the 1-4 van der Waals and 1-4 electrostatic terms should also be taken into account, which could further destabilise the $\chi_\mathrm{x} = 60°$ orientation or increase the barrier between the two orientations. It is likely due to the 1-4 van der Waals and 1-4 electrostatic terms that the $\chi_\mathrm{x} \approx 180°$ orientation is not observed in any of the simulations.

Unfortunately, this shows that the changes to the Thr128 residue upon binding of cAMP are not force field independent, and in fact the newest AMBER force field does not display any changes in the residue upon binding. It was shown in section 5.2.2 however that the ff12SB force field may not correctly simulate WT-CAP. Which could suggest that the signalling events, involving Thr128, seen in the ff99SB-ILDN simulations could in fact be an important pathway for the allosteric signal regardless of it not been seen in the ff12SB simulations.

5.2.6 Principal Component Analysis

The simulations were split into 20 ns windows, and PCA performed on each window to see how the principal components evolve throughout the simulation. Figure 5.26 shows the frequencies of the first 250 modes calculated using PCA on all of the 20 ns windows of WT apo-CAP. As it can be seen from the figure, the normal mode frequencies can change between windows. However, the lowest frequency normal modes tend to have a wavenumber of less than 1 cm^{-1}. This is on the timescale identified for hinge bend

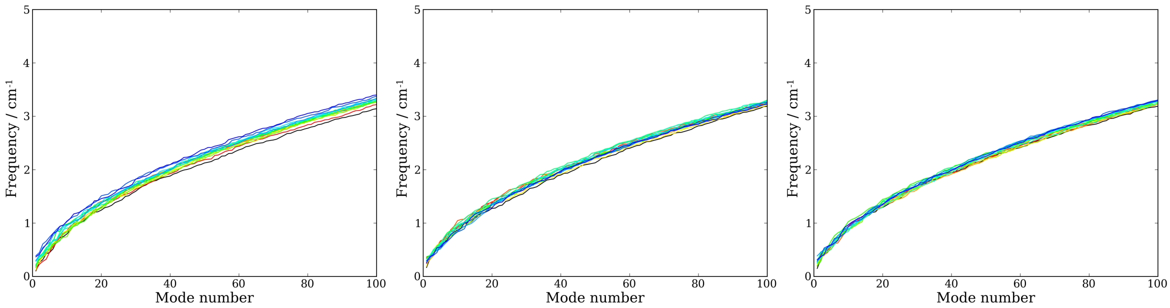


Figure 5.26: Frequency of the first 250 modes calculated using PCA on 20 ns windows of WT apo-CAP. — Window 1, — Window 2, — Window 3, — Window 4, — Window 5, — Window 6, — Window 7, — Window 8, — Window 9, — Window 10, — Window 11, — Window 12, — Window 13, — Window 14, — Window 15.

motions on domain interfaces [21]. The lowest frequency modes tend to be movement of the DNABDs relative to the LBDs about the hinge region, so fits well within this motion description.

The large differences between the frequencies of the normal modes for different windows of the apo-CAP simulation, suggests that the different windows are sampling different regions on the free energy surface. This would make sense due to the large structural rearrangement of the DNABD in this simulation. This could also be happening in the simulations of the other systems, but to a lesser extent. There is also a chance that diffusive motion is being sampled in all or some of the simulations, rather than correctly sampling the shape of an energy well. These hypotheses are investigated below, firstly by looking at how much the principal components resemble cosines; a trait shared with the principal components of diffusive motion [175]. And secondly by calculating the normalised overlaps between the covariance matrices of separate windows in each trajectory.

The cosine content of each mode was determined using the GROMACS simulation package, which uses equation 2.5.64 in section 2.5.6. These are shown for each window of each variant of CAP in figure 5.27. It is evident that for most of the 20 ns windows, the cosine content of the first PC is between 0.5 and 1, which would indicate that these PCs have a motion that can predominantly be described as diffusive motion. This means that for these windows, the protein moves fairly freely across the energy surface and is not kept within one quasiharmonic region of the energy surface.

As an example of the cosine shape of the PCs, the projection of the first window of the MD simulation of WT holo$_1$-CAP on the first four eigenvectors for the window

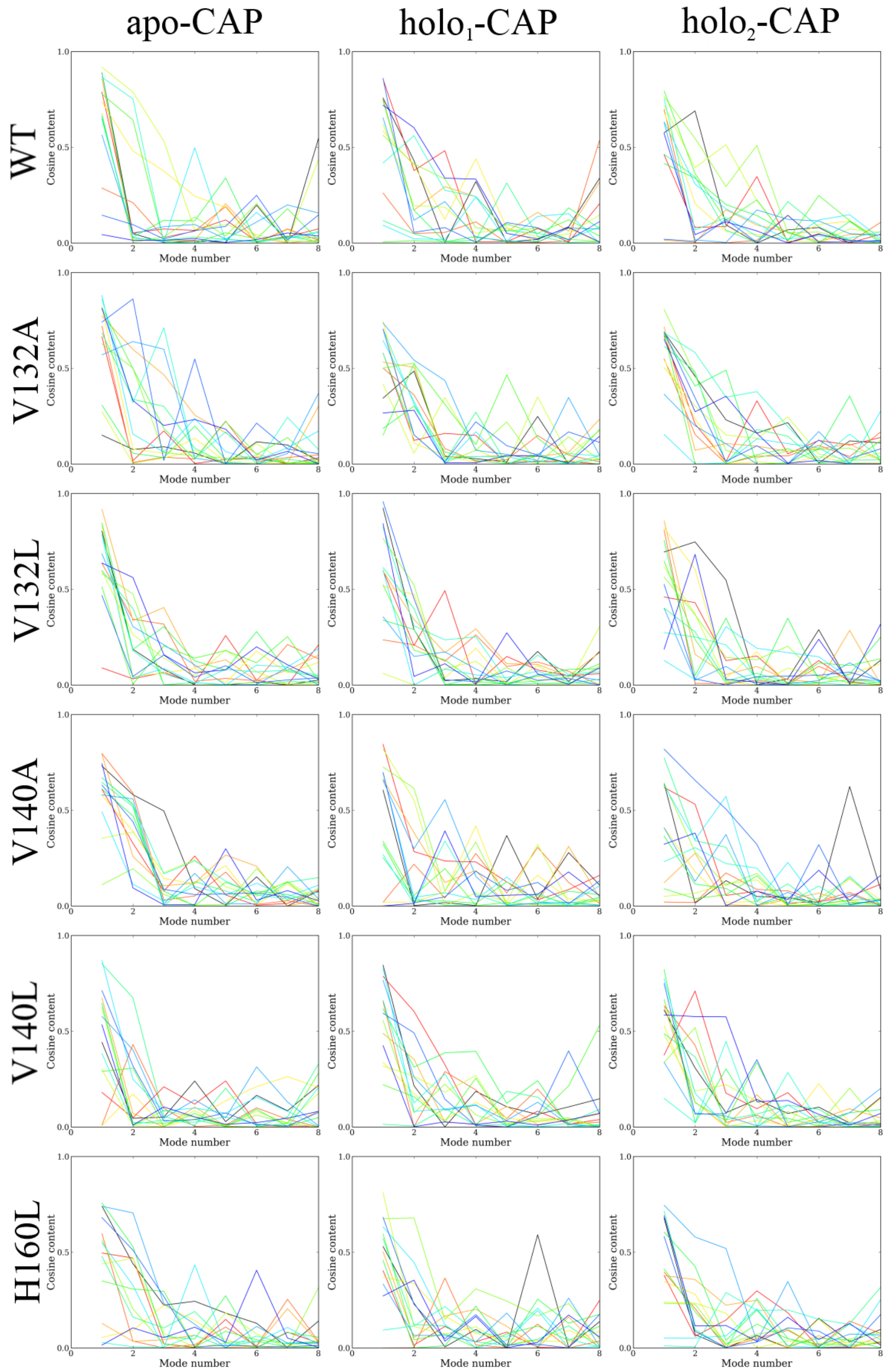


Figure 5.27: The cosine content for the first 8 Principal components calculated for each of the 20 ns windows; Shown for all three bound states of the variants of CAP. A cosine content around 1 for a PC indicates that it mainly samples diffusive motion. — Window 1, — Window 2, — Window 3, — Window 4, — Window 5, — Window 6, — Window 7, — Window 8, — Window 9, — Window 10, — Window 11, — Window 12, — Window 13, — Window 14, — Window 15.

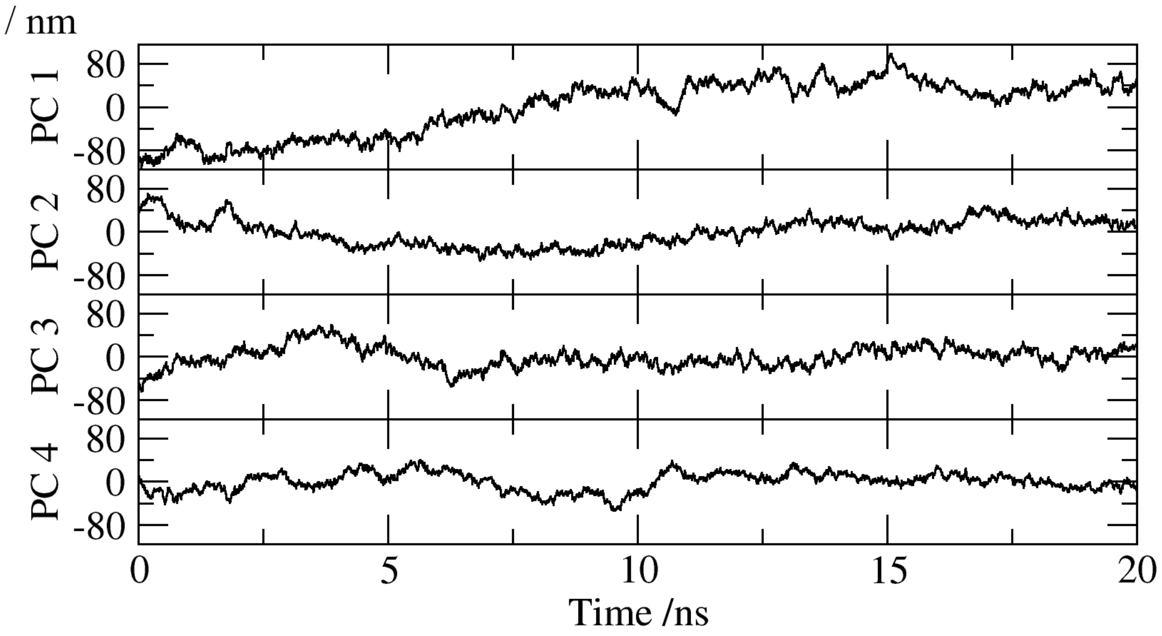


Figure 5.28: Projections of the MD simulation onto the first four PCs for the first window of the WT holo$_1$-CAP simulation. This window has a fairly high cosine content for the first PC (0.76), a low cosine content for PC 3 (2.5×10^{-3}) and cosine contents of 0.43 and 0.32 for PCs 2 and 4 respectively.

	apo-CAP	**holo$_1$-CAP**	**holo$_2$-CAP**
wt	0.212	0.223	0.249
v132a	0.214	0.236	0.234
v132l	0.225	0.228	0.239
v140a	0.246	0.264	0.263
v140l	0.247	0.233	0.246
h160l	0.247	0.255	0.270

Table 5.7: Average normalised overlaps between pairs of 20 ns windows of simulations for each system.

is shown in figure 5.28. This window was chosen as it shows PCs with varying degrees of cosine content, however most windows of all of the simulations could be shown to demonstrate at least one mode with a high cosine content. The cosine content of the first mode in this example is by no means the largest observed in any of the systems, however the 1/2 period cosine shape of the mode can still clearly be seen.

Next, the normalised subspace overlap was determined between each 20 ns window of a simulation using equation 2.5.62 in section 2.5.5 and plotted as a matrix map.

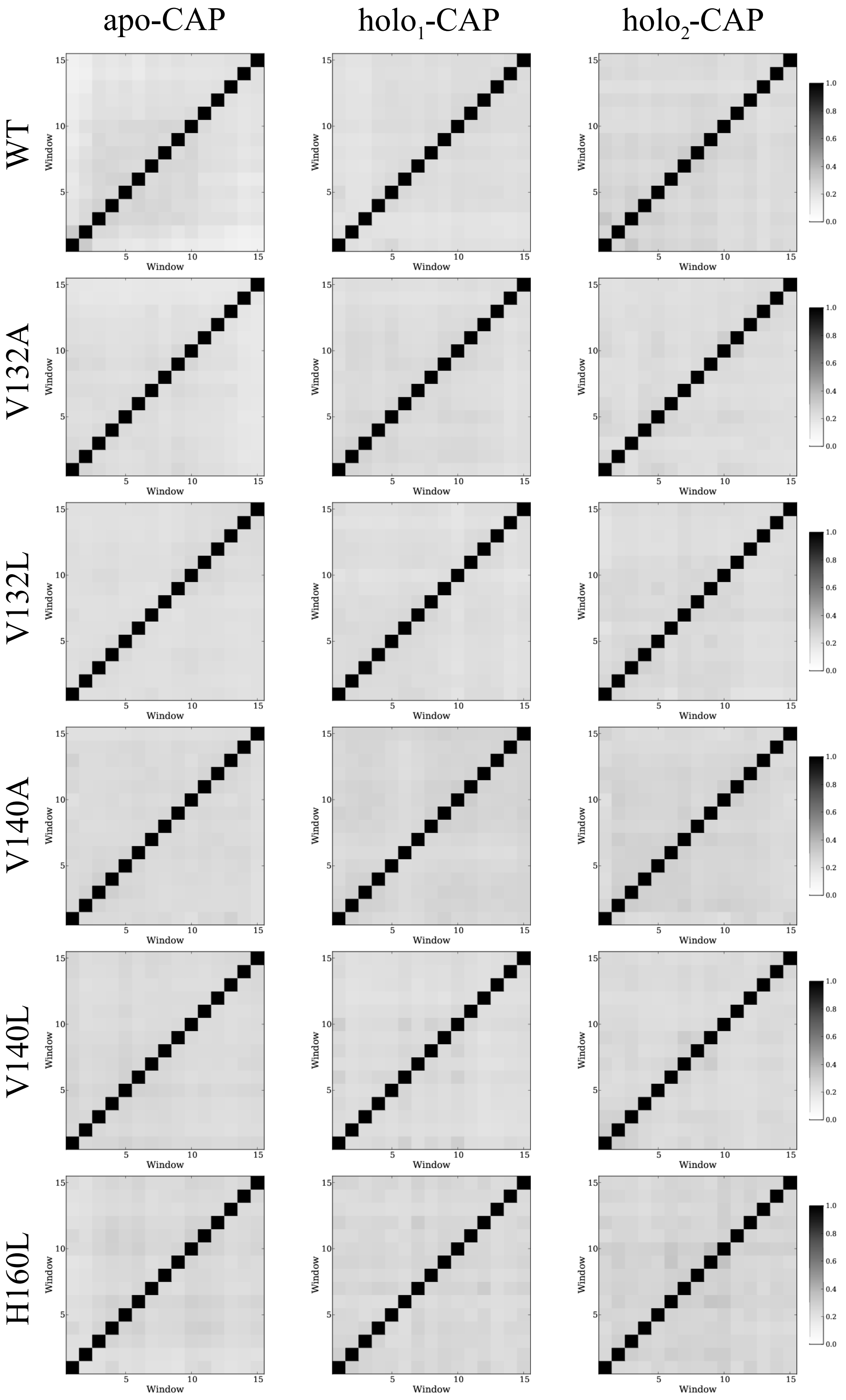


Figure 5.29: The normalised overlaps between different windows of a simulation of a system of CAP. Shown for the apo, holo$_1$ and holo$_2$ bound states of the WT, V132A, V132L, V140A, V140L and H160L variants of CAP.

This was performed for the simulations of the different bound states of the variants of CAP as shown in figure 5.29. This figure shows a measure of the similarities between the covariance matrices of two windows of the simulation, for every possible pair of windows. Black signifies a normalised overlap of 1, where the two covariance matrices are identical and white corresponding to a normalised overlap of 0 where the two covariance matrices are orthogonal.

The normalised overlap is 1 along the diagonal as is to be expected; the normalised overlap of a covariance matrix with itself is 1. The normalised overlaps between different windows of the simulations are much lower; typically between 0.2 and 0.3 for all pairs of windows. This can clearly be seen by looking at the average normalised overlaps between window pairs for each system in table 5.7.

The fact that the normalised overlaps is low between all of the windows of a simulation indicates that for the most part, the simulations are not in the same principal component space in each 20 ns window. This again can indicate either that the windows are sampling different regions of the energy surface or that they are not sampling the bottom of an energy well; or both.

For these reasons highlighted above, using PCA on 20 ns windows of an MD simulation as a technique to estimate the entropy of the system would yield values that one would have very little reason to have faith in. Therefore a different method was attempted to try and determine the entropies of the systems studied. This analysis is described in the following section.

5.2.7 Entropy of Binding from NMA Snapshots

This section looks at the techniques used to calculate the entropy of the cAMP binding events using the 300 ns simulations of the different CAP variants. The conformational entropy was determined by performing NMAs on snapshots of the simulation every 4 ns after a 100 ns equilibration time (totalling 50 snapshots).

For the NMAs, the snapshots used were stripped of all water molecules, reimaged so the two monomers were not split across the PBC and energy minimised using the L-BFGS method. For the energy minimisations that converged below the tolerance of 0.001 kJ mol^{-1} nm^{-1}, NMA was performed. Figure 5.30 shows the frequencies of the first 200 vibrational normal modes for all of the snapshots for WT apo-CAP. The

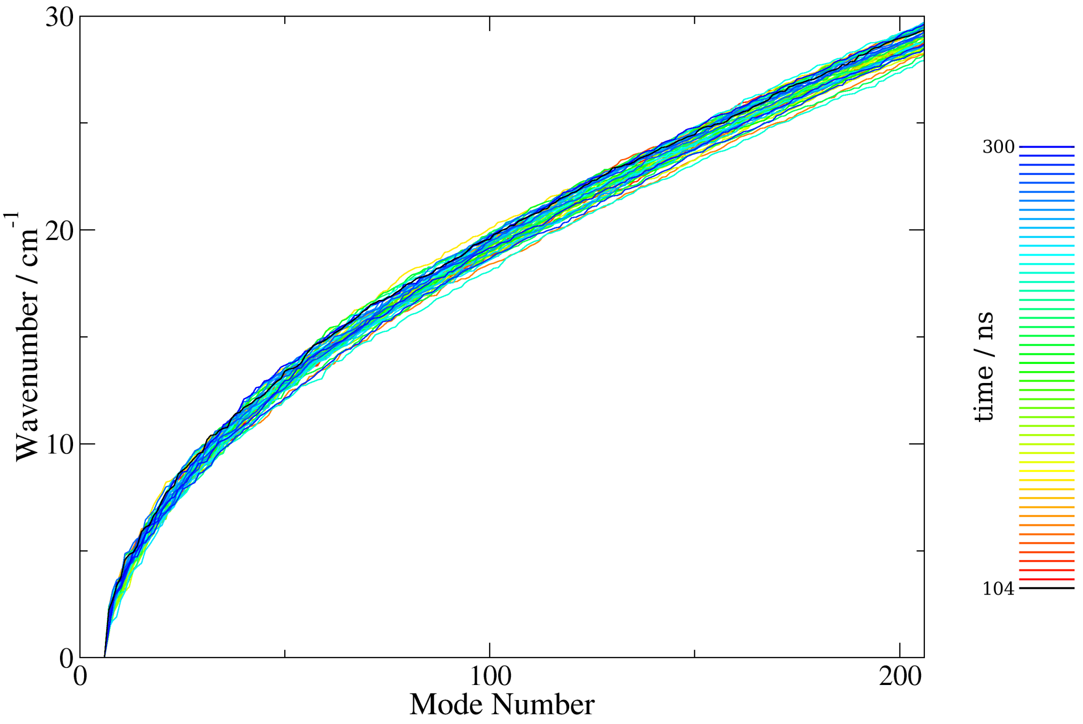


Figure 5.30: Frequencies of the first 200 normal modes of holo$_2$-CAP; calculated by NMA on 50 snapshots from the MD simulation.

distribution of these frequency signatures has a much smaller variance than for PCA. The frequencies calculated by NMA are also higher than those calculated by PCA. This is caused by a few factors. Firstly, the NMAs performed do not contain solvent molecules, which would have the effect of damping the normal modes in PCA. Secondly, the simulations used in PCA allow the system to move between local minima, so the resulting PCA gives a smoothing of the energy surface approximated as a harmonic well, which can incorporate multiple minima. NMA on the other hand is always confined to one local minimum. This effect is visualised in figure 5.31. Only if the simulation was confined to one local energy minimum, would the frequencies seen by NMA and PCA be very similar. The same protocol was also used to determine the entropy of cAMP from 50 snapshots of a 200 ns simulation.

The entropic contribution of each normal mode was determined from the frequencies collected from each snapshot of each system, using the classical method (see section 2.6.1). The average entropy for each mode in all of the systems was determined for each mode. The entropies of the systems were then determined by summing the entropies of up to the first 200 normal modes. Figure 5.32 shows plots of the average entropies of the normal modes for each system and a cumulative sum, which increases as the

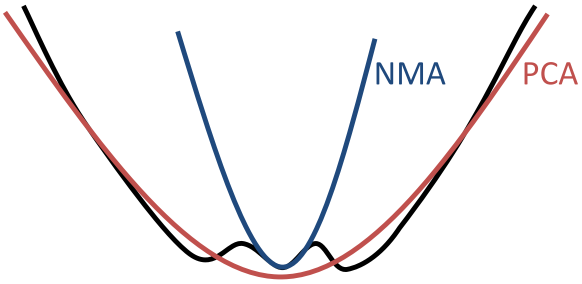


Figure 5.31: Differences between NMA and PCA: NMA creates a harmonic approximation of an energy minimum from the second derivative of that minimum. PCA samples the free energy surface throughout an MD simulation, creating a quasiharmonic approximation of the surface.

number of included modes increases.

These plots show a number of things. Firstly, the entropy of the systems is very large compared to the differences in entropy between the systems. Secondly, as would be expected the contribution to the entropy decreases as the mode number increases. The error bars in the cumulative entropy (given as the standard error of the mean calculated over the 50 4 ns snapshots) are displayed in inset plots of the systems, and show that they are very small relative to the entropies of the modes.

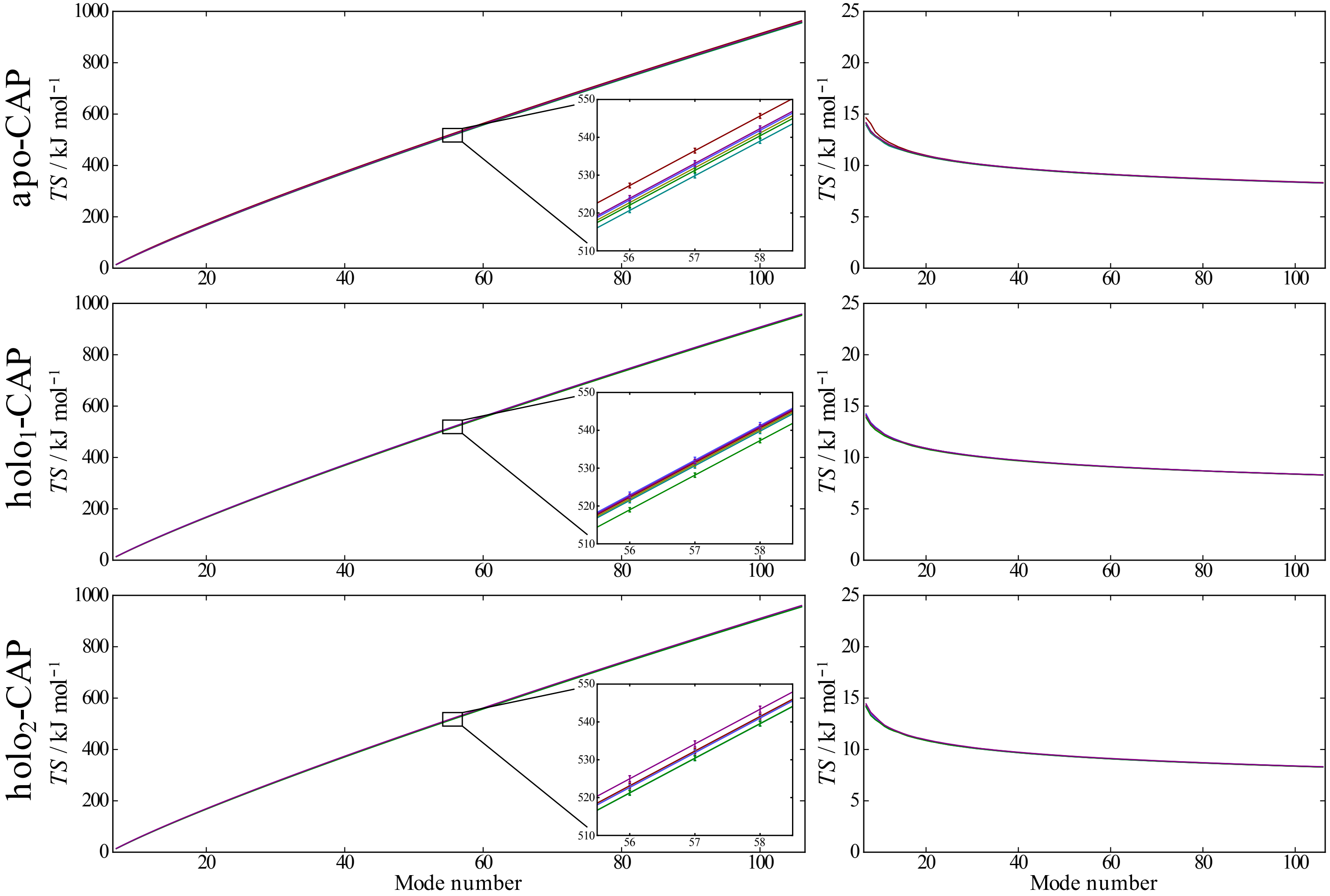


Figure 5.32: On the left, the average cumulative entropy for the first 100 none zero modes of apo, holo$_1$ and holo$_2$-CAP for each of the variants. The inset shows the size of the error bars, which are given as the standard error of the mean. The plots on the right show the average entropy of the individual modes for each system. — WT, — V132A, — V132L, — V140A, — V140L and — H160L.

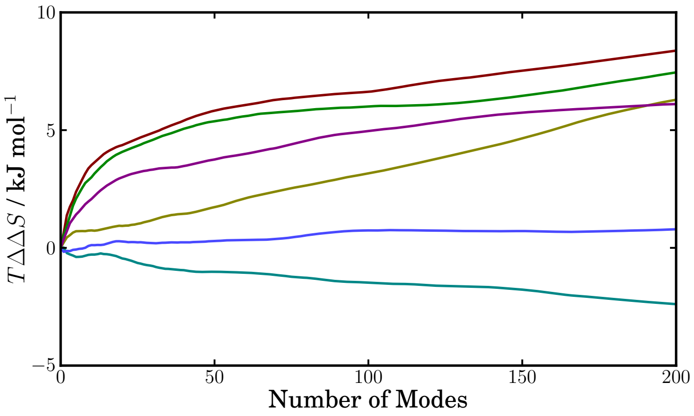


Figure 5.33: The cumulative allosteric entropy for the first 200 (none zero) modes for each variant: — WT, — V132A, — V132L, — V140A, — V140L and — H160L

The allosteric entropy for each variant was calculated with equation 2.6.71, using the average entropies for each system. The allosteric entropy as a function of the number of normal modes used in its calculation is shown for each variant in figure 5.33. This figure shows that after 200 normal modes, the entropies have not yet converged, so to get the absolute entropy, ideally more normal modes would be used in the calculation. However, only the first 200 none zero normal modes were calculated when running the NMAs (to obtain all of the normal modes would have been extremely time consuming for the 50 calculations per system performed here). The entropies calculated using this method show the entropy for four of the variants, WT, V132L, V140L and H160L are very similar.

Variant	$\boldsymbol{TS_{\mathrm{apo}}}$	$\boldsymbol{TS^{1}_{\mathrm{holo}}}$	$\boldsymbol{TS^{2}_{\mathrm{holo}}}$	$\boldsymbol{T\Delta\Delta S_{\mathrm{vib}}}$
	kJ mol^{-1}	**kJ mol^{-1}**	**kJ mol^{-1}**	**kJ mol^{-1}**
WT	1743.29(1.48)	1734.66(1.64)	1734.41(1.53)	8.38(2.69)
V132A	1730.17(1.16)	1732.06(1.36)	1731.56(1.51)	-2.39(2.34)
V132L	1735.90(1.35)	1732.18(1.35)	1734.75(1.61)	6.29(2.50)
V140A	1735.29(1.53)	1734.85(1.74)	1735.21(1.61)	0.79(2.82)
V140L	1733.40(1.47)	1728.78(1.48)	1731.61(1.46)	7.45(2.55)
H160L	1737.14(1.46)	1735.01(1.33)	1738.99(1.75)	6.11(2.64)

Table 5.8: Entropies for the three bound states of the CAP variants calculated with NMA on snapshots of the MD simulation. The allosteric entropy is also determined for each of the variants.

The trends seen in the conformational allosteric entropies calculated with this method are very different to the trends seen using the elastic network model. Using NMA on the snapshots from the MD simulation, the entropies of V132L, V140L and H160L are all similar to the WT value, the entropy of V140A is around 0 kJ mol^{-1} and the entropy of V132A around -2 kJ mol^{-1} when using the first 200 modes. The exact values are given in table 5.8. With the ENM, the main component of the free energy calculated is the conformational entropy, with an adjustment for the enthalpy. That model predicted that both V140 mutations would not affect the allostery in CAP, V132A would make cAMP more negatively cooperative and V132L and H160L would push the system towards positive cooperativity.

What is seen with the NMAs could imply that the conformational entropy is not the sole driving force behind the cooperativity in CAP, as is implied by the ENM. It is likely that other factors will come into play such as enthalpy-entropy compensation. The following section tries to investigate these other contributions to the allosteric free energy using the MM/PBSA method. Also, in chapter 6 the entropies calculated with this method are compared to the experimental entropies.

5.2.8 MM/PBSA

To try to determine the other contributions to the free energy of binding, the MM/PBSA method was used. The method used in this thesis was initially checked by running the calculations from a recent study of the stability of cyclic peptide (CP) nanotubes [235]. 200 snapshots were used from a 10 ns simulation of the *cyclo*[(D-Ala-L-Ala)$_4$] nanotube was used in this check, to see if the values obtained were comparable to the original study. Two different models for the nonpolar contribution to the free energy were used. The first was the SASA model and the second was the model proposed by Tan et al. [148], which models the repulsive contribution with the SASA and the attractive contribution with an approximation of the van der Waals attractive forces. This nonpolar model will be termed the SASA/vdW model from this point for clarity. Both of these models are described in more detail in section 2.1.12. Table 5.9 shows a quick comparison of the values obtained with the methods used in this thesis to the values obtained in the study.

There is strong agreement between the values calculated by the SASA model in

	Original study		**A**		**B**	
	$\Delta G_{\text{MM/PBSA}}$ / kcal mol^{-1}		**$\Delta G_{\text{MM/PBSA}}$ / kcal mol^{-1}**		**$\Delta G_{\text{MM/PBSA}}$ / kcal mol^{-1}**	
CP Unit	**AVG**	**SD**	**AVG**	**SD**	**AVG**	**SD**
CP1	-27.18	1.98	-24.77	4.70	-14.18	2.90
CP2	-54.63	2.79	-50.33	6.07	-31.25	3.80
CP3	-55.80	2.57	-52.19	4.84	-32.10	3.31
CP4	-55.85	2.77	-52.82	4.90	-32.67	3.06
CP5	-55.57	2.60	-52.74	4.84	-32.59	2.95
CP6	-55.57	2.66	-52.44	5.20	-31.86	3.47
CP7	-54.03	3.01	-50.99	5.62	-31.22	3.39
CP8	-26.48	2.30	-25.10	4.81	-14.33	2.49

Table 5.9: The MM/PBSA contribution to the free energy of assembly of CP nanotubes, $\Delta G_{\text{MM/PBSA}}$. Showing the values from the 2012 study by Subramanian *et al.* [235] and two values calculated in this study by two different methods; **A**, the SASA model to calculate the nonpolar contribution to the free energy and **B**, the SASA/vdW model [148].

this study and the values from the study by Subramanian *et al.* ($r^2 = 1.0$ and $p = 1.1 \times 10^{-9}$). This is the model that was used in the original study, which suggests that the method used in this study correctly reproduces the method used in the original study. The standard deviation in the values was larger for the calculations in this study as less snapshots were used.

The magnitudes of the values for ΔG calculated using the SASA/vdW model for the nonpolar contribution are smaller than the values calculated for the SASA model. This does not mean that this model should be abandoned though, with no experimental studies of the stability of the nanotube to compare the values to, it is not clear which methodology models the nanotube better. Both models are therefore used for modelling the nonpolar contributions of CAP-cAMP binding in this study.

The only differences between these calculations and the calculations for CAP-cAMP binding are the frequency of the snapshots used and that the CAP-cAMP calculations include an ionic strength of 300 mM to try to mimic the conditions for ITC.

Investigating CAP-cAMP Binding

MM/PBSA was then used to investigate the binding of cAMP to CAP. Tables 5.10 and 5.11 break down the MM/PBSA energy contributions for both of the binding events using the SASA and SASA/vdW models for the nonpolar contribution. Naturally the only differences between the models are in the nonpolar contributions. The SASA method only has one contribution, so in this case the ΔE_{SASA} label is used, whereas the SASA/vdW method has two contributions, so the ΔE_{SASA} and ΔE_{DISP} labels are used for the repulsive and attractive contributions respectively.

	Variant	$\Delta E_{\mathrm{bonded}}$	ΔE_{VDW}	ΔE_{EL}	ΔE_{PB}	ΔE_{SASA}	$\Delta G_{\mathrm{MM/PBSA}}$
		/ kJ mol^{-1}	/ kJ mol^{-1}	/ kJ mol^{-1}	/ kJ mol^{-1}	/ kJ mol^{-1}	/ kJ mol^{-1}
apo→holo$_1$	WT	-3.45 (1.87)	-70.63 (1.42)	-100.79 (5.16)	94.57 (4.26)	-4.84 (0.04)	-85.15 (2.04)
	V132A	5.08 (1.89)	-32.09 (1.44)	87.45 (5.00)	-88.72 (4.13)	-0.19 (0.05)	-28.47 (2.07)
	V132L	-0.01 (1.90)	-53.40 (1.46)	-36.75 (5.49)	40.87 (4.49)	-2.93 (0.05)	-52.21 (2.12)
	V140A	7.07 (1.88)	-47.65 (1.38)	-14.37 (5.23)	17.95 (4.53)	-1.29 (0.04)	-38.30 (2.09)
	V140L	-9.16 (1.91)	-49.43 (1.38)	-52.91 (5.15)	48.85 (4.46)	-3.01 (0.05)	-65.65 (2.09)
	H160L	-11.62 (1.87)	-51.24 (1.36)	-238.49 (5.30)	246.66 (4.47)	-3.55 (0.05)	-58.24 (2.08)
holo$_1$ →holo$_2$	WT	-4.38 (1.87)	-33.91 (1.40)	-59.53 (4.93)	72.89 (4.09)	-1.94 (0.04)	-26.87 (2.02)
	V132A	-4.80 (1.87)	-45.51 (1.44)	-184.86 (5.47)	170.43 (4.46)	-3.07 (0.05)	-67.81 (2.17)
	V132L	-19.76 (1.89)	-12.93 (1.49)	-184.51 (5.38)	187.71 (4.41)	-1.61 (0.05)	-31.09 (2.05)
	V140A	-9.15 (1.89)	-33.89 (1.36)	-170.51 (5.30)	160.74 (4.77)	-3.26 (0.05)	-56.07 (2.07)
	V140L	-4.53 (1.88)	-22.15 (1.37)	-171.91 (5.27)	172.40 (4.23)	-1.05 (0.05)	-27.24 (2.05)
	H160L	6.84 (1.89)	-22.13 (1.38)	108.74 (5.18)	-97.07 (4.51)	-0.01 (0.04)	-3.63 (2.09)

Table 5.10: Contributions to the free energy of both cAMP to CAP binding events. $\Delta E_{\mathrm{bonded}}$ contains the energy contributions calculated using differences in the bond length, bond angle and dihedral energy terms. ΔE_{VDW} contains the energy contributions calculated using differences in the vdW and 1-4 vdW energy terms. ΔE_{EL} contains the energy contributions calculated using differences in the electrostatic and 1-4 electrostatic energy terms. ΔG_{PB} contains the polar energy contributions to the differences in solvation free energy calculated using the PB implicit solvent model. The nonpolar contribution ΔG_{SASA} to the free energy difference was calculated using the SASA model (see section sub:NP-implicit-theory). The sum of all of these contributions, $\Delta G_{\mathrm{MM/PBSA}}$ does not include a vibrational entropy contribution.

	Variant	**$\Delta E_{\mathrm{bonded}}$ / kJ mol^{-1}**	**ΔE_{VDW} / kJ mol^{-1}**	**ΔE_{EL} / kJ mol^{-1}**	**ΔE_{PB} / kJ mol^{-1}**	**ΔE_{DISP} / kJ mol^{-1}**	**ΔE_{SASA} / kJ mol^{-1}**	**$\Delta G_{\mathrm{MM/PBSA}}$ / kJ mol^{-1}**
apo→holo$_1$	WT	-3.45 (1.87)	-70.63 (1.42)	-100.79 (5.16)	94.57 (4.26)	-47.28 (0.39)	83.84 (0.53)	-43.74 (2.11)
	V132A	5.08 (1.89)	-32.09 (1.44)	87.45 (5.00)	-88.72 (4.13)	-10.94 (0.43)	19.89 (0.67)	-19.32 (2.15)
	V132L	-0.01 (1.90)	-53.40 (1.46)	-36.75 (5.49)	40.87 (4.49)	-32.89 (0.46)	61.48 (0.59)	-20.69 (2.17)
	V140A	7.07 (1.88)	-47.65 (1.38)	-14.37 (5.23)	17.95 (4.53)	-23.93 (0.41)	40.25 (0.52)	-20.68 (2.13)
	V140L	-9.16 (1.91)	-49.43 (1.38)	-52.91 (5.15)	48.85 (4.46)	-31.03 (0.42)	58.17 (0.59)	-35.52 (2.16)
	H160L	-11.62 (1.87)	-51.24 (1.36)	-238.49 (5.30)	246.66 (4.47)	-35.02 (0.39)	62.50 (0.51)	-27.21 (2.14)
holo$_1$ →holo$_2$	WT	-4.38 (1.87)	-33.91 (1.40)	-59.53 (4.93)	72.89 (4.09)	-20.68 (0.39)	41.28 (0.50)	-4.33 (2.06)
	V132A	-4.80 (1.87)	-45.51 (1.44)	-184.86 (5.47)	170.43 (4.46)	-33.89 (0.41)	58.28 (0.58)	-40.34 (2.22)
	V132L	-19.76 (1.89)	-12.93 (1.49)	-184.51 (5.38)	187.71 (4.41)	-19.42 (0.44)	32.45 (0.59)	-16.44 (2.09)
	V140A	-9.15 (1.89)	-33.89 (1.36)	-170.51 (5.30)	160.74 (4.77)	-31.22 (0.38)	60.20 (0.53)	-23.82 (2.13)
	V140L	-4.53 (1.88)	-22.15 (1.37)	-171.91 (5.27)	172.40 (4.23)	-18.82 (0.41)	28.86 (0.55)	-16.15 (2.12)
	H160L	6.84 (1.89)	-22.13 (1.38)	108.74 (5.18)	-97.07 (4.51)	-9.34 (0.37)	18.72 (0.47)	5.76 (2.16)

Table 5.11: Contributions to the free energy of both cAMP to CAP binding events. This table contains all the same energy contributions as table 5.10 except for the nonpolar contribution. In this table the nonpolar energy contribution has two contributions; a repulsive contribution, ΔG_{SASA} determined using the SASA model and an attractive contribution, ΔG_{DISP}, an approximation of the vdW forces (see section 2.1.12). The values shown do not include the vibrational entropy contribution to the free energy, which is discussed in section 5.2.7.

Scrutinising the individual contributions to the free energy calculated by this method, the efficacy of the method for free energy calculations of binding events can be determined. The magnitude of the changes to the energy of ΔE_{bonded} (the bond length, angle and dihedral energies) is relatively small. This would be expected, as there no are bonds made or broken in the binding event. The vdW energy, ΔE_{VDW}, is negative for both binding events as one would expect, with new interactions being formed between the protein and the ligand.

What is surprising is the magnitude and range of values seen for ΔE_{EL}, the change in electrostatic energy for the binding events. The hydrogen bonds formed between the protein and the ligand would likely be the largest contribution to this electrostatic energy for both of the binding events. For both binding events, a similar number of hydrogen bonds would be made between the ligand and the protein. Making the liberal estimate of eight new hydrogen bonds forming on each binding event, and another liberal estimate of each of them contributing -20 kJ mol^{-1} to ΔE_{EL} would place the expected value of E_{EL} around -160 kJ mol^{-1}. This value is only a very rough estimate, as these new hydrogen bonds would not be the only changes to the electrostatic interactions of the system upon binding. However, the value would be expected to be not as dissimilar between different variants and between different binding events as is seen. For example for H160L, the first binding event has a ΔE_{EL} of -238 kJ mol^{-1} and the second binding event has a ΔE_{EL} of 109 kJ mol^{-1}. The value for the first binding event being favourable as would be expected and the value for the second binding event being unexpectedly quite strongly unfavourable. This is the most extreme example, but not the only example of unexpected values for ΔE_{EL}.

These inconsistencies between the values of ΔE_{EL}, strongly detract from the suitability of using the MM/PBSA method for determining the free energy of protein-ligand binding. This is even more true for determining allosteric energies, which requires calculating the difference between two binding events, thus increasing any errors further. A possible reason why such large fluctuations are seen in this electrostatic energy is again that the MD simulations may not have converged. It is also quite likely that with a protein as large as CAP, it is unlikely to ever converge to give reasonable results using a method such as MM/PBSA.

Regardless of the mistrust of the MM/PBSA method, the remaining energy contri-

Variant	$\boldsymbol{\Delta\Delta E_{\text{gas}}}$ **/ kJ mol^{-1}**	$\boldsymbol{\Delta\Delta G_{\text{solv}}}$ **/ kJ mol^{-1}**	$\boldsymbol{\Delta\Delta G_{\text{MM/PBSA}}}$ **/ kJ mol^{-1}**	$\boldsymbol{\Delta\Delta G_{\text{TOT}}}$ **/ kJ mol^{-1}**
WT	77.06 (6.68)	-19.87 (5.87)	57.19 (2.87)	48.81
V132A	-295.60 (7.07)	256.21 (6.04)	-39.38 (3.00)	-36.99
V132L	-127.04 (7.28)	148.21 (6.26)	21.16 (2.95)	14.87
V140A	-158.59 (7.39)	140.98 (6.54)	-17.61 (2.94)	-18.40
V140L	-87.09 (6.97)	124.43 (6.10)	37.34 (2.93)	29.89
H160L	394.80 (7.16)	-339.66 (6.32)	55.14 (2.95)	49.03

Table 5.12: $\Delta\Delta G_{\text{MM/PBSA}}$ values calculated using the SASA method for calculating the non-polar contribution. The values shown do not include the vibrational entropy contribution to the free energy, which is discussed in section 5.2.7.

Variant	$\boldsymbol{\Delta\Delta E_{\text{gas}}}$ **/ kJ mol^{-1}**	$\boldsymbol{\Delta\Delta G_{\text{solv}}}$ **/ kJ mol^{-1}**	$\boldsymbol{\Delta\Delta G_{\text{MM/PBSA}}}$ **/ kJ mol^{-1}**	$\boldsymbol{\Delta\Delta G_{\text{TOT}}}$ **/ kJ mol^{-1}**
WT	77.06 (6.68)	-38.74 (5.93)	38.32 (2.95)	29.94
V132A	-295.60 (7.07)	274.53 (6.13)	-21.06 (3.09)	-18.67
V132L	-127.04 (7.28)	131.34 (6.34)	4.30 (3.02)	-1.99
V140A	-158.59 (7.39)	155.62 (6.69)	-2.97 (3.01)	-3.76
V140L	-87.09 (6.97)	105.39 (6.23)	18.30 (3.03)	10.85
H160L	394.80 (7.16)	-361.31 (6.39)	33.49 (3.04)	27.38

Table 5.13: $\Delta\Delta G_{\text{MM/PBSA}}$ values calculated using the SASA/vdW method for calculating the non-polar contribution.

butions will be further analysed.

The PB solvation energy, ΔE_{PB}, as would be expected has the opposite sign to the electrostatic energy. Thus estimating the effect of the interactions being removed between the ligand or protein and the solvent when the new interactions between ligand and protein are made. However again, with some of the unexpected positive values of ΔE_{EL} come unexpected negative values of ΔE_{PB}.

The values calculated for $\Delta G_{\text{MM/PBSA}}$ using the SASA/vdW model for the nonpolar contribution to the solvation free energy seem slightly more reasonable than the SASA model alone. This is more apparent when looking at the value of $\Delta\Delta G_{\text{MM/PBSA}}$ and combining these values with $T\Delta\Delta S_{\text{vib}}$ (determined by NMA in section 5.2.7). The resulting value of $\Delta\Delta G_{\text{TOT}}$ (as shown in table 5.13) seems closer to the experimental

ITC values [59] than using just the SASA method for estimating the nonpolar contribution for WT-CAP (shown in table 5.12). The ITC values are discussed in chapter 6. Both models correctly show the negative cooperativity for WT-CAP, however the negative cooperativity is weaker and more reasonable using the SASA/vdW model for the nonpolar contribution to the solvation part of $\Delta\Delta G_{\mathrm{TOT}}$.

One thing that should be taken into account however is the margin of error of the values calculated using MM/PBSA. The values in the above tables show the standard error of the mean in parentheses. Each data point used to calculate the mean, however, is not independent from the rest as they all come from the same simulation. This means that the standard error of the mean underestimates the error by what could be a large amount. This weakens the argument that these methods correctly predict the negative allostery in WT-CAP. This is discussed further when the values of $\Delta\Delta G_{\mathrm{TOT}}$ determined with MM/PBSA and NMA are compared to the ITC values for each of the mutants in section 6.3.3.

5.2.9 Overall Discussion

In this chapter it was revealed that there was no significant change to the structure of the LBDs after 300 ns of simulation for apo-CAP or holo$_1$-CAP, for any of the CAP variants, when compared to holo$_2$-CAP. This further supports the hypothesis that CAP undergoes allostery without conformational change. The fact that making the mutations also does not change the LBD, further contributes to the hypothesis that allostery can be manipulated without a conformational change.

The DNABD, on the other hand is seen to undergo a structural rearrangement for one chain in the WT apo-CAP simulation. This rearrangement, although not occurring to the same extent as that seen by Steitz and coworkers in the crystal structure of a CAP variant [108], indicates the flexibility of the DNABD. As this rearrangement does not occur to the same extent in any of the variants suggests an energy barrier to this motion that is not crossed for these simulations.

A potential signalling pathway involving residue Thr128 was investigated, showing a change in its preferred orientation and a decrease in flexibility in this region upon cAMP binding. There was also shown to be less flexibility in the apo forms of the CAP variants (V132L and H160L) that were predicted by the ENM to undergo positively cooperative binding of cAMP.

Additionally, free energy calculations for the allostery in CAP were performed. These predicted weak negative cooperativity for WT-CAP as would be expected. However the margins of error on these calculations are large enough to throw doubt upon the value calculated. The thermodynamics of binding cAMP to the CAP variants is investigated further using ITC in the next chapter.

Chapter 6

Experimental

This chapter outlines the materials, methods and results for the experiments used to study the structure and dynamics of these variants and how cooperative binding of cAMP to CAP is affected by making these point mutations. The hypotheses and observations made using the CG and atomistic methods in chapters 4 and 5 respectively are investigated experimentally using X-ray crystallography and ITC. The experiments for the Q108A variant (performed by David Burnell) are shown in detail and the experiments for the remaining variants (performed by Phil Townsend) are provided with less detail. Comparisons and differences between the experimental results and the atomistic calculations are also investigated in this section.

6.1 Materials

Unless stated otherwise all chemicals were purchased from Sigma-Aldrich at highest purity and Milli-Q®water was used.

6.1.1 *Pfu* DNA Polymerase

Pfu (*pyrococcus furiosus*) DNA polymerase is a thermostable enzyme (Thermo Scientific) used in a polymerase chain reaction (PCR) in the presence of Mg^{2+} (see section 6.2.1). As well as DNA synthesis in the 5'-3' direction, it also exhibits proof reading activity in the 3'-5' direction reducing the likelihood of any unwanted mutations. [236–239].

6.1.2 BIOTAQ^{TM} Red DNA Polymerase

BIOTAQ^{TM} Red DNA polymerase (Bioline) is a mixture of the BIOTAQ^{TM} (*Taq*) DNA polymerase and a red dye. *Taq* DNA polymerase is a thermal stable enzyme purified from *Thermus aquaticus* [240]. The red dye allows direct loading onto an agarose gel without the need for DNA loading buffer.

6.1.3 T4 DNA Ligase

T4 DNA ligase (Thermo Scientific) catalyses the fusion of two DNA strands either in a cohesive ended or a blunt ended configuration. For example T4 DNA ligase can catalyse the insertion of a target gene into a plasmid [241].

6.1.4 HindIII and BamHI Restriction Enzymes

Restriction enzymes cut DNA at specific sites defined by the base sequence. The restriction site for BamHI (Thermo Scientific) is 5'-G|GATC-3', leaving a GATC overhang on the DNA chain [242]. The restriction site for HindIII (Thermo Scientific) is 5'-A|AGCT.T-3', leaving an AGCT overhang on the DNA.

6.1.5 Calf Intestinal Alkaline Phosphatase (CIAP)

Calf intestinal alkaline phosphatase (CIAP), purified from calf intestinal mucosa, (Invitrogen^{TM}) is a phosphomonoesterase that hydrolyses 5'-phosphate groups from DNA. It is used to remove 5'-phosphate groups from a linearised plasmid before performing a T4 DNA ligation reaction with an insert.

6.1.6 Plasmids

pCR^{TM}-Blunt II-TOPO® (Invitrogen) is a plasmid that allows the ligation of a blunt ended PCR product into the vector that contains kanamycin resistance (see appendix D.1) [243].

pQE30 (QIAGEN) is an IPTG (section 6.1.9) inducible expression vector that confers ampicillin resistance. It introduces a histidine (H_6) tag on the N-terminus of the expressed protein (see appendix D.3).

pREP4 (QIAGEN) is a plasmid (that is compatible with the pQE30 vector) which contains the *lacI*^{q} gene and confers ampicillin resistance (see appendix D.2).

6.1.7 *E. coli* Strains

Plasmids containing the target gene were transformed into chemically competent *E. coli* strains DH5α (Invitrogen^{TM}) and Mach1^{TM} T1 (Invitrogen^{TM}) when checking the ligation efficiency by colony PCR and DNA sequencing.

For protein expression, the *E. coli* strain M182 Δ CAP F$^-$ Δ(*lacIPOZY*)X74 *galE*15 *galK*16 *rpsL thi*$^+$ *lambda*$^-$ [pREP4] [244,245] was used. This *E. coli* strain has the WT-CAP gene removed to avoid the WT protein expressing at the same time as the mutated protein.

6.1.8 Antibiotics

Two antibiotics were used as selection agents when growing *E. coli* colonies. These were kanamycin (50 µg ml^{-1}) and ampicillin (100 µg ml^{-1}). The plasmids; pCR^{TM}-Blunt II-TOPO® and pREP4 contain resistance to kanamycin, while pQE30 is resistant to ampicillin.

6.1.9 Isopropyl β-D-thiogalactopyranoside (IPTG)

IPTG triggers transcription of the lac operon by binding to the lac repressor, releasing it from the lac operon. If a foreign gene has been introduced into the lac operon (as is the case in the pQE30 expression vector (section 6.1.6), it causes the transcription of this gene to occur and hence expression of the target protein.

6.1.10 CAP buffers

Table 6.1 contains the buffers used with CAP during protein purification, mass spectrometry (MS) and ITC.

6.1.11 PACT premier^{TM} HT-96

PACT premier^{TM} HT-96 (Molecular Dimensions) is a crystallisation screen containing 96 × 1 ml solutions. The screen contains three separate screens; a 24 well pH/PEG

Buffer	Contents	pH
Lysis Buffer	100 mM KH_2PO_4/K_2HPO_4, 200 mM KCl, 2 mM 2-thioglycerol	7.8
Wash Buffer	100 mM KH_2PO_4/K_2HPO_4, 200 mM KCl, 2 mM 2-thioglycerol, 15 mM imidazole	7.8
Elution Buffer	100 mM KH_2PO_4/K_2HPO_4, 200 mM KCl, 2 mM 2-thioglycerol, 300 mM imidazole	7.8
Storage Buffer	100 mM KH_2PO_4/K_2HPO_4, 200 mM KCl, 2 mM 2-thioglycerol, 15% (v/v) glycerol	7.8
ITC Buffer	100 mM KH_2PO_4/K_2HPO_4, 200 mM KCl, 2 mM 2-thioglycerol	7.8
Mass Spec. Buffer	20 mM Tris-HCl, 100 mM NaCl	7.8

Table 6.1: Buffers used in the purification, ITC and MS of CAP.

screen, a 24 well cation/PEG screen and a 48 well anion/PEG screen [246]. The composition of the solutions in the screen are given in appendix E.

6.2 Methods

6.2.1 Polymerase Chain Reaction (PCR)

PCR utilises DNA polymerase to amplify a specific target DNA sequence *in vitro*; multiplying as little as a few molecules of the target DNA to as much as a microgram [247].

In PCR, two short oligonucleotide primers (typically 20-30 base pairs long) bind the target sequence of DNA to be amplified; one primer at the 5' end of the target DNA in the forward reading direction and the other at the 5' end in the backward reading direction [248]. Each PCR experiment consists of three steps; denaturation, annealing and extension. This sequence of steps is repeated numerous times to generate enough DNA for the purpose [249]. In a cycle of the PCR experiment the number of target DNA copies would double if it were 100% efficient, however in reality the number of DNA copies does not quite double. The new DNA copies can then be used as DNA

Primer	Direction	Sequence	Plasmid/Gene
F-CAP$_{1-6}$	Forward	atggtgcttggcaaaccg	WT CAP
R-CAP$_{205-212}$	Reverse	ttattaacgagtggcgtaaacga	WT CAP
F-CAPQ108A$_{104-112}$	Forward	cgccaattgattGCGgtaaacccggac	Q108A CAP
R-CAPQ108A$_{104-112}$	Reverse	gtccgggtttacCGCaatcaattggcg	Q108A CAP
M13 Reverse	Reverse	caggaaacagctatgac	pCR-Blunt II-TOPO®™

Table 6.2: Primers used for the mutagenesis of Q108A CAP. Mutated bases are capitalised.

templates in the following cycles.

The primers used in PCR experiments determine the initiation site of the DNA synthesis, but not where it terminates. After one cycle, this creates new single strands of DNA extending beyond the end of the target sequence. After a small number of cycles, strands of DNA that have been synthesised in the forward direction in one step will be the templates for DNA synthesis in the backwards direction. The DNA synthesised in this cycle will have a precisely defined length, containing only the target sequence. After more cycles, the number of strands with this defined length will greatly outnumber the strands of DNA with undefined lengths. PCRs take place in a thermal block cycler, a device which automatically and accurately adjusts the temperature of the PCR mixtures according to a program defined by the user.

6.2.2 Primer Design

Two complementary DNA oligomers (primers) consisting of 20-30 base pairs were designed to contain the desired mutation. This length allows for sufficient specificity to ensure that the sequence is not going to appear elsewhere in the DNA template, or in any contaminant DNA. It is also a short enough length to ensure ease of binding to the DNA template at the annealing temperature.

The sequence of the forward primer was designed to match the DNA template for the nucleotides flanking the mutation site. However, the DNA base triplet at the mutation site itself was changed to match the genetic code for the point mutant. The sequence of the reverse primer was the reverse conjugate of the forward primer.

The primers were also designed to have a few G or C bases in the last five bases of the 3' end of the primer. The stronger bonding between the G and C bases compared

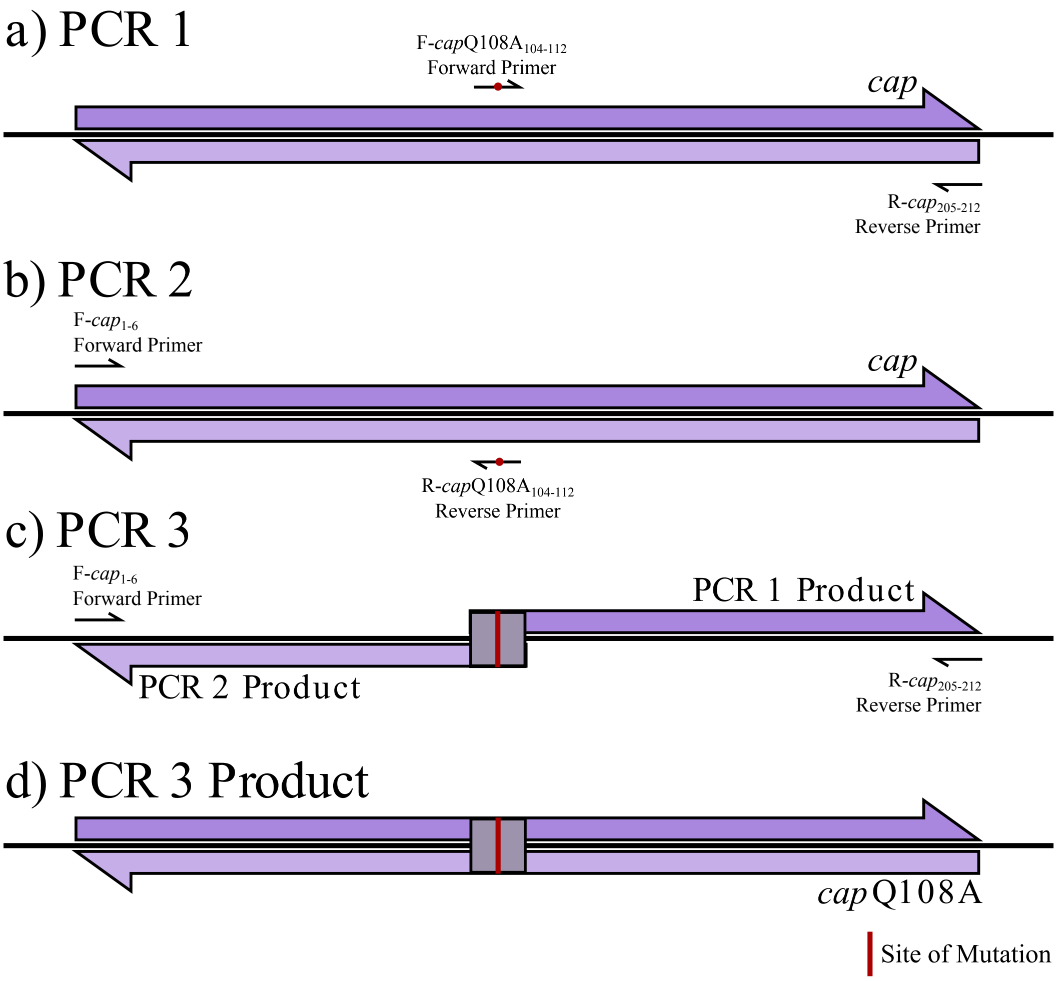


Figure 6.1: Site specific mutagenesis using three separate PCRs; (a) and (b) synthesise overlapping DNA fragments from the DNA template containing the desired mutation. (c) then combines the DNA fragments, creating an copy of the DNA template sequence with the desired mutation (d).

to the bonding between the A and T bases encourages specific binding of the primer. WT primers flanking the CAP gene were also used in numerous PCR experiments. All of the primers used for any type of PCR can be seen in table 6.2.

6.2.3 PCR Site Specific Mutagenesis

Complementary oligomers containing the desired mutation (section 6.2.2) were used in two separate polymerase chain reactions, with primers flanking the CAP gene to synthesise two overlapping fragments of DNA. These fragments were then used in a third PCR (PCR3) where the DNA fragments annealed at their overlap to act as primer for the 3' extension of the opposite strand [250]. This process can be seen in figure 6.1. The reaction mixtures for the first two PCRs (PCR1 and PCR2 in table 6.3) were set

Component	Volume / µl (PCR1)	Volume / µl (PCR2)	Volume / µl (PCR3)
Pfu DNA Polymerase 10× Buffer with $MgSO_4$*	2.5	2.5	2.5
dNTPs	2.5	2.5	2.5
DNA chain (WT CAP)	0.5	0.5	-
Primer (F-CAPQ108A$_{104-112}$)	1	-	-
Primer (R-CAP$_{205-212}$)	1	-	1
Primer (F-CAP$_{1-6}$)	-	1	1
Primer (R-CAPQ108A$_{104-112}$)	-	1	-
DNA chain (PCR1 output)	-	-	1
DNA chain (PCR2 output)	-	-	1
Pfu DNA Polymerase	0.5	0.5	0.5
H_2O	17	17	15.5

Table 6.3: Polymerase Chain Reaction mixtures with a total volume of 25 µl used for site specific mutagenesis

	Cycle Step	Temp / °C	Time (PCR1/2)	Time (PCR 3)
1	Initial denaturation	95	2 minutes	2 minutes
2	Denaturation	95	45 seconds	1 minute
3	Annealing	55	45 seconds	1 minute
4	Elongation	72	45 seconds	1 minute
5	Final Elongation	72	5 minutes	5 minutes
6	Storage	4	-	-

Table 6.4: PCR cycles used for overlap extension PCRs. Steps 2, 3 and 4 were repeated in a cycle 30 times.

up on ice and the reaction took place in a thermal block cycler with the cycle outlined in table 6.4. The denaturation-annealing-elongation cycle was repeated 30 times.

Agarose gel electrophoresis (section 6.2.4) was used to distinguish the synthesised DNA fragments from the WT gene. These fragments were then extracted (section 6.2.5 from the agarose and used in the next PCR. The reaction mixtures and the cycle used for this reaction are (PCR3) in tables 6.3 and 6.4.

6.2.4 Agarose Gel Electrophoresis

0.75 g of electrophoresis grade agarose gel powder was added to 50 ml TAE buffer (40 mM Tris-acetate, 1 mM EDTA, pH 8.0) per each agarose gel required and microwaved until dissolved. 12.5 µl of 2.5×10^{-2} M ethidium bromide per gel was added before the gel set. DNA samples were mixed 4:1 with loading buffer (2% (w/v) orange G and 20% (w/v) sucrose in TAE buffer) then loaded into a gel. A DNA marker ladder was loaded alongside, for the purpose of estimating the length of the DNA chains. The gels were run at a voltage between 100 V and 140 V. DNA was visualised using a UV transilluminator.

6.2.5 DNA Extraction from Agarose Gels

To obtain pure target DNA, the DNA bands were cut from the agarose gel and dissolved in 600 µl binding buffer (6 M sodium perchlorate, 10 mM EDTA, 50 mM Tris-HCl pH 8.0) at 65 °C. 10 µl of 166 mg ml^{-1} silica suspension in water was added and the mixture was held at room temperature for 30 minutes.

The suspension was centrifuged for 30 seconds and the supernatant discarded (All centrifugation steps were run at 12000 g) The pellet was re-suspended in 200 µl binding buffer and centrifuged again for 30 seconds. The supernatant was again discarded and the pellet suspended in 750 µl wash buffer (400 mM NaCl, 20 mM Tris-HCl pH 7.5, 2 mM EDTA and 50% (v/v) ethanol) before being centrifuged again. The fluid was discarded using a pump and the solid residue left to air dry at room temperature for 30 minutes.

The solid residue was then suspended in 15 µl H_2O and incubated at 37 °C for 15 minutes, before being centrifuged again for 2 minutes. As much of the supernatant as possible was removed without disturbing the silica. This fluid, containing the DNA was then stored at -20 °C until needed.

6.2.6 Zero Blunt® TOPO® PCR cloning reaction

This reaction is a cloning strategy for inserting blunt ended PCR products into a plasmid vector. The ligation reaction was prepared on ice with 1 µl pCR™-Blunt II-TOPO®, 3 µl PCR product, 1 µl 6× salt solution (200 mM NaCl, 10 mM $MgCl_2$) and

1 µl H_2O. This was held at room temperature for 30 minutes before being transformed into chemically competent *E. coli* (section 6.2.7). The *E. coli* was then left to grow overnight on LB agar plates with the antibiotic Kanamycin (section 6.2.9).

In this reaction, the DNA is inserted into the pCR™-Blunt II-TOPO® plasmid in either the forward or reverse direction, so the directionality of the insert for a number of colonies on the agar plate was checked by colony PCR (section 6.2.10) using the M13 reverse primer (from the vector) and primers F-CAPQ108A$_{104-112}$ and R-CAPQ108A$_{104-112}$ (from the insert).

6.2.7 Transformation into Chemically Competent *E. coli*

50 µl of chemically competent *E. coli* was thawed then immediately put on ice. 2 µl of the ligation mixture was added and then it was kept on ice for 30 minutes. The cells were then heat shocked at 42 °C for 45 seconds, then put on ice for two more minutes. Next 950 µl of 37 °C Lysogeny broth (LB) (10 g dm^{-3} tryptone, 5 g dm^{-3} yeast extract and 5 g dm^{-3} NaCl) media was added and the bacteria was cultured at 37 ° C for 1-2 hours. The culture was then centrifuged and 850 µl of the fluid removed. The bacteria were then re-suspended, spread on an LB agar plate (section 6.2.9) and incubated overnight at 37 °C.

Once the mutant CAP gene was ligated into the pQE30 vector, 1 ml of super optimal broth with catabolite repression (SOC) (20 g dm^{-3} tryptone, 5 g dm^{-3} yeast extract, 0.5 g dm^{-3} NaCl, 0.186 g dm^{-3} KCl, 0.952 g dm^{-3} $MgCl_2$ and 3.603 g dm^{-3} glucose) was used instead of LB and it was cultured at 32 °C for 1-2 hours. It was then grown on LB agar plates overnight at 32 °C. The reason for this is highlighted in section 6.3.2.

6.2.8 Media Preparation

All growth media was mixed according to manufacturer's instructions, then autoclaved at 121 °C for 15 minutes to sterilise. Any antibiotics were added at the time the media was being used.

6.2.9 LB Agar Plate Preparation

A stock solution of LB agar was made by autoclaving a mixture of 35 g LB agar powder per litre of H_2O at 121 °C for 15 minutes. 25 ml of LB agar per plate was then extracted from the stock in sterile conditions and selection agents were added to prevent non-specific growth before being poured into a Petri dish. To melt the LB agar stock for future uses it was micro waved.

6.2.10 Colony PCR

Approximately five cultures from an overnight agar plate were chosen to investigate whether they contained the insert. A small sample of each colony from the plate was added to separate copies of the PCR mixture from table 6.5. to check the directionality of the insert from the Zero Blunt® TOPO® PCR cloning reaction, two PCR reactions were run per colony.

Component	**Volume / µl**
Buffer (10 × NH_4)*	2.5
dNTPs	2.5
$MgCl_2$	0.75
Forward Primer	1
Reverse Primer	1
BIOTAQ Red DNA polymerase	1
H_2O	16.25

Table 6.5: Polymerase Chain Reaction mixtures with a total volume of 25 µl used for colony PCR. *(200 mM Tris-HCl, 100 mM KCl, 100 mM $(NH_4)_2SO_4$, 20 mM $MgSO_4$, 1 mg ml^{-1} nuclease-free BSA, 1% Triton® X-100).

The PCR solutions were heated to 95 °C for 2 minutes for initial denaturation. The PCR cycle of 30 seconds at 95 °C (denaturation), 30 seconds at 55 °C (primer annealing) and 30 seconds at 72 °C (elongation) was repeated 30 times before a final elongation step for 5 minutes at 72 °C. The reaction mixtures were then held at 4 °C until required.

To check whether the mutated CAP gene was present in the PCR mixtures, (and hence the *E. coli* colony) agarose electrophoresis was used (section 6.2.4).

6.2.11 Plasmid DNA Purification

An 8 ml overnight LB culture of the *E. coli* shown to contain the plasmid by colony PCR was centrifuged at 5000 g for 10 minutes. The supernatant was discarded and the plasmid purified from the pellet using a commercial miniprep kit, following the manufacturer's instructions.

6.2.12 Restriction Digest

A restriction digest reaction, removing the inserted gene from the pCR^{TM}-Blunt II-TOPO$^{®}$ vector (15 µl pCR-Blunt-CAP, 2 µl HindIII, 1 µl BamHI, 4 µl 10× BamHI buffer(100 mM Tris-HCl, 50 mM $MgCl_2$, 1 M KCl, 0.2% Triton X-100, 1 mg ml^{-1} BCA, pH 8.0), 18 µl H_2O) was set up on ice in parallel to a restriction digest incising the pQE30 expression vector (10 µl pQE30, 2 µl HindIII, 1 µl BamHI, 1 µl CIAP, 4 µl BamHI buffer, 22 µl H_2O). The restriction digest reaction solutions were incubated at 37 °C for 2 hours. Agarose electrophoresis followed by extraction was then used to separate out the CAP gene and the pQE30 vector.

6.2.13 DNA Ligation with T4 DNA Ligase

The mutant CAP insert and pQE30 vector were ligated using T4 DNA ligase, inserting the CAP gene into the BamHI and HindIII sites of the pQE30 vector. A ligation solution was made to a total volume of 10 µl. This mixture contained 1 µl of T4 DNA ligase and 1 µl of 10× buffer (300 mM Tris-HCl pH 7.8, 100 mM $MgCl_2$, 100 mM DTT and 10 mM ATP). The rest of the solution was composed so that the mass ratio of the insert to vector was approximately 3:1. The solution was made up to a final volume of 10 µl with water. Where possible, the final concentration of the vector (measured using a NanoDrop 2000) was no less than 10 ng ml^{-1}.

The ligation mixture was then transformed into *E. coli* (section 6.2.7), and then colony PCR with both kanamycin and ampicillin as selection agents (section 6.2.10) followed by agarose gel electrophoresis (section 6.2.5) were used to determine whether the ligation and transformation were successful.

Two successful cultures were then chosen to make 8 ml SOC overnight cultures with both kanamycin and ampicillin as selection agents. The plasmid DNA was then purified using a miniprep kit (section 6.2.11), which was then sent off for DNA sequencing to see if the sequence of the insert was correct and in the correct reading frame.

A plasmid with the correct insert was then transformed into the expression strain of *E. coli*; (M182 Δ CAP [pREP4]) and plated on agar plates with kanamycin and ampicillin. Colony PCR was then used to confirm successful transformation.

6.2.14 Protein Expression

A 100 ml overnight starter culture of Terrific Broth (TB) (21 g dm^{-3} tryptone, 24 g dm^{-3} yeast extract, 9.4 g dm^{-3} K_2HPO_4, 2.2 g dm^{-3} KH_2PO_4 and 8 ml dm^{-3} glycerol) was grown at 32 °C with the antibiotics kanamycin and ampicillin.

10 ml of this starter culture was then added to each of six flasks containing 1 litre of LB broth containing both kanamycin and ampicillin as selection agents. These were then incubated in a shaker (180 rpm) at 37 °C until the optical density at a wavelength of 600 nm (OD_{600}) reached 0.6. Protein expression was induced by the addition of 1 ml of 1 M IPTG to each flask. The cultures were incubated for a further two hours at 37 °C.

The bacteria was harvested by centrifugation (4500 *g*, 4 °C) for 15 minutes. Each 3 litres of LB culture made one *E. coli* pellet. Each pellet was re-suspended in 30 ml bacterial wash buffer (50 mM Tris-HCl pH 8.5, 1 mM EDTA), then centrifuged (5500 *g*, 4 °C) for 20 minutes. The supernatant was then discarded and the pellet frozen at -80 °C until needed.

6.2.15 Protein Purification

The frozen *E. coli* pellets were thawed out in a 37 °C water bath, then kept on ice. Lysis of the *E. coli* cells was performed with a cOmplete, Mini, EDTA-free Protease Inhibitor Cocktail Tablet (Roche Applied Science) dissolved in 2 ml lysis buffer (section 6.1.10). More lysis buffer was then added to each pellet to make a total volume of 30 ml. The pellets were then re-suspended and the cells disrupted by sonication for 150 seconds before being centrifuged (50000 g, 4 °C) for 45 minutes.

The protein was then purified using nickel-chelated sepharose affinity. 5 µl samples were collected of the supernatant and flow through for electrophoresis (section 6.2.17 and 6.2.18). Here the His-tagged target protein selectively bound to the nickel chelated sepharose in the column, while other proteins passed through the column. A number of wash steps including one wash (wash 1) using elution buffer (section 6.1.10) and at least two washes using wash buffer (wash 2+) were performed to remove unwanted contaminant protein. The washes were monitored for protein content by adding a 20 µl sample of the flow through to 1 ml Bradford reagent. 5 µl samples of the wash steps were kept for analysis by electrophoresis.

The target protein was then eluted by the addition of 5 ml elution buffer. The efficacy of the protein purification was checked using either SDS-PAGE gel electrophoresis (section 6.2.17) or Amersham^{TM} ECL^{TM} gel electrophoresis (section 6.2.18).

A buffer exchange with storage buffer was performed using a Zeba^{TM} Desalt Spin Column (Thermo Scientific), following the manufacturer's instructions. The protein was then stored at -20 °C until needed.

To check that the correct protein had been expressed, MS was used. The MS was performed by the MS service at Durham University using electrospray ionization (ESI) to ionise the proteins. For crystallisation trials, the protein was purified further using a Superdex 75 16/60 size exclusion column (GE Healthcare).

6.2.16 Protein Concentration Determination

The absorbance of the protein solution at a wavelength of 280 nm was measured with a NanoDrop 2000 spectrophotometer with Nanodrop 2000 software version 1.4.2 (Thermo Scientific), using the mean of five repeat measures. The protein concentration was then determined using the Beer-Lambert Law and a molar extinction coefficient of 20,065 M^{-1} cm^{-1} at 280 nm [59].

6.2.17 SDS-PAGE Gel Electrophoresis

A pre-prepared SDS-PAGE gel was placed in a gel tank with 500 ml SDS buffer. Samples collected from the protein purification were prepared by adding 10 µl H_2O, 5 µl LDS and 1 µl DTT, and then boiled for 3-5 minutes to denature the protein. Lanes of the gel were loaded with the prepared samples and the gel was run at 120 V. Afterwards, the protein bands were stained using InstantBlue^{TM} protein stain.

6.2.18 Amersham™ ECL™ Gel Electrophoresis

Amersham™ ECL™ gels are pre-packed gels for performing polyacrylamide gel electrophoresis (PAGE). Samples were prepared in the same manner as for SDS gels. The Amersham™ ECL™ gel was placed in the gel box and a voltage of 180 V run across it for 30 minutes. The lanes were then loaded with the prepared samples and the gel run at 180 V. Afterwards, the protein bands were stained using InstantBlue™ protein stain.

6.2.19 Degassing Samples

Samples for ITC were placed in small test tubes with a small magnetic stirrer into a MicroCal™ ThermoVac machine. Samples were then degassed in the machine at a temperature of 25 °C for approximately 10 minutes with the stirring mechanism running.

6.2.20 Isothermal Titration Calorimetry (ITC)

The protein was dialysed into ITC buffer (see section 6.1.10) at 4 °C and the concentration of protein in this solution then determined (section 6.2.16).

A stock of cyclic adenosine monophosphate (cAMP) in degassed ITC buffer (section 6.2.19) was made to a concentration of 20 mM. A working solution of cAMP with a concentration between 1 mM and 5 mM was then made by diluting the stock solution in degassed ITC buffer. The concentration of cAMP used was altered depending on the concentration of the protein available, and calculated using the Beer-Lambert Law and a molar extinction coefficient of 14,650 M^{-1} cm^{-1} at 259 nm [59].

Both the protein and the cAMP solutions were de-gassed before being used for ITC. The ITC machine was prepared by cleaning and loading the samples as described in the manufacturer's manual, loading the cAMP into the syringe and the protein into the cell.

For the experiment a MicroCal iTC200 (with control software version 1.25.5) was used to implement 40 consecutive injections of 1 µl of ligand into 202 µl protein. The data for the first injection was discarded, as this was subjected to possible diffusion between the syringe and the protein solution prior to data collection.

Ligand binding for cAMP to CAP was described by a sequential three-site model; with 2 major and 1 minor binding site [251]. The free ligand concentration of each injection, [L], was calculated for each injection using the bisection method; allowing the calculation of the fraction of protein in each bound state, F_i.

$$[\mathrm{L}]_t = [\mathrm{L}] + \sum_{i=1}^{3} iF_i \tag{6.2.1}$$

$$F_i = \frac{\left(\prod_i K_i\right)[L]^i}{\sum_{j=0}^{3}\left([L]^j \prod_{i=0}^{j} K_i\right)} \quad , \quad i = 0, 3 \tag{6.2.2}$$

The binding constants, K_i, and binding enthalpies, ΔH_i, were determined by fitting the calculated heat constant,

$$Q = [P]_t V_0 \sum_{i=1}^{3}\left(\sum_{j=1}^{i} \Delta H_j\right) \tag{6.2.3}$$

to the experimental value using the solver package on Microsoft® Excel®.

6.2.21 Protein Crystallisation

Two separate protein concentrations (7 mg ml^{-1} and 14 mg ml^{-1}) were used for crystallisation trials. Automated 96 well sitting drop vapour diffusion plates were prepared using an Innovadyne Screenmaker 96+8$^{\mathrm{TM}}$ for each protein concentration using a PACT premier$^{\mathrm{TM}}$ HT-96 screen (see section 6.1.11) [246].

Crystallisation screens were also set up in 24 well hanging drop vapour diffusion plates. These screens were at pH 6.5 with 7-10% (w/v) polyethylene glycol 3350 and 15-20% (v/v) 2-methyl-2,4-pentanediol. These screens were used as it was under these conditions that WT-CAP was known to crystallise [59]. All crystallisation trays were left at room temperature for crystallisation to occur.

6.2.22 Crystal Structure Determination

Protein crystals were cryoprotected using a 1:1 mixture of mother liquor and 50% (v/v) glycerol and flash cooled in liquid nitrogen. [252] The diffraction data was collected at the Diamond Light Source and processed using Mosflm (version 7.0) [253] and Scala (version 3.3) [254].

The intensity of diffraction peaks decrease rapidly with resolution, so the signal to noise ratio can be monitored as an indicator of data quality. The signal to noise ratio for a resolution shell, $\langle |I|/\sigma(I) \rangle$, is determined by summing the intensities I over the standard deviations $\sigma(I)$ of all N reflections in the shell:

$$\langle |I|/\sigma(I) \rangle = \frac{1}{N} \sum_{\mathbf{h}}^{N} \frac{\left| I_{(\mathbf{h})} \right|}{\sigma\left(I_{\mathbf{h}}\right)} \qquad (6.2.4)$$

Another common indicator of data quality when merging N occurrences of reflection $\mathbf{h}$ in a resolution shell, is the linear merging R-value, R_{merge}:

$$R_{\text{merge}} = \frac{\sum_{\mathbf{h}} \sum_{i=1}^{N} \left| I_{(\mathbf{h})i} - \bar{I}_{(\mathbf{h})} \right|}{\sum_{\mathbf{h}} \sum_{i=1}^{N} I_{(\mathbf{h})i}} \qquad (6.2.5)$$

where the summation is for all N repeat observations of a reflection $\mathbf{h}$. $\bar{I}_{(\mathbf{h})}$ is the averaged intensity of each reflection. The merging of weak reflections in higher resolution shells, causes a rapid increase in the R_{merge}, due to an increase in relative error.

Both of the values of $\langle |I|/\sigma(I) \rangle$ and R_{merge} were therefore used to help determine which data set was used to build the structural model from.

The phase problem was solved and the initial structural model was built using molecular replacement (as described in section sub:phase-problem) with Phaser (version 2.1) using the WT-CAP structure(PDB 1G6N [107]).

Model building and refinement was then completed iteratively using *Coot* [219] and Refmac (version 5.5) [197] in CCP4 (version 6.1) [254]. For the initial refinement, 10 steps of rigid body refinement was used. Restrained refinement was used for all future steps, using bond length, bond angle and B-factor restraints as described in section 3.1.7. *Coot* was then used after every 10 to 20 steps to manually refine the structure.

During the automatic refinement, the quality of the model in comparison to the diffraction data was quantified by the R-value (see equation 3.1.27). In the refinement, 95% of the data was used for the fitting of the model, so only this data was included in calculation of the R-value. The free data was used to calculate R_{free}, which helps determine if the model was being over fit with too many parameters [200]. Other measures used for the quality of the model were the root mean square values for bond lengths, bond angles and chiral volumes.

When the model was being built graphically using *Coot*, both the the combined electron density map $(2F_{\text{obs}} - F_{\text{calc}})\exp(i\phi_{\text{calc}})$ and the basic difference map $(F_{\text{obs}} - F_{\text{calc}})\exp(i\phi_{\text{calc}})$ were used. Both of these maps are introduced in section 3.1.7. The combined electron density map was used for most of the visual modelling, with the difference map clearly identifying where the model and the observed electron densities differed. For the majority of the modelling a σ value of 1 was used for the electron density maps, meaning that only the electron density that was one standard deviation above the mean was visualised.

A number of other tools were used with *Coot* when building the model, including "Real space refine zone", a tool which automatically fit and refined a region of the protein based on the electron density map, "Regularise zone", a tool which optimises the bond lengths and angles of a region of the protein, tools for changing residue rotamers and a number of other tools for performing transformations. *Coot* was also used to build the Q108A mutations into the model and mutate any residues where there was no side chain electron density to alanine residues to improve the model. As well as using the electron density maps, the quality of the model was monitored using the Ramachandran plot to ensure the backbone dihedrals were within the allowed region, a density fit plot, a geometry analysis plot and a rotamer likelihood analysis plot.

6.3 Results and Discussion

6.3.1 Mutagenesis

Three PCRs were run to include the desired mutation in the CAP gene (see section 6.2.3). The DNA chains synthesised by the first two PCRs were identified in and extracted from the agarose gel (figure 6.2). After the second PCR, two bands appeared in the agarose gel, although only one was expected. These were both removed from the gel separately and analysed separately. Only one of these samples (the larger band) was then taken further into the cloning process as the *E. coli* containing this insert showed positive results in colony PCR, whereas the colony PCR for the other band exhibited ambiguous results (figure 6.3).

Whenever the mutant CAP gene was ligated into one of the vectors used, colony PCR was used to check whether it was successful. For the ligation into the pCR$^{\text{TM}}$-

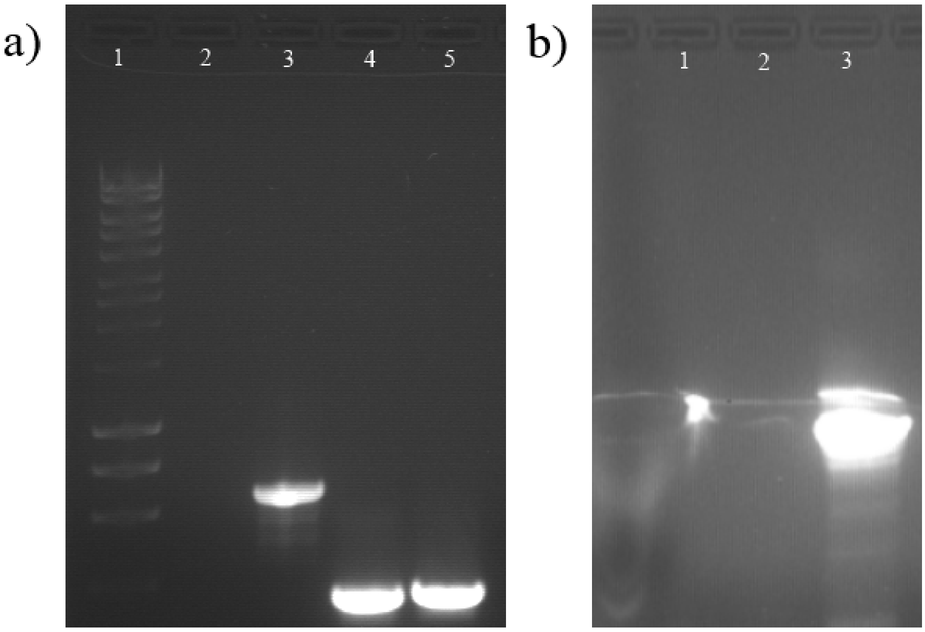


Figure 6.2: (a) Agarose gel electrophoresis UV images of the first two PCRs: (1) 250 base pair DNA ladder, (2) negative control, (3) positive control, (4) PCR2, PCR1 (table 6.3). (b) UV image from the agarose gel electrophoresis of PCR3 solution: (1) positive control (2) negative control and (3) PCR3 solution. In this case a split in the gel caused a bad image, and hence the positive control is only visible as a spot. A DNA ladder was also accidentally omitted from the gel.

Blunt II-TOPO® plasmid, the directionality of the insert was also checked at the same time by using the M13 reverse primer. Figure 6.3 shows the agarose gel for these PCRs. Lanes 8 and 9 both appeared to show positive results, implying that for this colony the insert got ligated into two plasmids which were both in opposite directions. The *E. coli* colony used for this spot was then used to continue the cloning process. Row B showed inconclusive results, so these were ignored.

Restriction digests of this plasmid were set up, and the mutant CAP gene was separated out using agarose gel electrophoresis and ligated into pQE30. The insert was checked, using colony PCR and DNA sequencing.

6.3.2 Protein Expression and Purification

A problem arose at this stage, as at some point during the cloning procedures, a mutation occurred removing part of the promoter sequence from the pQE30 vector preventing the mutated CAP gene from expressing (confirmed by DNA sequencing). To try to prevent this occurring, SOC was used and cultures were grown at 32 °C rather than using LB at 37 °C, which had previously been used for other *E. coli*

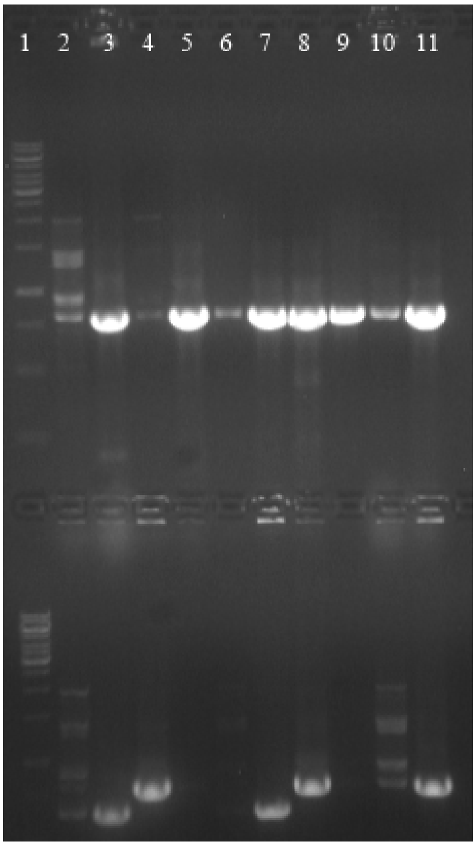


Figure 6.3: Colony PCR of 5 colonies after the blunt CAP gene was ligated into pCR^{TM}-Blunt II-TOPO$^{®}$. The left hand row of lanes (row A) correspond to the colonies grown from the large band from figure 6.2(b) and the right hand row (row B) to the colonies grown from the small band. Lane pairs correspond to one PCR using the F-CAP$_{1-6}$ primer and one PCR using the R-CAP$_{205-212}$ primer for each colony, both with the M13 reverse primer. Lane 1 contains the 250 base pair DNA ladder.

cultures. When SOC was used, the expression problems were solved, and Q108A CAP was successfully purified (see figure 6.4). The concentration of protein obtained was typically around 10 mg ml^{-1}.

6.3.3 ITC

The binding model as described in section 6.2.20 was fitted to data from 8 ITC experiments. An example fit can be seen in figure 6.5. The thermodynamic parameters calculated in this manner are exhibited in table 6.6. The allostery coefficient for this variant is determined to be $\frac{K_{a1}}{K_{a2}} = 1.8$, which is similar to the value seen for WT-CAP (1.6) and is within the error for the experimental technique (see table 6.7). This confirms that as was predicted by the ENM, the variant Q108A does not affect the allostery in CAP.

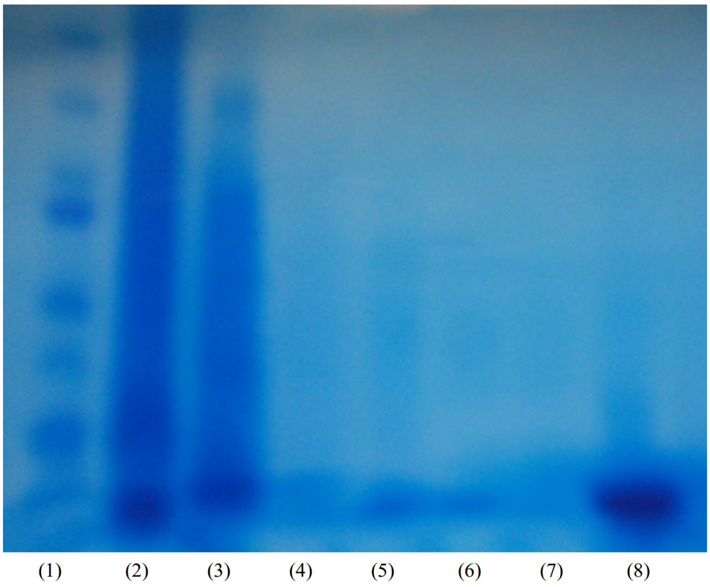


Figure 6.4: Protein electrophoresis (using Amersham ECL gel) of protein purification solutions (from left to right): (1) protein Ladder, (2) *E. coli* lysis solution, (3) flow through, (4) wash 1, (5) wash 2, (6) wash 3, (7) wash 4, (8) eluted protein.

Run No.	$\boldsymbol{K}_{\mathrm{a1}}$ **/ M^{-1}×10^{4}**	$\boldsymbol{K}_{\mathrm{a2}}$ **/ M^{-1}×10^{4}**	$\boldsymbol{\Delta H_1}$ **/ kJ mol^{-1}**	$\boldsymbol{\Delta H_2}$ **/ kJ mol^{-1}**
1	22.0	9.18	-2.66	35.1
2	19.4	12.9	-8.30	44.1
3	10.6	9.32	-13.7	42.7
4	7.20	7.67	-2.66	35.1
5	19.0	11.1	-4.68	42.3
6	18.7	10.9	-8.16	43.5
7	25.3	9.22	-4.64	43.1
8	16.5	7.55	-12.1	2.96
Avg.	17.40 (2.1)	9.41 (0.5)	-6.94 (1.47)	39.6 (2.00)

Table 6.6: Binding association constants and binding enthalpies for CAP Q108A as determined by ITC. The standard error of the mean for the average thermodynamic values are given in parentheses adjacent to the mean.

A comparison of the thermodynamic parameters for this variant with the other variants is discussed in the following subsection.

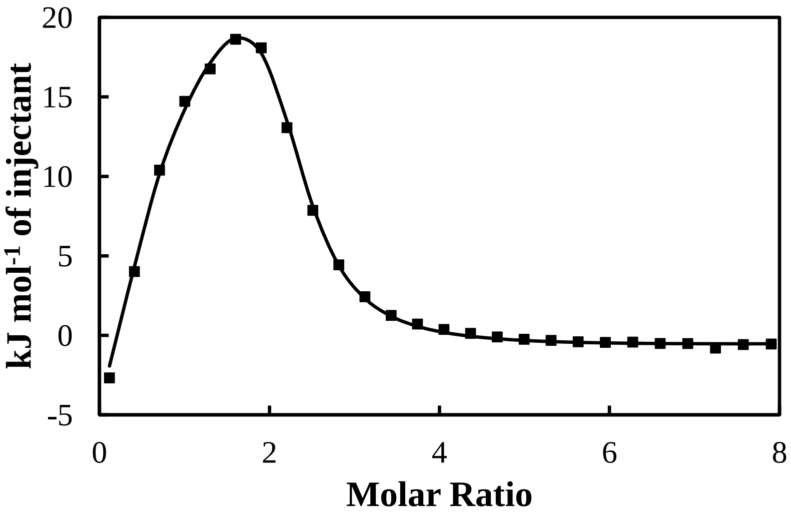


Figure 6.5: An experimental ITC curve shown as black squares with the corresponding modelled curve for one of the runs for Q108A.

6.3.4 Comparison of ITC Results for all Variants

ITC was performed on the other variants by Phil Townsend, revealing the thermodynamic properties of cAMP binding for these variants of CAP. These thermodynamic parameters can be seen in table 6.7. This table shows that as was predicted by the ENM, mutating residue V132 can manipulate the allosteric binding of cAMP. The V132A mutation pushes the system further into negative cooperativity from the usual weakly negatively cooperativity of WT-CAP (making $\Delta\Delta G$ larger) and the V132L mutation makes the system a positively cooperative system (with $\Delta\Delta G < 0$). The V132A variant reduces the hydrophobic interactions between the two monomers as it has a smaller hydrophobic group for the interaction to occur across the interface compared to WT. The V132L has the opposite effect, increasing the hydrophobic interactions between the two chains. This effect can be seen in figure 6.6. For V132L, this is the equivalent of increasing k_{xx} in the SCG models or increasing the relative spring constants of the springs originating from residue 132 in the ENM. The opposite is true for the V132A mutation.

Variant	$\boldsymbol{\Delta H_1}$ /	$\boldsymbol{\Delta H_2}$ /	$\boldsymbol{\Delta\Delta H}$ /	$\boldsymbol{-T\Delta S_1}$ /	$\boldsymbol{-T\Delta S_2}$ /	$\boldsymbol{-T\Delta\Delta S}$ /
	kJ mol^{-1}	**kJ mol^{-1}**	**kJ mol^{-1}**	**kJ mol^{-1}**	**kJ mol^{-1}**	**kJ mol^{-1}**
WT	-8.97 (0.31)	34.3 (0.99)	43.3 (1.22)	-19.9 (0.29)	-62.1 (1.02)	-42.2 (1.21)
V132A	-11.1 (1.06)	2.88 (0.82)	16.7 (3.14)	-16.3 (1.96)	-29.3 (1.34)	-13.0 (3.22)
V132L	-7.80 (0.39)	25.4 (0.53)	33.2 (0.89)	-22.8 (0.51)	-57.4 (0.56)	-34.6 (0.98)
V140A	-20.3 (1.27)	52.6 (1.59)	72.8 (2.76)	-6.08 (1.64)	-84.4 (1.60)	-78.3 (3.16)
V140L	-9.26 (0.24)	30.0 (0.54)	39.3 (0.63)	-20.7 (0.35)	-58.9 (0.34)	-38.2 (0.66)
H160L	-12.0 (0.45)	40.3 (0.34)	52.3 (0.76)	-16.6 (0.45)	-68.2 (0.34)	-51.6 (0.78)
Q108A	-6.94 (1.47)	39.6 (2.00)	46.6 (2.46)	-22.8 (1.57)	-68.0 (2.10)	-45.2 (2.51)

Variant	$\boldsymbol{K_{a1}}$	$\boldsymbol{K_{a2}}$	$\boldsymbol{K_{a1}/K_{a2}}$	$\boldsymbol{\Delta G_1}$	$\boldsymbol{\Delta G_2}$	$\boldsymbol{\Delta\Delta G}$
	M^{-1} × 10^4	**M^{-1} × 10^4**		**kJ mol^{-1}**	**kJ mol^{-1}**	**kJ mol^{-1}**
WT	12.2 (1.0)	7.5 (0.5)	1.6 (0.2)	-28.9 (0.14)	-27.7 (0.12)	1.15 (0.18)
V132A	14.1 (1.3)	3.1 (0.3)	4.8 (0.6)	-29.1 (0.24)	-25.4 (0.26)	3.76 (0.35)
V132L	24.6 (1.6)	42.5 (2.1)	0.58 (0.05)	-30.6 (0.17)	-32.0 (0.13)	-1.39 (0.21)
V140A	10.6 (0.6)	19.6 (0.9)	0.59 (0.04)	-28.3 (0.18)	-30.3 (0.09)	-1.96 (0.20)
V140L	20.7 (2.1)	13.5 (1.4)	1.6 (0.2)	-30.0 (0.25)	-28.9 (0.25)	1.07 (0.35)
H160L	10.2 (0.4)	7.7 (0.2)	1.3 (0.1)	-28.5 (0.08)	-27.8 (0.08)	0.71 (0.11)
Q108A	17.4 (2.1)	9.4 (0.5)	1.8 (0.2)	-29.7 (0.36)	-28.3 (0.13)	1.38 (0.38)

Table 6.7: Experimental thermodynamic parameters for WT CAP and its variants; obtained by ITC. Mean values are given for the first and second cAMP binding events and the allosteric quantities, with the standard error of the mean provided in parentheses.

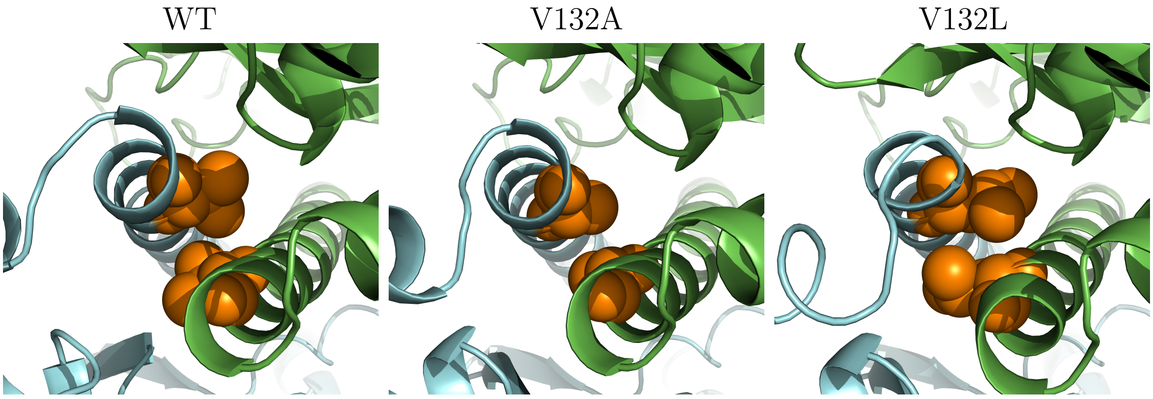


Figure 6.6: Hydrophobic interactions across the dimer interface for WT, V132A and V132L, showing residue 132 as orange spheres.

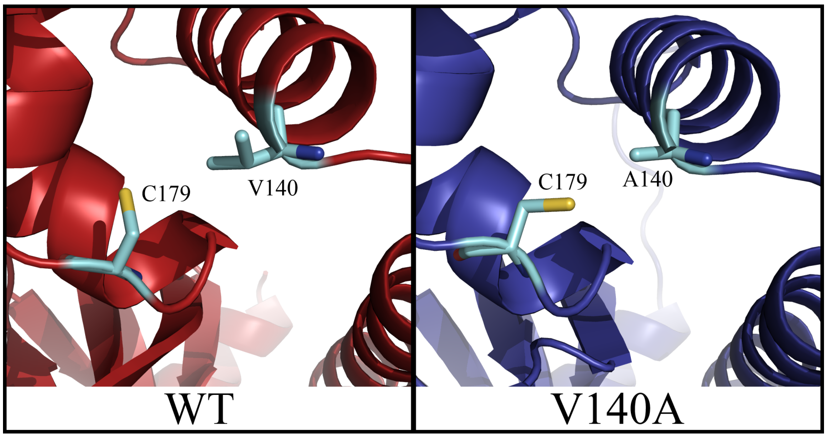


Figure 6.7: Local structure of residue Cys179 in WT CAP compared to V132A.

The H160L mutation also reduces the negative cooperativity of the system when compared to WT-CAP. It does not however, make CAP positively cooperative as is the case for V132L. This is again in agreement with the predictions made by the ENM.

As for the V140 mutations, the ENM predicted that mutating this residue would not affect the allostery of the system. For the V140L mutation, this is seen to be the case. However, for the V140A mutation, the allostery of the system becomes positively cooperative, contrary to the prediction made by the ENM. This change to the allostery cannot be explained by a global change in structure as is discussed in section 6.3.7. Nevertheless, there is a significant local conformational change in the crystal structure; a rotation of the side chain of the nearby Cys179 residue. This local structural change, shown in figure 6.7, allows the mutated V140A residue to form a new hydrophobic interaction with residue Cys179. Including this mutation in the ENM by using the spring constant $k_{\mathrm{Cys179}}/k = 4$ for the springs connected to Cys179, the system was pushed towards positive cooperativity. This therefore qualitatively justifies the change in allostery seen for the V140A variant.

The description of allostery in CAP is actually more complex than just looking at the changes in conformational entropy, unlike what the ENM indicates. It appears in table 6.7, that both the entropic and enthalpic contributions are equally important for the allostery of the system and enthalpy-entropy compensation occurs to some extent. For example, for all of the variants of CAP, the enthalpic contribution to the allosteric free energy is always positive (which would indicate negative cooperativity). On the other hand, the entropic contribution to the free energy ($-T\Delta\Delta S$) is always negative (which treated alone would indicate positive cooperativity). This is contrary to what was observed by Kalodimos *et al.* when they studied the truncated form of CAP [17]. They made the conclusion that allostery in the truncated form of CAP was dynamically driven as only the entropic term led to the negative cooperativity of the system, the opposite to the results in this section.

In this system, the enthalpy and entropy contributions to the allosteric free energy are a similar magnitude, with opposite signs, leading to a weakly positive or weakly negative allosteric free energy. The entropic contributions to the allosteric free energy do not scale in the same way as the allosteric free energy calculated using the ENM (which calculates the free energy as a crude enthalpic adjustment to the vibrational entropy); mutations that are not supposed to make changes to the allostery in CAP do make changes to the absolute values of the enthalpy and the entropy terms. It is only when these contributions are combined that there is no change from the allosteric free energy of WT-CAP. This and the differences observed between the full and truncated forms of CAP imply that the story behind the allostery of full length CAP is more complicated than being just entropically or enthalpically driven. Rather is the consequence of enthalpy-entropy compensation.

Looking at the energy contributions separately, it is possible to compare these ITC results to the entropies calculated from the atomistic simulations. This is performed in the following subsection.

Variant	$T\Delta\Delta S^{\text{vib}}_{\text{NMA}}$ / kJ mol^{-1}	$T\Delta\Delta S_{\text{ITC}}$ / kJ mol^{-1}
WT	8.38	42.2
V132A	-2.39	13.0
V132L	6.29	34.6
V140A	0.79	78.3
V140L	7.45	38.2
H160L	6.11	51.6

Table 6.8: Comparison of NMA entropies to ITC entropies. The vibrational entropies as determined by NMA ($T\Delta\Delta S^{\text{vib}}_{\text{NMA}}$) are displayed with the entropies determined by ITC ($T\Delta\Delta S_{\text{ITC}}$)

6.3.5 Comparison of ITC Results to Atomistic Simulations

In chapter 5 a couple of techniques, MM/PBSA and NMA were used to try calculate the allosteric free energy for the variants of CAP.

NMA

Firstly, the contribution of the conformational entropy to the allosteric signalling in CAP was calculated using NMA on snapshots of a long MD simulation. Table 6.8 shows how these simulation results compared to the ITC results. It is evident that the entropies calculated by NMA have a lower magnitude when compared to the ITC entropies. This can be explained, as the NMA calculations only return the conformational entropies, whereas the entropies determined using ITC contain all of the extra entropic contributions such as the effects of the solvent. Additionally, only the first 200 modes were used for the calculation of the entropy by NMA, which was before the entropy had converged.

There only appears to be any correlation between the ITC results and the NMA results when ignoring the data point for the V140A variant. The best fit line for this reduced data set 6.8 has a Pearson's correlation coefficient of 0.857. The variant V140A, however shows no agreement between the two methods; including this data point when calculating the correlation (solid line) reduces the Pearson's correlation coefficient to 0.149. This variant also did not behave as expected in the ITC when compared to the ENM (see section 6.3.4). In that case, the difference was explained by a structural

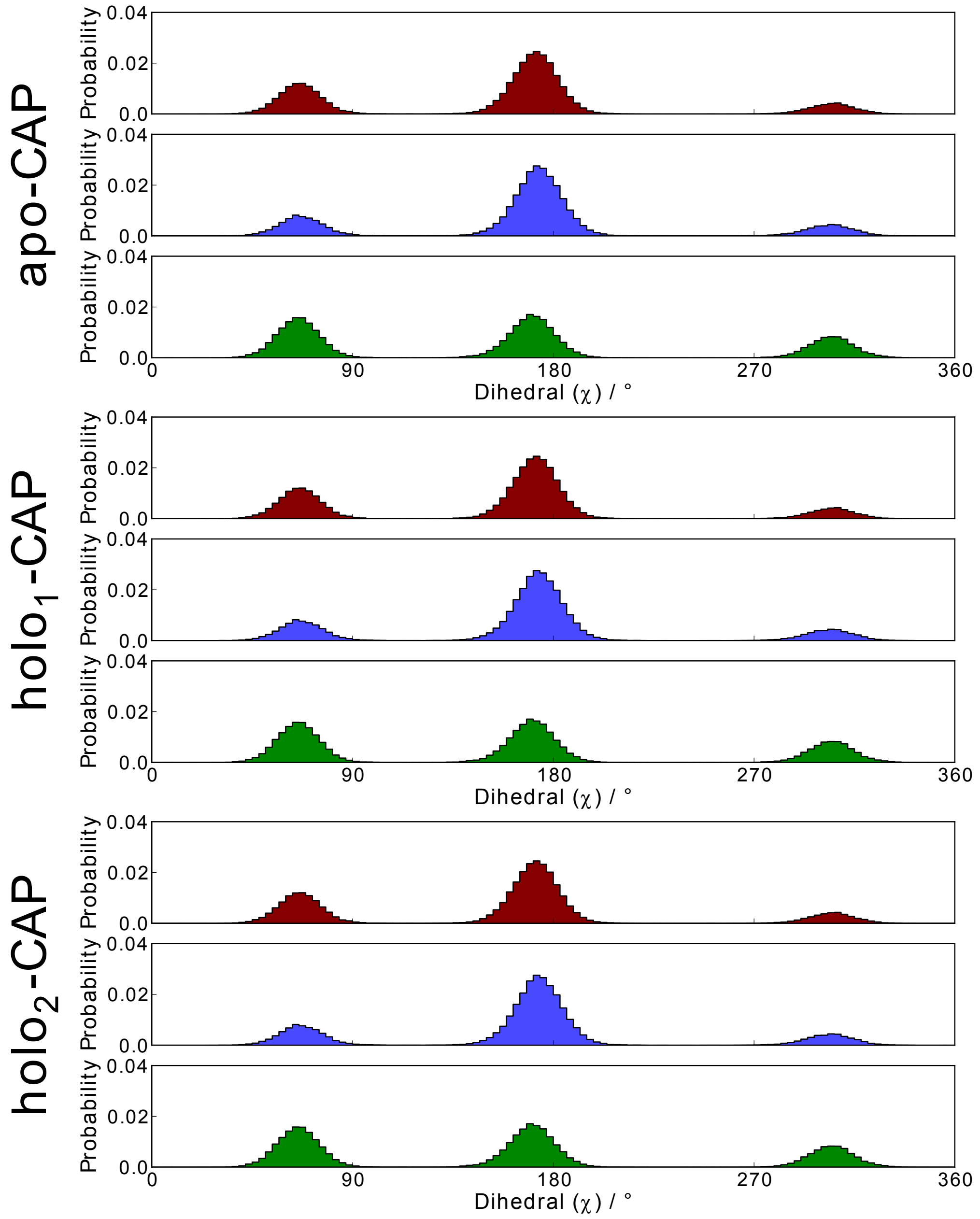


Figure 6.8: The C-C$_\alpha$-C$_\beta$-S$_\gamma$ dihedral angle for Cys179 of WT, V132A and V132L. 🟥 WT, 🟦 V140A, 🟩 V140L

change that had not been captured by the ENM.

To investigate if the reason that the MD shows different results to the ITC is because it does not capture the structural change seen in the X-ray crystallography, the C179 residue was investigated further for the WT, V140A and V140L simulations.

To distinguish between the two different orientations of Cys179, the C-C$_\alpha$-C$_\beta$-S$_\gamma$ dihedral was investigated. For WT (4HZF) and V140L (4I02), this dihedral has a value around -50 °. The PDB for V140A contains 6 monomers, of which only one has a dihedral around -50 °with the other 5 chains having the Cys179 residue in a different orientation with dihedrals of around 165 °. This local structural difference between the WT crystal structure and the V140A crystal structure can be seen in figure 6.7 and is discussed earlier in section 6.3.4.

The distribution of this dihedral angle (χ) was then plotted for all three bound states of WT, V132A and V132L CAP as shown in figure 6.8. This figure shows that for all of the simulations, three different dihedral angle distributions are well sampled; one centred at around 170 °, one centred at around 60 ° and one centred at around 310 ° (or -50 °). What is quite surprising is that for all of the simulations, the least sampled dihedral angle in the simulations is the -50 ° dihedral; the dihedral observed most in the X-ray structures.

The -50 ° dihedral of Cys179 gets sampled slightly more for V140L, but not significantly. Also, the V140A variant does appear to sample the $\chi = 160$ ° orientation for Cys179, however the differences between the dihedral distributions are almost negligible when compared to the large differences seen in the crystal structures. This therefore does not confirm that the Cys179 residue samples the 160 ° dihedral significantly more for the V140A mutation than the other variants. In fact, it appears as if none of the variants sample the experimentally observed value for this dihedral angle to the extent seen in the experimental data, which could suggest that this region of the protein is not properly represented by the ff99SBildn force field.

The explanation as to why the entropy for the V140A variant is so different to the others and the value determined by ITC is therefore still unknown.

MM/PBSA

MM/PBSA was used to try determine the other contributions to the allosteric free energy, as was discussed in 5.2.8. Combining $\Delta\Delta G_{\text{MM/PBSA}}$ (the contributions to the allosteric free energy calculated using MM/PBSA) and $T\Delta\Delta S$ (the vibrational entropy determined with NMA) the resulting $\Delta\Delta G_{\text{COMP}}$ can be compared to the ITC values for each of the CAP variants($\Delta\Delta G_{\text{EXP}}$). This is shown in table 6.9 using the energy values determined with the SASA/vdW model for the nonpolar contribution to the solvation energies in the MM/PBSA calculations.

This table shows that there is not any convincing agreement between the computational values of $\Delta\Delta G$ and the experimental values. For example, using computa-

Variant	$\Delta\Delta G_{\mathrm{COMP}}$	$\Delta\Delta G_{\mathrm{EXP}}$
WT	29.94	1.15
V132A	-18.67	3.76
V132L	-1.99	-1.39
V140A	-3.76	-1.96
V140L	10.85	1.07
H160L	27.38	0.71

Table 6.9: Comparison between computational and experimental $\Delta\Delta G$ values. The computational values ($\Delta\Delta G_{\mathrm{COMP}}$) are determined using MM/PBSA and NMA, whilst the experimental values ($\Delta\Delta G_{\mathrm{EXP}}$) are determined using ITC.

tional methods, V132A is predicted to exhibit strongly positive allostery, not negative allostery as expected. As discussed in 5.2.8, the MM/PBSA contributions to the allosteric free energy carry large errors, hence the lack of agreement. A large contribution to these errors can probably be attributed to the simulations not being properly converged as is discussed in section 5.2.6. These simulations would not have been able to explore as much of phase space as is necessary to get convincing results. As well as this, the values calculated for the contributions to the vibrational entropy seemed a little low. As discussed earlier in this section, this could be because not enough normal modes are included in the calculation.

The lack of agreement with the computational values with experiment after 300 ns of simulation makes one wonder how long a simulation would need to be run to achieve proper convergence and whether force fields are accurate enough over such long time scales, where errors could creep in. Also, even when there are fairly small errors in calculating the free energy of one system, the percentage error is much larger when looking at an energy difference and larger still when looking at the difference between two energy differences as is the case when studying $\Delta\Delta G$. These arguments all point towards MM/PBSA perhaps not being an accurate enough method for investigating cooperative binding.

6.3.6 Crystal Structure Determination of Q108A CAP

Two manual crystal screens using hanging drop crystal trays (described in section 6.2.21) exhibited no crystallisation in any of the wells. However, crystallisation did

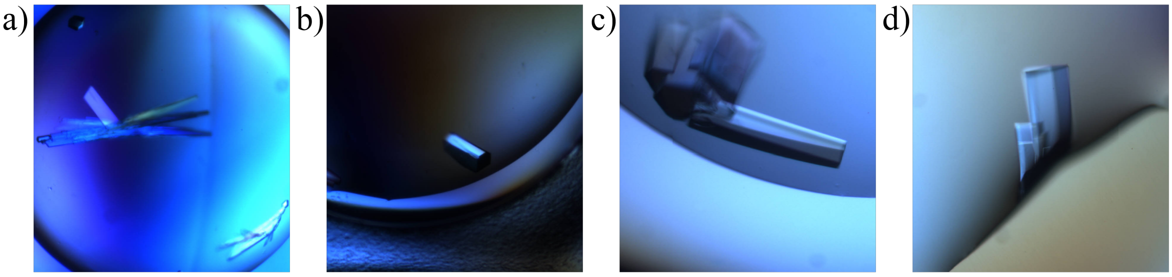


Figure 6.9: Crystals of Q108A CAP grown in an PACT *premier*$^{\text{TM}}$ screen, with conditions: (a) 0.2 M sodium bromide, 0.1 M Bis Tris Propane, 20% (w/v) PEG 3350 pH 6.5. (b) 0.2 M sodium iodide, 0.1 M Bis Tris propane, 20% (w/v) PEG 3350 pH 8.5. (c) 0.2 M sodium nitrate, 0.1 M Bis Tris propane, 20% (w/v) PEG 3350 pH 8.5. (d) 0.2 M potassium/sodium tartrate, 0.1 M Bis Tris propane, 20% (w/v) PEG 3350 pH 8.5.

Resolution / Å	R_{merge}	$\langle I \rangle$	$\sigma(I)$	$\langle \lvert I \rvert / \sigma(I) \rangle$	N_{meas}	N_{ref}
5.63	0.032	9578	633	30.1	7234	1337
3.98	0.049	12603	855	30.2	12395	2299
3.25	0.044	8424	618	25.9	14569	2974
2.81	0.055	3414	303	20.5	18119	3606
2.52	0.069	1840	211	15.7	20566	4120
2.30	0.090	1260	186	12.5	22812	4538
2.13	0.115	988	189	9.8	25141	4957
1.99	0.182	651	178	7.1	26974	5236
1.88	0.293	348	164	4.4	29109	5566
1.78	0.486	202	159	2.7	30611	5810

Table 6.10: Statistics from the resolution shells for the CAP Q108A variant diffraction pattern output by Scala [254]. $\langle I \rangle$, $\sigma(I)$, $R_{\mathbf{merge}}$ and $\langle |I|/\sigma(I) \rangle$ are described in section 6.2.22, whilst N_{meas} and N_{ref} are number of measurements and the number of independent reflections respectively.

occur in a number of wells of the automated 96 well sitting drop screens. Figure 6.9 shows images of a number of crystals from the automated screen with the conditions used to promote crystallisation.

A selection of these crystals (plus a few more) were taken to the Diamond Light Source, where X-ray diffraction data was collected. MOSFLM was used to process the diffraction patterns to a resolution of 1.78 Å. The diffraction peaks were integrated and analysed using Scala [254]. An analysis of the statistics of the data against resolution can be seen in table 6.10.

Parameter	**Value**
Resolution / Å	1.78
Wavelength	0.9795
No. reflections	41452
Completeness / %	99.4
a / Å	46.0
b / Å	93.7
c / Å	104.0
α / °	90.0
β / °	90.0
γ / °	90.0
No. protein atoms	3121
No. ligand atoms	66
No. water atoms	315
No. protein chains	2
Space group	$P2_12_12_1$
R-value	0.187
R_{free}-value	0.239
RMS bond length / Å	0.0159
RMS bond angle / °	1.57
RMS chiral volume / Å^3	0.137
Average B-factor	29.3

Table 6.11: Crystallographic data collection and refinement statistics for the crystal structure of CAP Q108A.

The value for R_{merge} obtained for the highest resolution shell for CAP Q108A was 0.486 and $\langle|I|/\sigma(I)\rangle$ was 2.7. A value of R_{merge} less than 0.5 and a value of $\langle|I|/\sigma(I)\rangle$ greater than 2 for the highest resolution shell are usually considered acceptable [255]. Therefore, they indicate that the data collected is of high enough quality to use to solve the crystal structure.

The final crystal structure contained one CAP dimer in the asymmetric unit and was in the $P2_12_12_1$ space group, which is the same space group as the V132A, V132L and V140L mutants [59]. Both of the chains were complete, contained no missing loop residues, however at the termini, 5 residues from the N terminus and 4 residues from the C terminus were unresolved from chain A. Chain B had 7 residues from the N terminus and 3 residues from the C terminus unresolved. As well as the two monomers, the unit cell also contained three cAMP molecules and 315 water molecules.

The parameters defining the unit cell and the refinement parameters for the crystal structure (R-value,R_{free}-value, RMS values and average B-factor) are shown in table

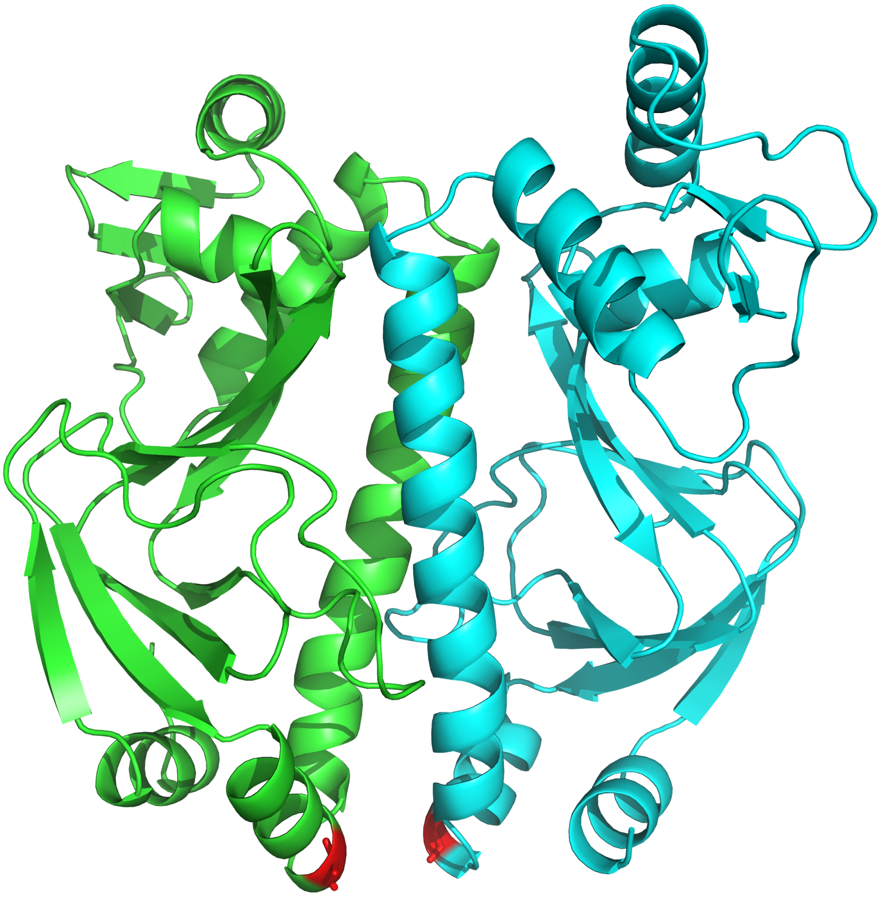


Figure 6.10: The crystal structure of CAP Q108A. Residue 108 is coloured red to indicate the position of the mutation.

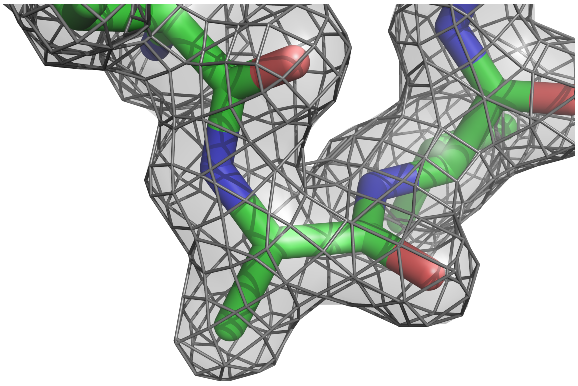


Figure 6.11: A close-up view of the mutated residue, 108, in the crystal structure of the Q108A variant of CAP. The $2F_{\text{obs}} - F_{\text{calc}}$ electron density map shown here at a level of $1\,\sigma$ indicates the presence of the mutation. Should the mutation be missing a large region of electron density would be present corresponding to the glutamine residue in the WT.

6.11.

The solved crystal structure can be seen in figure 6.10 and a close up of the mutated residue with electron density can be seen in figure 6.11. There is very little global structural change of CAP when mutating this residue, as is discussed more in 6.3.7.

6.3.7 Comparison of Crystal Structures for all Variants

The crystal structure for holo$_2$-CAP was solved for WT, V132A, V132L, V140A, V140L and H160L by Phil Townsend. The RMSD comparison between the WT CAP and the variants were calculated for the full length protein and just the LBD using the Kabsch algorithm [232]. These RMSD calculations were performed using PyMOL [234]. The RMSD values and the number of atoms used in the calculations can be seen in table 6.12. This shows that the RMSD between WT and the variants is less than 2 Å for all variants except V132L, which has an RMSD of 2.80 Å. These values are low enough to show that there no major changes to the global structure of CAP upon making the mutations.

The RMSDs between the LBDs for the variants and WT-CAP are much lower than for full length CAP; below 0.8 Å for all of the variants. This again shows that the main differences between the structures of the CAP variants are small rotations in the DNABD. The variant crystal structures are displayed in figure 6.12, showing how similar the structures are and the small changes to the orientations of the DNABDs. The next section relates these X-ray structures to the MD simulations and the NMA first shown in chapter 5.

PDB	**RMSD$_{\mathbf{full}}$**	**n$_{\mathbf{full}}$**	**RMSD$_{\mathbf{LBD}}$**	**n$_{\mathbf{LBD}}$**
	/ Å		/ Å	
V132A (4I0A)	1.90	1575	0.53	1024
V132L (4I09)	2.80	1582	0.80	1030
V140A (4I02{A,E})	1.54	1566	0.51	1025
V140A (4I02{B,C})	1.49	1569	0.55	1028
V140A (4I02{D,F})	1.40	1573	0.49	1028
V140L (4I01)	1.72	1581	0.55	1028
H160L (4I0B)	1.99	1570	0.53	1021
Q108A	1.90	1576	0.59	1028

Table 6.12: RMSDs of the X-ray structures of CAP variants from the aligned WT structure using the backbone atoms for all overlapping residues in the calculation; performed using PyMOL [234]. The number of overlapping atoms used in the calculation differs for each variant and is displayed as $\mathrm{n}_{\mathrm{full}}$. The RMSDs of only the LBD is also included ($\mathrm{RMSD}_{\mathrm{LBD}}$), where the LBDs of the variant and the reference WT structure are aligned before the RMSD calculated for the backbone atoms. $\mathrm{n}_{\mathrm{LBD}}$ shows the number of atoms used in the calculation. PDB IDs of structures available on the PDB are provided in parentheses.

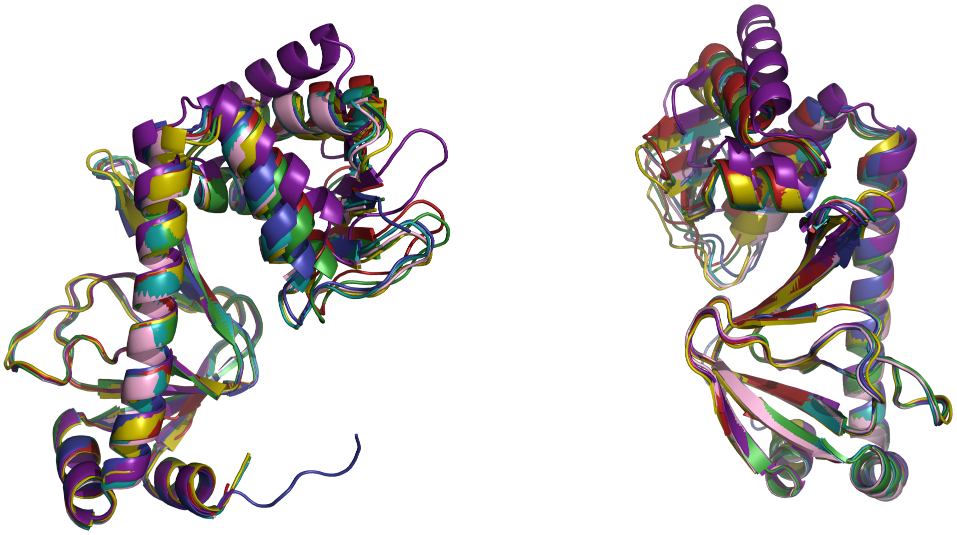


Figure 6.12: Overlay of the X-ray crystal structures of one monomer of the CAP variants. The alignment is performed on the LBDs of both chains using PyMOL [234]. Chain A for each variant is shown: ■ WT, ■ V132A, ■ V132L, ■ V140A, ■ V140L, ■ H160L and ■ Q108A.

6.3.8 Comparison of Protein Crystallography Results to Atomistic Simulations

The crystal structures of all of the variants studied in this thesis are all very similar, and the main differences in structure are due to rotation of the DNABDs relative to the LBDs. These differences are similar to the differences between the final structures of the variants after 300 ns of MD simulation. This section investigates whether these rotational differences observed between the different variant crystal structures can be described by the motions observed by the atomistic techniques; NMA and PCA.

NMA

First NMA was investigated by looking at how similar the first six normal modes of WT holo$_2$-CAP are to the differences seen between the variant crystal structures. This was done by calculating difference vectors between all nine PDBs to return a set of 36 vectors. These difference vectors were calculated by subtracting the coordinates of one PDB from the coordinates of the other after all of the structures had been aligned to the WT structure. The overlaps of these difference vectors with each of the first 6 normal modes were then calculated (using C_α coordinates). The structure from the NMA also had to be aligned with the WT crystal structure and the eigenvectors transformed by the same amount before the overlaps were calculated. These alignments and overlaps were performed using the author's own code written in Python (given in appendix C); the alignments performed using the Kabsch algorithm [232] and the overlaps determined using the equations described in section 2.5.3.

The results of this analysis are shown in table 6.13. In this table the normal modes that have a significant overlap with the difference vectors between two PDBs are highlighted. The definition of a 'significant overlap' in this study is considered to be an overlap of 0.5 or above, a value where the motion described by the normal modes will be similar to the difference vectors between the two PDBs.

This table shows that most of the PDBs tend to follow the 4th mode closest of all the normal modes calculated. Figure 6.13 shows the vector of the 4th normal mode and how it compares to the difference vector between the PDBs 4I09 (V132L) and 4I0B (H160L), which have a high overlap of 0.790. The figure shows vector arrows to compare the direction of the motion for both vectors and a plot of the residue wise

PDB		Mode Number						$s(\mathbf{u}_{1-6}, \mathbf{W})$
PDB1	**PDB2**	**1**	**2**	**3**	**4**	**5**	**6**	
A	B	0.443	0.308	0.174	0.669	0.077	0.145	0.892
A	C	0.552	0.015	0.228	0.705	0.060	0.109	0.932
A	D	0.180	0.548	0.413	0.157	0.413	0.075	0.839
A	E	0.195	0.475	0.417	0.211	0.405	0.121	0.813
A	F	0.022	0.447	0.387	0.209	0.433	0.148	0.777
A	G	0.480	0.219	0.101	0.585	0.106	0.350	0.874
A	H	0.002	0.313	0.392	0.634	0.041	0.014	0.810
A	I	0.430	0.272	0.112	0.694	0.108	0.155	0.888
B	C	0.499	0.372	0.218	0.500	0.018	0.024	0.828
B	D	0.250	0.103	0.415	0.658	0.205	0.169	0.865
B	E	0.240	0.044	0.401	0.675	0.187	0.193	0.865
B	F	0.339	0.011	0.359	0.649	0.185	0.199	0.860
B	G	0.014	0.229	0.173	0.291	0.040	0.366	0.550
B	H	0.269	0.003	0.132	0.791	0.072	0.079	0.853
B	I	0.042	0.189	0.357	0.202	0.194	0.073	0.499
C	D	0.393	0.235	0.384	0.676	0.134	0.129	0.921
C	E	0.387	0.193	0.377	0.692	0.123	0.147	0.919
C	F	0.453	0.167	0.347	0.672	0.123	0.151	0.918
C	G	0.427	0.193	0.272	0.570	0.006	0.170	0.805
C	H	0.387	0.136	0.024	0.790	0.062	0.071	0.895
C	I	0.506	0.328	0.295	0.452	0.025	0.008	0.810
D	E	0.054	0.528	0.054	0.280	0.131	0.252	0.666
D	F	0.555	0.484	0.205	0.117	0.062	0.209	0.803
D	G	0.244	0.194	0.348	0.544	0.190	0.313	0.805
D	H	0.120	0.105	0.616	0.436	0.316	0.064	0.836
D	I	0.239	0.128	0.360	0.676	0.176	0.177	0.850
E	F	0.716	0.213	0.211	0.057	0.017	0.076	0.782
E	G	0.233	0.133	0.335	0.564	0.172	0.334	0.802
E	H	0.128	0.046	0.622	0.412	0.307	0.094	0.824
E	I	0.229	0.070	0.347	0.693	0.158	0.200	0.850
F	G	0.335	0.099	0.297	0.545	0.172	0.339	0.807
F	H	0.012	0.011	0.601	0.431	0.317	0.109	0.812
F	I	0.327	0.037	0.308	0.667	0.157	0.205	0.845
G	H	0.280	0.072	0.192	0.745	0.088	0.195	0.849
G	I	0.001	0.171	0.049	0.386	0.031	0.360	0.558
H	I	0.257	0.022	0.164	0.790	0.089	0.084	0.856

Table 6.13: Overlap of (none zero) NMA modes to difference vectors between two PDBs. PDBs used are: A: WT (4HZF), B: V132A (4I0A), C: V132L (4I09), D: V140A (4I02 chains A and E), E: V140A (4I02 chains B and C), F: V140A (4I02 chains D and F), G: V140L (4I01), H: H160L (4I0B), I: Q108A (unpublished). The difference vectors were calculated by aligning all of the structures to the WT structure and subtracting PDB2 from PDB1. The subspace overlap, $s(\mathbf{u}_{1-6}, \mathbf{W})$, (see equation 2.5.61 in section 2.5.4), taking into account the overlap between the first six normal modes, $\mathbf{u}_{1-6}$, and the PDB difference vector, $\mathbf{W}$, is also shown for each difference vector.

cross correlation of these vectors in real space. The residue-wise cross correlation was calculated with an adaptation of equation 2.5.58 using the individual PDB difference vector between the two PDBs and looking at only the C_α atoms in the calculation. These cross correlations are shown in figure 6.13 in real space, superimposed onto the structure of PDB 4HZF. These show that for most regions of the protein, the fourth normal mode from the NMA describes the differences between the two variant structures very well, with only small regions where the two vectors do not correlate well.

Table 6.13 also contains a column for the subspace overlap, $s(\mathbf{u}_{1-6}, \mathbf{W})$, defined by equation 2.5.61 in section 2.5.4. This value quantifies how much of the difference between the two PDBs can be described by the six modes in the table. If all $3N-6$ normal modes of a protein with N atoms were used, the subspace overlap would have a value of one as the normal modes form an orthogonal basis. Therefore, a good measure of how much a reduced set of the normal modes contribute to the motion seen in a reference vector is how close this subspace overlap is to a value of 1 for this reduced set.

The value of the subspace overlap for the first six modes from NMA appears to be over 0.8 for most of the difference vectors between the different PDBs, showing that these six modes capture most of motion necessary for the structural changes between each of the PDBs to occur. This implies that the structural differences seen between the variants are again not large global changes and are instead more likely due to the motion along the lowest energy motion getting frozen out at different points along the mode during crystal packing.

PCA

The same analysis was performed on the first six modes calculated on the 200 ns WT holo$_2$-CAP MD simulation using PCA. In this case, the mode which best describes the differences seen in the PDBs is the second mode, as can be seen in table 6.14. However, the overlap of this mode with the PDB difference vectors is typically not as great as was seen with the top mode calculated with NMA. Only 9 PDB difference vectors having an overlap greater than 0.5 for mode 2 from PCA compared to the 22 difference vectors that have an overlap greater than 0.5 for mode 4 from NMA. This

PDB		**Mode Number**						$\boldsymbol{s(\mathbf{u}_{1-6}, \mathbf{W})}$
PDB1	**PDB2**	**1**	**2**	**3**	**4**	**5**	**6**	
A	B	0.031	0.527	0.289	0.188	0.093	0.106	0.646
A	C	0.022	0.562	0.195	0.203	0.075	0.208	0.667
A	D	0.046	0.218	0.145	0.047	0.275	0.158	0.416
A	E	0.052	0.210	0.101	0.095	0.257	0.176	0.404
A	F	0.039	0.156	0.143	0.156	0.196	0.139	0.358
A	G	0.034	0.451	0.258	0.318	0.150	0.150	0.646
A	H	0.600	0.419	0.197	0.048	0.099	0.121	0.775
A	I	0.018	0.532	0.236	0.195	0.097	0.129	0.635
B	C	0.083	0.404	0.001	0.157	0.030	0.267	0.517
B	D	0.001	0.301	0.140	0.191	0.103	0.193	0.441
B	E	0.003	0.305	0.168	0.216	0.085	0.197	0.463
B	F	0.005	0.334	0.144	0.241	0.038	0.165	0.468
B	G	0.130	0.240	0.112	0.207	0.086	0.063	0.376
B	H	0.343	0.579	0.058	0.082	0.003	0.138	0.694
B	I	0.290	0.080	0.290	0.052	0.026	0.140	0.444
C	D	0.033	0.391	0.097	0.201	0.061	0.252	0.521
C	E	0.034	0.395	0.118	0.220	0.049	0.256	0.536
C	F	0.028	0.413	0.102	0.236	0.015	0.231	0.539
C	G	0.002	0.464	0.057	0.025	0.021	0.192	0.506
C	H	0.291	0.600	0.045	0.119	0.004	0.202	0.708
C	I	0.017	0.383	0.063	0.146	0.027	0.235	0.478
D	E	0.021	0.091	0.278	0.271	0.152	0.072	0.433
D	F	0.038	0.269	0.048	0.338	0.336	0.108	0.561
D	G	0.055	0.204	0.096	0.274	0.071	0.217	0.426
D	H	0.483	0.504	0.067	0.076	0.106	0.005	0.713
D	I	0.044	0.304	0.098	0.196	0.099	0.208	0.442
E	F	0.062	0.264	0.141	0.219	0.301	0.183	0.515
E	G	0.057	0.209	0.125	0.297	0.053	0.221	0.450
E	H	0.490	0.505	0.101	0.109	0.089	0.014	0.725
E	I	0.046	0.308	0.126	0.220	0.081	0.212	0.461
F	G	0.047	0.242	0.102	0.322	0.006	0.189	0.459
F	H	0.512	0.471	0.081	0.147	0.043	0.016	0.717
F	I	0.037	0.337	0.104	0.245	0.035	0.180	0.468
G	H	0.401	0.534	0.025	0.154	0.024	0.165	0.706
G	I	0.028	0.286	0.010	0.200	0.082	0.013	0.360
H	I	0.367	0.569	0.028	0.084	0.003	0.146	0.698

Table 6.14: Overlap between PCA normal modes and the difference vectors between two PDBs. PDBs used are: A: WT (4HZF), B: V132A (4I0A), C: V132L (4I09), D: V140A (4I02 chains A and E), E: V140A (4I02 chains B and C), F: V140A (4I02 chains D and F), G: V140L (4I01), H: H160L (4I0B), Q108A (unpublished). The subspace overlap, $s(\mathbf{u}_{1-6}, \mathbf{W})$, between each individual difference vector, $\mathbf{W}$, and the six PCA modes, $\mathbf{u}_{1-6}$, is also shown.

is not surprising as the NMA was performed on one of the crystal structures, whereas in the MD simulation the protein was free to explore a larger region of conformational space. However, there is still significant overlap between the difference vector between H160L and the other PDBs and the second mode.

Figure 6.13 shows the vector for the second principal component as arrows for comparison with the PDB difference vector. The residue wise cross correlation between these two vectors is also displayed, showing that again large regions of the protein move in a similar manner to the difference seen between the X-ray structures. It does highlight that there is less agreement with PCA to experiment than with NMA however.

The subspace overlap between the H160L PDB and all of the other variants is around 0.7 (see table 6.14), showing that the first six principal components are responsible for most of the differences seen between these PDBs. The subspace overlap with the principal components is also around 0.5 for the difference vectors between most of the PDBs, which indicates that although the lowest modes from PCA do not capture as much of the motion seen as NMA, they do still capture a large proportion of the motion in the first six modes. Increasing the number of modes investigated would only increase this subspace overlap.

Therefore both NMA and PCA are able to capture a large proportion of the motion that would be required to cause the differences in structure seen in all of the variants of CAP. However, it appears that the NMA method is able to capture the motion more effectively.

6.3.9 Overall Discussion

In this chapter, the thermodynamic parameters of the cooperative binding of cAMP to CAP were investigated using ITC. The variants, V132A, V132L and H160L, predicted to affect allostery using the ENM in chapter 4 were shown by ITC to affect allostery in the same way as predicted. Similarly, variants V140L and Q108A predicted by the ENM to not affect allostery were also shown to not affect allostery by ITC. There was some agreement between the entropies determined by atomistic NMA, although not quite to the extent of the agreement seen between the ENM and ITC. There was less agreement still between the thermodynamic parameters determined using MM/PBSA

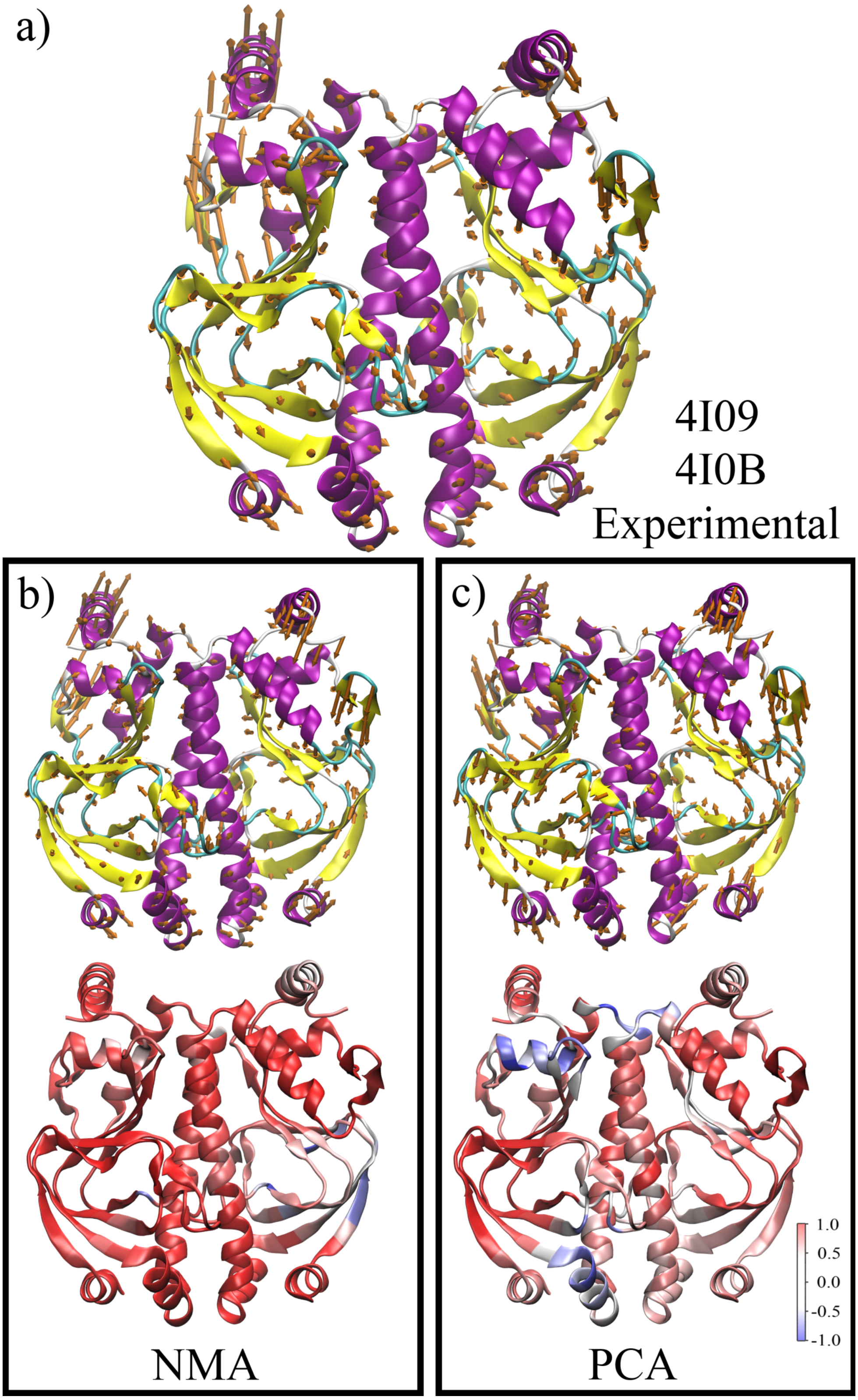


Figure 6.13: Comparison of difference vectors between PDBs 4I09 (V132L) and 4I0B (H160L) to normal modes from NMA and PCA. a) The difference vectors between the two X-ray crystal structures. b) and c) show the eigenvectors from a single normal mode as calculated from each theoretical method above a plot of the residue wise cross correlation (see section 2.5.2) between this vector and the PDB difference vector in a); red corresponding to a cross correlation of 1 and blue corresponding to a cross correlation of -1. b) shows the 4th none zero normal mode calculated by NMA and c) shows the 2nd normal mode calculated by PCA. All arrows are calculated to a length that would give the protein an RMSD of 2 Å.

on atomistic simulations and the experimental values. The errors using these particular atomistic methods to calculate the allosteric free energy difference has the potential of being significantly large, so if future study were to be performed on this system, another more accurate method would be recommended. Repeating the chosen method a large number of times would also be required to get statistically significant average values.

The X-ray structure for the Q108A variant of holo$_2$ CAP was solved. It was shown that there were no major structural differences between this variant and any of the other variants of holo$_2$ CAP studied. Although this does not directly show allostery without conformational change, combined with the ITC results it shows that the allostery in CAP can be modified without changing the global structure of the protein.

The main structural differences that were observed between the holo$_2$ CAP variants were rearrangements of the DNABDs relative to the LBDs, again showing the flexibility of the hinge region. These structural differences were shown to agree with the motions seen in the principal components of atomistic MD simulations and NMAs; agreeing more with the modes seen in NMA than PCA.

Chapter 7

Conclusions and Further Work

7.1 Conclusions

In this thesis, SCG models and the ENM were used to show that WT CAP is on a region of the allosteric free energy landscape that is very sensitive to small changes in the interactions within the protein's LBD. An advantage of CAP being located at this region of the free energy landscape is that changes in the environment that make an unfavourable change to the cooperative binding of cAMP to CAP could be counteracted by a mutation in CAP. Being at this location on the free energy landscape allows CAP to adapt when there are changes to the environment. The SCG models (and to a lesser extent the ENM) also showed that CAP was much less sensitive to changes in the interactions within the protein's DNABDs. This has the consequence that changes to the homotropic allostery with the allosteric effector cAMP (which binds to the LBDs) are unlikely to affect the DNA recognition helix, keeping these two mechanisms separate.

The ENM was also used to identify individual residues, which when mutated were likely to affect the allostery within CAP. Variants of CAP incorporating these point mutations were investigated using atomistic computational methods and experimental techniques. ITC for example showed a selection of the variants identified by the ENM to alter the allostery within CAP, did in fact change the allostery as predicted. NMA and MM/PBSA were used together to try determine the allosteric free energies for these variants computationally. Although the computed entropies appeared to show a fair agreement with the ITC values, the enthalpies showed less of an agreement and the errors with the values were argued to be too large for the trends seen to be statistically significant.

Atomistic MD simulations were used to study some of the global and local motions observed within the variants of CAP, with a potential pathway for the homotropic allostery within CAP being identified through the Thr128 residues of each monomer.

The protein motion observed through MD simulations and NMA calculations was also shown to agree with the ensemble of structures for the variants of CAP observed using X-ray crystallography. The inherent flexibility of the hinge region (residues 132-138) was seen, with both simulation and crystallography. Indicating a range of orientations of the DNABD relative to the LBD are possible.

Although the overall goal of showing that allostery within CAP occurs without a significant conformational change could not be proved, this thesis did show that it could be manipulated without a significant conformational change. It also showed that both the enthalpy and entropy were important contributions to the cooperativity of binding cAMP and the resulting allosteric free energy was a result of enthalpy-entropy compensation. The following further work section outlines some additional work that could be done that could help tackle some of the questions that are still open after this thesis.

7.2 Further Work

In chapter 4, SCG models were used to study the allostery in CAP. It was mentioned that it would be possible to use these models to study other allosteric homodimers by adjusting the parameters of the model, and a larger range of proteins could be covered by including the springs excluded in the current model.

The ENM, also discussed in chapter 4, allowed the addition of mutations to the protein by adjusting spring constants. However with this model, all residues are treated equally except the mutated residue. It would be interesting is to investigate an ENM that has varying cut-offs for residues with different length side chains; longer side chains naturally corresponding to larger cut-offs. Additionally, the spring constants could be parametrised so they are different depending on the residues they connect. The point mutations could then be made based on the residue parametrisations. Similar ENMs with variable spring constants have previously been studied to look at motion in proteins [256, 257], so there would be a good starting point for such a study.

As was discussed in chapter 5, the force field ff12SB is a new AMBER force field. It was only used to study the WT variant of CAP. In the studies in this thesis, it did not appear to simulate apo-CAP as would be expected. However, a more comprehensive study of the force field compared to the other older AMBER force fields would be useful for the entire scientific community.

The computational methods used in this thesis to determine the allosteric free energy, NMA and MM/PBSA, were found to not be of sufficient accuracy to determine the small free energy difference involved in allostery. In a future study, other methods such as thermodynamic integration could be used to determine the free energy of the cAMP-CAP binding events. Such simulations could be run over a range of temperatures to determine entropic changes. Other alternatives could be steered molecular dynamics or umbrella sampling. These methods are much more computationally expensive than the methods used in this study, and as a lot of repeats are needed for statistically significant results, they were deemed too computationally expensive for this study. However, with increasing GPU support for these types of calculations with each AMBER release, running long studies on proteins of the size of CAP or larger, using computationally expensive methods such as these is becoming increasingly possible.

Future experimental work could include continuing with the target of obtaining a good quality crystal structure of apo-CAP. This could help answer a lot of the remaining questions about whether the allostery in CAP occurs without a conformational change.

Finally, in this thesis it was shown that it was possible to manipulate allostery in a system that already exhibits negative allostery without a conformational change by making point mutations in the protein. An interesting future study could be to see if allostery can also be manipulated by binding another ligand to the protein rather than making physical changes to the protein, or in other words another allosteric event. Alternatively, a protein that does not normally exhibit cooperative binding could be manipulated so that it does show this, either by making mutations or binding ligands to distant sites on the protein.

Bibliography

[1] C. Branden and J. Tooze. *Introduction to Protein Structure. - 2nd Ed.* Garland Science, 1999. [p. 3, 5]

[2] G. Weber. Ligand binding and internal equilibiums in proteins. *Biochemistry*, 11:864–878, 1972. [p. 3, 6, 21]

[3] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–99, 1963. [p. 4]

[4] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry: fourth edition.* W. H. Freeman and Company, 2005. [p. 4, 18]

[5] M. A. Mart-Renom, A. C. Stuart, A. Fiser, R. Snchez, F. Melo, and A. ali. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, 29:291–325, 2000. [p. 6]

[6] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. 3-Dimensional model of the Myoglobin molecule obtained buy X-ray analysis. *Nature*, 181:662–666, 1958. [p. 6]

[7] F. Bernstein, T. Koetzle, G. Williams, E. J. Meryer, M. Brice, J. Rodgers, K. O., T. Shimanouchi, and T. M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977. [p. 6]

[8] RSCB Protein Data Bank. *http://www.rcsb.org/*. [p. 6]

Style: [i] Authors. Article Title. *Journal*, Volume:Pages, Year. [p. Page(s) in thesis]

[9] R. H. Austin, K. W. Beeson, L. Eisenstein, H. Frauenfelder, and I. C. Gunsalus. Dynamics of ligand binding to myoglobin. *Biochemistry*, 14:5355–5373, 1975. [p. 6]

[10] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.*, 104:4546–4559, 1982. [p. 7]

[11] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *J. Am. Chem. Soc.*, 104:4559–4570, 1982. [p. 7]

[12] M. Levitt and A. Warshel. Computer-simulation of protein folding. *Nature*, 253:5494, 1975. [p. 7]

[13] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand. Conformational entropy in molecular recognition by proteins. *Nature*, 448:325–329, 2007. [p. 7]

[14] D. M. Korzhnev, X. Salvatella, M. Vendruscolo, A. A. Di Nardo, A. R. Davidson, C. M. Dobson, and L. E. Kay. Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR. *Nature*, 430:586–590, 2004. [p. 7]

[15] J. S. Fraser, M. W. Clarkson, S. C. Degnan, R. Erion, D. Kern, and T. Alber. Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, 462:669–673, 2009. [p. 7]

[16] K. A. Baker, C. Tzitzilonis, W. Kwiatkowski, S. Choe, , and R. Riek. Conformational dynamics of the KcsA potassium channel governs gating properties. *Nat. Struct. Mol. Biol.*, 14:1089 – 1095, 2007. [p. 7]

[17] N. Popovych, S. Sun, R. Ebright, and C. Kalodimos. Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.*, 13:831–838, 2006. [p. 7, 8, 25, 26, 30, 32, 33, 96, 108, 131, 196]

[18] S.-R. Tzeng and C. G. Kalodimos. Dynamic activation of an allosteric regulatory protein. *Nature*, 462:368–372, 2009. [p. 7, 27, 30]

[19] N. Tokuriki and D. S. Tawfik. Protein Dynamism and Evolvability. *Science*, 324:203–207, 2009. [p. 7]

[20] D. Kern and K. Henzler-Wildman. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007. [p. 7, 8]

[21] S. J. Benkovic and S. Hammes-Schiffer. A Perspective on Enzyme Catalysis. *Science*, 301:1196–1202, 2003. [p. 8, 25, 155]

[22] R. Daniel, R. Dunn, J. Finney, and J. Smith. The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.*, 32:69–92, 2003. [p. 8]

[23] B. F. Volkman, D. Lipson, D. E. Wemmer, and D. Kern. Two-State Allosteric Behavior in a Single-Domain Signaling Protein. *Science*, 291:2429–2433, 2001. [p. 8]

[24] A. J. Wand. On the Dynamic Origins of Allosteric Activation. *Science*, 293:1395, 2001. [p. 8]

[25] Q. Cui and M. Karplus. Allostery and cooperativity revisited. *Protein Sci.*, 17:1295–1307, 2008. [p. 8, 32]

[26] E. Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.*, 27:2985–2993, 1894. [p. 8]

[27] D. E. Koshland. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.*, 44:98–104, 1958. [p. 9]

[28] J.-P. Changeux and S. J. Edelstein. Allosteric Mechanisms of Signal Transduction. *Science*, 308:1424–1428, 2005. [p. 12]

[29] R. Nussinov and C.-J. Tsai. Allostery in Disease and in Drug Discovery. *Cell*, 153:293 – 305, 2013. [p. 12]

[30] N. Smith and G. Milligan. Allostery at G protein-coupled receptor homo- and heteromers: uncharted pharmacological landscapes. *Pharmacol. Rev.*, 62:701725, 2010. [p. 12]

[31] M. Wood, C. Hopkins, J. Brogan, P. Conn, and C. Lindsley. ”Molecular switches” on mGluR allosteric ligands that modulate modes of pharmacology. *Biochemistry*, 50:2403–2410, 2011. [p. 12]

[32] A. V. Hill. Proceedings of the Physiological Society. *J. Physiol. (Lond.)*, 40:i–vii, 1910. [p. 13]

[33] K. E. van Holde, W. C. Johnson, and P. S. Ho. *Principles of Physical Biochemistry: Second Edition*. Pearson Prentice Hall, 2006. [p. 14, 18]

[34] J. Monod, J. Wyman, and J.-P. Changeux. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.*, 12:88 – 118, 1965. [p. 16, 17, 95]

[35] D. Koshland Jr, G. Neméthy, and D. Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5:365–368, 1966. [p. 16, 18, 19, 95]

[36] A. Cooper and D. Dryden. Allostery without conformational change. *Eur. Biophys. J.*, 11:103–109, 1984. [p. 17, 21, 110]

[37] A. Brzozowski, Z. Derewenda, E. Dodson, G. Dodson, M. Grabowski, R. Liddington, T. Skarzynski, and D. Vallely. Bonding of molecular oxygen to T state human haemoglobin. *Nature*, 307:74–76, 1984. [p. 20]

[38] A. Arnone, P. Rogers, N. V. Blough, J. L. McGourty, and B. M. Hoffman. X-ray diffraction studies of a partially liganded hemoglobin, [α(FeII-CO)β(MnII)]2. *J. Mol. Biol.*, 188:693 – 706, 1986. [p. 20]

[39] B. Luisi, B. Liddington, G. Fermi, and N. Shibayama. Structure of deoxy-quaternary haemoglobin with liganded β subunits. *J. Mol. Biol.*, 214:7 – 14, 1990. [p. 20]

[40] S. Ogawa, A. Mayer, and R. Shulman. High resolution proton magnetic resonance study of the two quaternary states in fully ligated hemoglobin Kansas. *Biochem. Biophys. Res. Commun.*, 49:1485 – 1491, 1972. [p. 20]

[41] J. M. Salhany, S. Ogawa, and R. G. Shulman. Correlation between quaternary structure and ligand dissociation kinetics for fully liganded hemoglobin. *Biochemistry*, 14:2180–2190, 1975. [p. 20]

[42] A. Szabo and M. Karplus. A mathematical model for structure-function relations in hemoglobin. *J. Mol. Biol.*, 72:163 – 197, 1972. [p. 20]

[43] J. Hbezfeld and H. Stanley. A general approach to co-operativity and its application to the oxygen equilibrium of hemoglobin and its effectors. *J. Mol. Biol.*, 82:231 – 265, 1974. [p. 20]

[44] G. K. Ackers and M. L. Johnson. Linked functions in allosteric proteins: Extension of the concerted (MWC) model for ligand-linked subunit assembly and its application to human hemoglobins. *J. Mol. Biol.*, 147:559 – 582, 1981. [p. 20]

[45] A. W. Lee and M. Karplus. Structure-specific model of hemoglobin cooperativity. *Proc. Natl. Acad. Sci.*, 80:7055–7059, 1983. [p. 20]

[46] M. L. Johnson, B. W. Turner, and G. K. Ackers. A quantitative model for the cooperative mechanism of human hemoglobin. *Proc. Natl. Acad. Sci.*, 81:1093–1097, 1984. [p. 20]

[47] G. K. Ackers, M. L. Doyle, D. Myers, and M. A. Daugherty. Molecular Code for Cooperativity in Hemoglobin. *Science*, 255:54–63, 1992. [p. 20]

[48] J. F. Swain and L. M. Gierasch. The changing landscape of protein allostery. *Curr. Opin. Struct. Biol.*, 16:102 – 108, 2006. [p. 21]

[49] G. Manley and J. P. Loria. NMR insights into protein allostery. *Arch. Biochem. Biophys.*, 519:223 – 231, 2012. [p. 21]

[50] D. Kern and E. R. Zuiderweg. The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.*, 13:748 – 757, 2003. [p. 21]

[51] A. del Sol, C.-J. Tsai, B. Ma, and R. Nussinov. The Origin of Allosteric Functional Modulation: Multiple Pre-existing Pathways. *Structure*, 17:1042 – 1050, 2009. [p. 21]

[52] A. Cooper, A. McAlpine, and P. G. Stockley. Calorimetric studies of the energetics of protein-DNA interactions in the E. coli methionine repressor (MetJ) system. *FEBS Letters*, 348:41 – 45, 1994. [p. 21]

[53] C.-J. Tsai, A. del Sol, and R. Nussinov. Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play. *J. Mol. Biol.*, 378:1 – 11, 2008. [p. 21]

[54] H. Toncrova and T. McLeish. Substrate-Modulated Thermal Fluctuations Affect Long-Range Allosteric Signaling in Protein Homodimers: Exemplified in CAP. *Biophys. J.*, 98:2317–2326, 2010. [p. 21, 34, 59, 60, 96, 110]

[55] H. Toncrova. *Coarse-Grained Models of Biomolecule Dynamics and Allostery.* PhD thesis, University of Leeds, 2011. [p. 21, 34, 59, 60, 96]

[56] R. Hawkins and T. C. B. McLeish. Coarse-Grained Model of Entropic Allostery. *Phys. Rev. Lett.*, 93:098104–1 – 098104–4, 2004. [p. 21, 59]

[57] R. Hawkins and T. C. B. McLeish. Coupling of Global and Local Vibrational Modes in Dynamic Allostery of Proteins. *Biophys. J.*, 91:2055–2062, 2006. [p. 21, 59]

[58] S.-R. Tzeng and C. Kalodimos. Protein activity regulation by conformational entropy. *Nature*, 488:236–240, 2012. [p. 21]

[59] T. L. Rodgers, P. D. Townsend, D. Burnell, M. L. Jones, S. A. Richards, T. C. B. McLeish, E. Pohl, M. R. Wilson, and M. J. Cann. Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR Family Transcription Factors. *PLoS Biol.*, 11:e1001651, 2013. [p. 21, 27, 28, 34, 95, 96, 99, 100, 101, 108, 110, 111, 112, 113, 171, 185, 186, 187, 202]

[60] W. L. Peticolas. Low frequency vibrations and the dynamics of proteins and polypeptides. *Methods Enzymol.*, 61:425–58, 1979. [p. 22]

[61] B. Jacrot, S. Cusack, A. J. Dianoux, and D. M. Engelman. Inelastic Neutron-scattering Analysis of Hexokinase Dynamics and Its Modification On Binding of Glucose. *Nature*, 300:84–86, 1982. [p. 22]

[62] H. D. Middendorf. Biophysical Applications of Quasi-Elastic and Inelastic Neutron Scattering. *Annu. Rev. Biophys. Bioeng.*, 13:425–451, 1984. [p. 22, 24]

[63] F. R. Gurd and T. M. Rothgeb. Motions in proteins. *Adv. Protein Chem.*, 33:73–165, 1979. [p. 22]

[64] N. Go. Thermodynamics of small-amplitude conformational fluctuations in native globular proteins. *Proc. Jap. Acad. Ser. B*, 56:414–419, 1980. [p. 22]

[65] N. Go, T. Noguti, and N. T. Dynamics of A Small Globular Protein In Terms of Low-frequency Vibrational-modes. *Proc. Natl. Acad. Sci. USA*, 80:3696–3700, 1983. [p. 22, 54]

[66] K. M. Brooks B. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci.*, 80:6571–6575, 1983. [p. 22]

[67] J. A. McCammon and M. Karplus. The dynamic picture of protein structure. *Acc. Chem. Res.*, 16:187–193, 1983. [p. 22]

[68] M. Levitt. Molecular dynamics of native protein. I. Computer simulation of trajectories. *J. Mol. Biol.*, 168:595–620, 1983. [p. 22, 24]

[69] M. Levitt. Molecular dynamics of native protein. II. Anaysis and nature of motion. *J. Mol. Biol.*, 168:621–657, 1983. [p. 22, 24]

[70] J. M. Sturtevant. Heat capacity and entropy changes in processes involving proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 74:2236–2240, 1977. [p. 22]

[71] J. A. McCammon, B. R. Gelin, M. Karplus, and P. G. Wolynes. The hinge-bending mode in lysozyme. *Nature*, 262:325–326, 1976. [p. 24]

[72] R. Yirdaw and H. Mchaourab. Direct Observation of T4 Lysozyme Hinge-Bending Motion by Fluorescence Correlation Spectroscopy. *Biophys. J.*, 103:1525–1536, 2012. [p. 24]

[73] M. Akke, R. Brueschweiler, and A. G. Palmer. NMR order parameters and free energy: an analytical approach and its application to cooperative calcium(2+) binding by calbindin D9k. *J. Am. Chem. Soc.*, 115:9832–9833, 1993. [p. 24]

[74] M. D. Daily, T. J. Upadhyaya, and J. J. Gray. Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins: Struct., Funct., Bioinf.*, 71:455–466, 2008. [p. 25]

[75] C. M. Petit, J. Zhang, P. J. Sapienza, E. J. Fuentes, and A. L. Lee. Hidden dynamic allostery in a PDZ domain. *Proc. Natl. Acad. Sci.*, 106:18249–18254, 2009. [p. 25]

[76] J. O. Wrabl, J. Gu, T. Liu, T. P. Schrank, S. T. Whitten, and V. J. Hilser. The role of protein conformational fluctuations in allostery, function, and evolution. *Biophys. Chem.*, 159:129 – 141, 2011. [p. 25]

[77] T. P. Schrank, D. W. Bolen, and V. J. Hilser. Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. *Proc. Natl. Acad. Sci.*, 106:16984–16989, 2009. [p. 25]

[78] V. J. Hilser, J. O.Wrabl, and H. N. Motlagh. Structural and Energetic Basis of Allostery. *Annu. Rev. Biophys.*, 41:585–609, 2012. [p. 25]

[79] H. N. Motlagh and V. J. Hilser. Agonism/antagonism switching in allosteric ensembles. *Proc. Natl. Acad. Sci.*, 109:4134–4139, 2012. [p. 25]

[80] T. Liu, S. T. Whitten, and V. J. Hilser. Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble. *Proc. Natl. Acad. Sci.*, 104:4347–4352, 2007. [p. 25]

[81] K. Reynolds, R. McLaughlin, and R. Ranganathan. Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell*, 147:1564 – 1575, 2011. [p. 25]

[82] A. Zhuravleva and L. M. Gierasch. Allosteric signal transmission in the nucleotide-binding domain of 70-kDa heat shock protein (Hsp70) molecular chaperones. *Proc. Natl. Acad. Sci.*, 108:6987–6992, 2011. [p. 25]

[83] A. Zhuravleva, E. M. Clerico, and L. Gierasch. An Interdomain Energetic Tug-of-War Creates the Allosterically Active State in Hsp70 Molecular Chaperones. *Cell*, 151:1296 – 1307, 2012. [p. 25]

[84] U. Gether. Uncovering Molecular Mechanisms Involved in Activation of G Protein-Coupled Receptors. *Endocr. Rev.*, 21:90–113, 2000. [p. 25]

[85] S. T. Menon, M. Han, and T. P. Sakmar. Rhodopsin: Structural Basis of Molecular Physiology. *Physiol. Rev.*, 81:1659–1688, 2001. [p. 25]

[86] F. C. Peterson, R. R. Penkert, B. F. Volkman, and K. E. Prehoda. Cdc42 Regulates the Par-6 PDZ Domain through an Allosteric CRIB-PDZ Transition. *Mol. Cell*, 13:665 – 676, 2004. [p. 25]

[87] P. K. Agarwal, S. R. Billeter, P. T. R. Rajagopalan, S. J. Benkovic, and S. Hammes-Schiffer. Network of coupled promoting motions in enzyme catalysis. *Proc. Natl. Acad. Sci.*, 99:2794–2799, 2002. [p. 25]

[88] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Boscom, J. J. Skalicky, L. E. Kay, and D. Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438:117–121, 2005. [p. 25]

[89] J. S. Fraser, M. W. Clarkson, S. C. Degnan, R. Erion, D. Kern, and T. Albe. Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, 462:669–673, 2009. [p. 25]

[90] S. W. Lockless and R. Ranganathan. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, 286:295–299, 1999. [p. 25]

[91] A. D. Ferguson, C. A. Amezcua, N. M. Halabi, Y. Chelliah, M. K. Rosen, R. Ranganathan, and J. Deisenhofer. Signal transduction pathway of TonB-dependent transporters. *Proc. Natl. Acad. Sci.*, 104:513–518, 2007. [p. 25]

[92] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell*, 138:774 – 786, 2009. [p. 25]

[93] M. E. Hatley, S. W. Lockless, S. K. Gibson, A. G. Gilman, and R. Ranganathan. Allosteric determinants in guanine nucleotide-binding proteins. *Proc. Natl. Acad. Sci.*, 100:14445–14450, 2003. [p. 25]

[94] A. I. Shulman, C. Larson, D. J. Mangelsdorf, and R. Ranganathan. Structural Determinants of Allosteric Ligand Activation in {RXR} Heterodimers. *Cell*, 116:417 – 429, 2004. [p. 25]

[95] G. Süel, S. Lockless, M. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Mol. Biol.*, 10:59–69, 2002. [p. 25]

[96] J. Monod. *Nobel Lectures, Physiology or Medicine 1963-1970.* Elsevier Publishing Company, Amsterdam, 1972. [p. 26]

[97] B. Magasanik. *The Lactose Operon.* Cold Spring Harbour Laboratory, 1970. [p. 26]

[98] A. Kolb, S. Busby, H. Buc, S. Garges, and S. Adhya. Transcriptional Regulation By cAMP and its Receptor Protein. *Annu. Rev. Biochem.*, 62:749–795, 1993. [p. 26]

[99] J. Harwood and A. Peterkofsky. Glucose-sensitive adenylate cyclase in toluene-treated cells of Escherichia coli B. *J. Biol. Chem.*, 250:4656–4662, 1975. [p. 26]

[100] B. Feucht and M. Saier Jr. Fine control of adenylate cyclase by the phosphoenolpyruvate:sugar phosphotransferase systems in Escherichia coli and Salmonella typhimurium. *J. Bacteriol.*, 141:603–610, 1980. [p. 26]

[101] S. Busby and R. H. Ebright. Transcription activation by catabolite activator protein (CAP). *J. Mol. Biol.*, 293:199 – 213, 1999. [p. 27]

[102] K. Hollands, S. J. Busby, and G. S. Lloyd. New targets for the cyclic AMP receptor protein in the Escherichia coli K-12 genome. *FEMS Microbiol. Lett.*, 274:89–94, 2007. [p. 27]

[103] S. Schultz, G. Shields, and S. T.A. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, 253:1001–1007, 1991. [p. 27, 32]

[104] W. Anderson, A. Schneide, M. Emmer, R. Perlman, and I. Pastan. Purification of and properties of cyclic adenosine 3',5'-Monophosphate receptor protein which mediates cyclic adenosine 3',5'-monophosphate-dependent gene transcription in escherichia-coli. *J. Biol Chem.*, 246:5929–5937, 1971. [p. 27]

[105] H. Won, Y. Lee, S. Lee, and B. Lee. Structural overview on the allosteric activation of cyclic AMP receptor protein. *Biochim. Biophys. Acta*, 1794:1299 – 1308, 2009. [p. 27]

[106] J. Li, X. Cheng, and J. Lee. Structure and Dynamics of the Modular Halves of Escherichia coli Cyclic AMP Receptor Protein. *Biochemistry*, 41:14771–14778, 2002. [p. 27]

[107] J. Passner, S. Schultz, and T. Steitz. Modeling the cAMP-induced allosteric transition using the crystal structure of CAP-cAMP at 2.1 angstrom resolution. *J. Mol. Biol.*, 304:847–859, 2000. [p. 27, 101, 188]

[108] H. Sharma, S. Yu, J. Kong, J. Wang, and T. Steitz. Structure of apo-CAP reveals that large conformational changes are necessary for DNA binding. *Proc. Natl. Acad. Sci. USA*, 106:16604–16609, 2009. [p. 27, 29, 30, 31, 171]

[109] J. Harman. Allosteric regulation of the cAMP receptor protein. *Biochim. Biophys. Acta*, 1547:1 – 17, 2001. [p. 28]

[110] I. T. Weber and T. A. Steitz. Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Å resolution. *J. Mol. Biol.*, 198:311 – 326, 1987. [p. 28, 29]

[111] N. Popovych, S.-R. Tzeng, M. Tonelli, R. H. Ebright, and C. G. Kalodimos. Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proc. Natl. Acad. Sci.*, 106:6927–6932, 2009. [p. 29, 30, 95]

[112] H. Won, T. Yamazaki, T. Lee, M. Yoon, S. Park, Y. Kyogoku, and B. Lee. Structural Understanding of the Allosteric Conformational Change of Cyclic AMP Receptor Protein by Cyclic AMP Binding. *Biochemistry*, 39:13953–13962, 2000. [p. 29]

[113] G. Tan, P. Kelly, J. Kim, and R. Wartell. Comparison of cAMP receptor protein (CRP) and a cAMP-independent form of CRP by Raman spectroscopy and DNA binding. *Biochemistry*, 30:5076–5080, 1991. [p. 30]

[114] H. DeGrazia, J. Harman, G. Tan, and R. Wartell. Investigation of the cAMP receptor protein secondary structure by Raman spectroscopy. *Biochemistry (Mosc.)*, 29:3557–3562, 1990. [p. 30]

[115] E. Heyduk, T. Heyduk, and J. Lee. Global conformational changes in allosteric proteins. A study of Escherichia coli cAMP receptor protein and muscle pyruvate kinase. *J. Biol. Chem.*, 267:3200–3204, 1992. [p. 30]

[116] J. Passner and T. Steitz. The structure of a CAP-DNA complex having two cAMP molecules bound to each monomer. *Proc. Natl. Acad. Sci. USA*, 94:2843–2847, 1997. [p. 32]

[117] G. Parkinson, C. Wilson, A. Gunasekera, Y. W. Ebright, R. E. Ebright, and H. M. Berman. Structure of the CAP-DNA Complex at 2.5 Resolution: A

Complete Picture of the Protein-DNA Interface. *J. Mol. Biol.*, 260:395 – 408, 1996. [p. 32]

[118] S. Chen, J. Vojtechovsky, G. Parkinson, R. Ebright, and H. Berman. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: DNA binding specificity based on energetics of DNA kinking. *J. Mol. Biol.*, 314:63–74, 2001. [p. 32]

[119] B. Benoff, C. Yang, H.and Lawson, G. Parkinson, J. Liu, E. Blatter, Y. Ebright, H. Berman, and R. Ebright. Structural basis of transcription activation: the CAP-alpha CTD-DNA complex. *Science*, 297:1562–1566, 2002. [p. 32]

[120] M. Takahashi, B. Blazy, and A. Baudras. An equilibrium study of the cooperative binding of adenosine cyclic 3',5'-monophosphate and guanosine cyclic 3',5'-monophosphate to the adenosine cyclic 3',5'-monophosphate receptor protein from Escherichia coli. *Biochemistry*, 19:5124–5130, 1980. [p. 34, 111]

[121] M. Takahashi, B. Blazy, A. Baudras, and W. Hillen. Ligand-modulated binding of a gene regulatory protein to DNA. Quantitative analysis of cyclic-AMP induced binding of CRP from Escherichia coli to non-specific and specific DNA targets. *J. Mol Biol*, 207:783–796, 1989. [p. 34, 111]

[122] T. Heyduk and J. Lee. Escherichia coli cAMP receptor protein: evidence for three protein conformational states with different promoter binding affinities. *Biochemistry*, 28:6914–6924, 1989. [p. 34, 111]

[123] S. Leu, C. Baker, E. Lee, and J. Harman. Position 127 Amino Acid Substitutions Affect the Formation of CRP:cAMP:lacP Complexes but Not CRP:cAMP:RNA Polymerase Complexes at lacP. *Biochemistry*, 38:6222–6230, 1999. [p. 34]

[124] E. Heyduk, T. Heyduk, and J. Lee. Intersubunit Communications in Escherichia coli Cyclic AMP Receptor Protein: Studies of the Ligand Binding Domain. *Biochemistry*, 31:3682–3688, 1992. [p. 34]

[125] M. Allen and T. D.J. *Computer Simulation of Liquids.* Oxford Science Publications, 1987. [p. 39]

[126] J. Ponder and D. Case. Force fields for protein simulations. *Adv. Prot. Chem.*, 66:27–85, 2003. [p. 40]

[127] C. Oostenbrink, A. Villa, A. Mark, and W. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25:1656–1676, 2004. [p. 40]

[128] L. Monticelli and E. Salonen, editors. *Biomolecular Simulations: Methods and Protocols*. Springer, 2013. [p. 41, 42, 44]

[129] L. Verlet. Computer ”Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.*, 159:98–103, 1967. [p. 41]

[130] D. Case, T. Darden, T. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang, K. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvai, K. Wong, F. Paesani, J. Vanicek, J. Liu, S. B. X. Wu, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. Roe, D. Mathews, M. Seetin, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. Kollman. *AMBER 12*. University of California, San Francisco, 2012. [p. 42, 48, 49, 50, 117, 118, 120, 135, 143]

[131] M. R. Hestenes and E. Stiefel. Methods of Conjugate Gradients for Solving Linear Systems. *J. Res. Natl. Bureau Stand.*, 49:409–436, 1952. [p. 43]

[132] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *Siam J. Sci. Comp.*, 16:1190–1208, 1995. [p. 43, 54]

[133] P. H. Hünenberger. Thermostat Algorithms for Molecular Dynamics Simulations. In C. Holm and K. Kremer, editors, *Advanced Computer Simulation*, volume 173 of *Advances in Polymer Science*, pages 105–149. Springer Berlin Heidelberg, 2005. [p. 45, 46]

[134] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984. [p. 47]

[135] W. Jorgensen, J. Chandrasekhar, J. Madura, R. Impey, and M. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983. [p. 47]

[136] W. J. M.W. Mahone. A five-site model for liquid water and the reproduction of the density anomaly by rigid nonpolarizable potential functions. *J. Chem. Phys.*, 112:8910–8922, 2000. [p. 47]

[137] J. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comp. Phys.*, 23:327–341, 1977. [p. 48]

[138] B. Honig and A. Nicholls. Classical Electrostatics in Biology and Chemistry. *Science*, 268:pp. 1144–1149, 1995. [p. 48]

[139] M. K. Gilson, K. A. Sharp, and B. H. Honig. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comput. Chem.*, 9:327–335, 1988. [p. 48]

[140] J. Warwicker and H. Watson. Calculation of the electric potential in the active site cleft due to -helix dipoles. *J. Mol. Biol.*, 157:671 – 679, 1982. [p. 48]

[141] I. Klapper, R. Hagstrom, R. Fine, K. Sharp, and B. Honig. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *Proteins*, 1:47–59, 1986. [p. 48]

[142] A. Onufriev, D. A. Case, and D. Bashford. Effective Born radii in the generalized Born approximation: The importance of being perfect. *J. Comput. Chem.*, 23:1297–1304, 2002. [p. 49]

[143] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990. [p. 49]

[144] J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor. Chem. Acc.*, 101:426–434, 1999. [p. 49]

[145] B. Jayaram, D. Sprous, and D. L. Beveridge. Solvation Free Energy of Biomacromolecules: Parameters for a Modified Generalized Born Model Consistent with the AMBER Force Field. *J. Phys. Chem. B*, 102:9571–9576, 1998. [p. 49]

[146] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986. [p. 50]

[147] K. A. Sharp, A. Nicholls, R. F. Fine, and B. Honig. Reconciling the Magnitude of the Microscopic and Macroscopic Hydrophobic Effects. *Science*, 252:106–109, 1991. [p. 50]

[148] C. Tan, Y.-H. Tan, and R. Luo. Implicit Nonpolar Solvent Models. *J. Phys. Chem. B*, 111:12263–12274, 2007. [p. 50, 120, 164, 165]

[149] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.*, 5:350–358, 2009. [p. 50]

[150] L. R. Pratt and D. Chandler. Hydrophobic solvation of nonspherical solutes. *J. Chem. Phys.*, 73:3430–3433, 1980. [p. 50]

[151] D. M. Huang and D. Chandler. The Hydrophobic Effect and the Influence of Solute-Solvent Attractions. *J. Phys. Chem. B*, 106:2047–2053, 2002. [p. 50]

[152] H. Goldstein. *Classical Mechanics*. Addison Wesley, Reading, Massachusetts, USA, 1950. [p. 51]

[153] M. Levitt, C. Sander, and P. Stern. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *Int. J. Quantum Chem.*, 24:181–199, 1985. [p. 51]

[154] I. Bahar, T. Lezon, A. Bakan, and H. Shrivastava. Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chem. Rev.*, 110:1463–1497, 2010. [p. 51]

[155] *GROMACS Groningen Machine for Chemical Simulations: User Manual - Version 4.5*. [p. 54, 118]

[156] M. Tirion. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996. [p. 54, 56, 95]

[157] T. Haliloglu, I. Bahar, and B. Erman. Gaussian Dynamics of Folded Proteins. *Phys. Rev. Lett.*, 79:3090–3093, 1997. [p. 54]

[158] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, 2:173 – 181, 1997. [p. 54, 95]

[159] P. Doruker, A. R. Atilgan, and I. Bahar. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor. *Proteins: Struct., Funct., Bioinf.*, 40:512–524, 2000. [p. 54, 55]

[160] A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.*, 80:505–515, 2001. [p. 54, 55]

[161] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998. [p. 54]

[162] K. Hinsen, A. Thomas, and M. Field. Analysis of domain motions in large proteins. *Proteins*, 34:369–382, 1999. [p. 54]

[163] P. J. Flory, M. Gordon, and N. G. McCrum. Statistical Thermodynamics of Random Networks [and Discussion]. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, 351:351–380, 1976. [p. 54]

[164] L. Yang, X. Liu, C. Jursa, M. Holliman, A. Rader, H. Karimi, and I. Bahar. iGNM: a database of protein functional motions based on Gaussian Network Model. *Bioinformatics*, 21:2978–2987, 2005. [p. 55]

[165] E. Eyal, L. Yang, and I. Bahar. Anisotropic Network Model: Systematic Evaluation and a New Web Interface. *Bioinformatics*, 22:2619–2627, 2006. [p. 56]

[166] F. Tama, F. Gadea, O. Marques, and Y. Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, 41:1–7, 2000. [p. 57]

[167] M. Karplus and J. Kushick. Model for Estimating the Configurational Entropy of Macromolecules. *Macromolecules*, 14:325–332, 1981. [p. 57]

[168] A. Amadei, A. Linssen, and H. J. C. Berendsen. Essential Dynamics of Proteins. *Proteins: Struct., Funct., Bioinf.*, 17:412–425, 1993. [p. 57]

[169] M. Balsera, W. Wriggers, Y. Oono, and K. Schulten. Principal Component Analysis and Long Time Protein Dynamics. *J. Phys Chem.*, 100:2567–2572, 1996. [p. 57]

[170] E. Dykeman and O. Sankey. Normal Mode Analysis and Applications in Biological Physics. *J. Phys.: Comndens. Matter*, 22:423202–423227, 2010. [p. 57]

[171] J. Schlitter. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chemical Physics Letters*, 215:617 – 621, 1993. [p. 58, 67, 68]

[172] T. Rodgers, D. Burnell, P. Townsend, E. Pohl, M. Cann, M. Wilson, and T. McLeish. ΔΔ PT: a comprehensive toolbox for the analysis of protein motion. *BMC Bioinformatics*, 14:183, 2013. [p. 58, 95, 119]

[173] E. C. Dykeman and O. F. Sankey. Normal mode analysis and applications in biological physics. *Journal of Physics: Condensed Matter*, 22:423202, 2010. [p. 58]

[174] B. Hess. Convergence of sampling in protein simulations. *Phys. Rev. E*, 65:031910:1–10, 2002. [p. 64, 65]

[175] B. Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E*, 62:8438–8448, 2000. [p. 65, 155]

[176] H. Schfer, A. E. Mark, and W. F. van Gunsteren. Absolute entropies from molecular dynamics simulation trajectories. *The Journal of Chemical Physics*, 113:7809–7817, 2000. [p. 68]

[177] S. A. Harris, E. Gavathiotis, M. S. Searle, M. Orozco, and C. A. Laughton. Cooperativity in DrugDNA Recognition: A Molecular Dynamics Study. *J. Am. Chem. Soc.*, 123:12658–12663, 2001. [p. 68]

[178] S. A. Harris and C. A. Laughton. A simple physical description of DNA dynamics: quasi-harmonic analysis as a route to the configurational entropy. *J. Phys.: Condens. Matter*, 19:076103, 2007. [p. 68]

[179] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.*, 33:889–897, 2000. [p. 68]

[180] B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.*, 8:3314–3321, 2012. [p. 68, 69, 120, 143]

[181] B. Rupp. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology.* Garland Science, 2010. [p. 72, 73, 74, 75, 76, 77, 79, 80, 82, 87]

[182] C. A. Schall, E. Arnold, and J. M. Wiencek. Enthalpy of crystallization of hen egg-white lysozyme. *Journal of Crystal Growth*, 165:293 – 298, 1996. [p. 73]

[183] S.-T. Yau, D. N. Petsev, B. R. Thomas, and P. G. Vekilov. Molecular-level thermodynamic and kinetic parameters for the self-assembly of apoferritin molecules into crystals. *J. Mol. Biol.*, 303:667 – 678, 2000. [p. 73]

[184] D. N. Petsev, B. R. Thomas, S.-T. Yau, D. Tsekova, C. Nanev, W. W. Wilson, and P. G. Vekilov. Temperature-independent solubility and interactions between apoferritin monomers and dimers in solution. *J. Cryst. Growth*, 232:21–29, 2001. [p. 73]

[185] O. Gliko, N. Neumaier, W. Pan, I. Haase, M. Fischer, A. Bacher, S. Weinkauf, and P. G. Vekilov. A Metastable Prerequisite for the Growth of Lumazine Synthase Crystals. *J. Am. Chem. Soc.*, 127:3433–3438, 2005. [p. 73]

[186] Z. S. Derewenda and P. G. Vekilov. Entropy and surface engineering in protein crystallization. *Acta Crystallogr. D Biol. Crystallogr.*, 62:116–124, 2006. [p. 73]

[187] E. Pohl, R. K. Holmes, and W. G. J. Hol. Motion of the DNA-binding Domain with Respect to the Core of the Diphtheria Toxin Repressor (DtxR) Revealed in the Crystal Structures of Apo- and Holo-DtxR. *J. Biol. Chem.*, 273:22420–22427, 1998. [p. 74]

[188] A. McPherson. Crystallization of proteins from polyethylene glycol. *J. Biol. Chem.*, 251:6300–6303, 1976. [p. 76]

[189] R. G. F. Jr., A. L. Perryman, and C. T. Samudzi. Re-clustering the database for crystallization of macromolecules. *Journal of Crystal Growth*, 183:653 – 668, 1998. [p. 77]

[190] M. Tung and D. T. Gallagher. The Biomolecular Crystallization Database Version 4: expanded content and new features. *Acta Crystallogr. D Biol. Crystallogr.*, 65:18–23, 2009. [p. 77]

[191] A. McPherson and B. Cudney. Searching for silver bullets: An alternative strategy for crystallizing macromolecules. *J. Struct. Biol.*, 156:387 – 406, 2006. [p. 77]

[192] E. A. Stura, G. R. Nemerow, and I. A. Wilson. Strategies in the crystallisation of glycoproteins and protein complexes. *J. Cryst. Growth*, 122:273–285, 1992. [p. 77]

[193] J. Jancarik and S.-H. Kim. Sparse matrix sampling: a screening method for crystallization of proteins. *J. Appl. Cryst.*, 24:409–411, 1991. [p. 77]

[194] E. Prince, editor. *International Tables for Crystallography, Volume C, 3rd Edition, Mathematical, Physical and Chemical Tables.* Kluwer Academic Publishers, 2004. [p. 79]

[195] A. L. Patterson. A Fourier Series Method for the Determination of the Components of Interatomic Distances in Crystals. *Phys. Rev.*, 46:372–376, 1934. [p. 84]

[196] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read. *Phaser* crystallographic software. *J. Appl. Crystallogr.*, 40:658–674, 2007. [p. 85, 86]

[197] G. N. Murshudov, A. A. Vagin, and E. J. Dodson. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr. D Biol. Crystallogr.*, 53:240–255, 1997. [p. 85, 87, 188]

[198] R. J. Read. *International Tables for Crystallography F*, chapter Model phases: Probabilities, bias and maps, pages 325–331. Springer, 2001. [p. 86]

[199] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin. *REFMAC*5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.*, 67:355–367, 2011. [p. 87, 88]

[200] A. T. Brunger and M. Nilges. Computational Challenges for Macromolecular Structure Determination by X-Ray Crystallography and Solution NMR-Spectroscopy. *Q. Rev. Biophys.*, 26:49–125, 1993. [p. 88, 188]

[201] R. A. Engh and R. Huber. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr. A*, 47:392–400, 1991. [p. 88]

[202] I. J. Tickle. Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. *Acta Crystallogr. D Biol. Crystallogr.*, 63:1274–1281, 2007. [p. 88]

[203] E. A. Lewis and K. P. Murphy. *Protein Ligand Interactions: Methods and Applications*, chapter Isothermal Titration Calorimetry, pages 1–17. Humana Press, 2005. [p. 89, 91]

[204] J. J. Christensen, R. M. Izatt, L. D. Hansen, and J. A. Partridge. Entropy Titration. A Calorimetric Method for the Determination of G, H, and S from a Single Thermometric Titration. *J. Phys. Chem.*, 70:2003–2010, 1966. [p. 90]

[205] J. J. Christensen, D. P. Wrathall, J. O. Oscarson, and R. M. Izatt. Theoretical evaluation of entropy titration method for calorimetric determination of equilibrium constants in aqueous solution. *Anal. Chem.*, 40:1713–1717, 1968. [p. 90]

[206] J. J. Christensen, R. M. Izatt, and D. Eatough. Thermodynamics of Metal Cyanide Coordination. V. Log K, H, and S Values for the Hg2+-CN- System. *Inorg. Chem.*, 4:1278–1280, 1965. [p. 90]

[207] D. Eatough. Calorimetric determination of equilibrium constants for very stable metal-ligand complexes. *Anal. Chem.*, 42:635–639, 1970. [p. 90]

[208] N. V. Beaudette and N. Langerman. An improved method for obtaining thermal titration curves using micromolar quantities of protein. *Anal. Biochem.*, 90:693–704, 1978. [p. 90]

[209] T. Wiseman, S. Williston, J. F. Brandts, and L.-N. Lin. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Anal. Biochem.*, 179:131 – 137, 1989. [p. 90]

[210] *MicroCal$^{\text{TM}}$ iTC200 System: Getting Started Booklet.* [p. 90]

[211] S. Leavitt and E. Freire. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Curr. Opin. Struct. Biol.*, 11:560 – 566, 2001. [p. 90]

[212] A. Brown. Analysis of Cooperativity by Isothermal Titration Calorimetry. *Int. J. Mol. Sci.*, 10:3457–3477, 2009. [p. 91]

[213] *MicroCal iTC200 System Manual/Getting Started.* [p. 92]

[214] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual Molecular Dynamics. *J. Mol. Graph.*, 14:33–38, 1996. [p. 99, 127, 143]

[215] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65:712–725, 2006. [p. 117, 124]

[216] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.*, 78:1950–1958, 2010. [p. 117]

[217] D.-W. Li and R. Brschweiler. NMR-Based Protein Potentials. *Angew. Chem.*, 49:6778–6780, 2010. [p. 117]

[218] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge

force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem*, 24:1999–2012, 2003. [p. 117]

[219] P. Emsley and K. Cowtan. *Coot*: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.*, 60:2126–2132, 2004. [p. 117, 119, 188]

[220] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and Development of Coot. *Acta Crystallogr. D Biol. Crystallogr.*, 66:486–501, 2010. [p. 117]

[221] A. Jakalian, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem*, 23:16231641, 2002. [p. 117]

[222] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25:247 – 260, 2006. [p. 117]

[223] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.*, 98:11623–11627, 1994. [p. 117, 121, 122]

[224] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.*, 72:650–654, 1980. [p. 117, 121, 122]

[225] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E.

Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, , and D. J. Fox. *Gaussian 09, Revision A.02.* Gaussian, Inc., Wallingford CT, 2009. [p. 117, 121, 122]

[226] A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.*, 8:1542–1555, 2012. [p. 118]

[227] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.*, 9:3878–3888, 2013. [p. 118]

[228] D. R. Roe and T. E. Cheatham. PTRAJ and CPPTRAJ: Software for processing and analysis of Molecular Dynamics trajectory data. *J. Chem. Theory Comput.*, 9:3084–3094, 2013. [p. 119, 129]

[229] University of California, San Francisco. *AMBER Tools 12*, 2010. [p. 119]

[230] B. Mennucci, J. Tomasi, R. Cammi, J. R. Cheeseman, M. J. Frisch, F. J. Devlin, S. Gabriel, and P. J. Stephens. Polarizable Continuum Model (PCM) Calculations of Solvent Effects on Optical Rotations of Chiral Molecules. *J. Phys. Chem. A*, 106:6102–6113, 2002. [p. 122]

[231] L. Xu, H. Sun, Y. Li, J. Wang, and T. Hou. Assessing the Performance of MM/PBSA and MM/GBSA Methods. 3. The Impact of Force Fields and Ligand Charge Models. *J. Phys. Chem. B*, 117:8408–8421, 2013. [p. 124]

[232] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, 34:827–828, 1978. [p. 127, 140, 141, 142, 204, 206, 249]

[233] W. Wriggers and K. Schulten. Protein Domain Movements: Detection of Rigid Domains and Visualization of Hinges in Comparisons of Atomic Coordinates. *Proteins*, 29:1–14, 1997. [p. 127, 143]

[234] The PyMOL Molecular Graphics System, Version 1.3r1. Schrödinger LLC, 2010. [p. 139, 140, 141, 142, 204, 205]

[235] R. Vijayaraj, S. Van Damme, P. Bultinck, and V. Subramanian. Structure and stability of cyclic peptide based nanotubes: a molecular dynamics study of the influence of amino acid composition. *Phys. Chem. Chem. Phys.*, 14:15135–15144, 2012. [p. 164, 165]

[236] K. S. Lundberg, D. D. Shoemaker, M. W. Adams, J. M. Short, J. A. Sorge, and E. J. Mathur. High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. *Gene*, 108:1 – 6, 1991. [p. 173]

[237] P. Andr, A. Kim, K. Khrapko, and W. G. Thilly. Fidelity and Mutational Spectrum of Pfu DNA Polymerase on a Human Mitochondrial DNASequence. *Genome Res.*, 7:843–852, 1997. [p. 173]

[238] J. Cline, J. C. Braman, and H. H. Hogrefe. PCR Fidelity of Pfu DNA Polymerase and Other Thermostable DNA Polymerases. *Nucleic Acids Res.*, 24:3546–3551, 1996. [p. 173]

[239] J.-M. Flaman, T. Frebourg, V. Moreau, F. Charbonnier, C. Martin, C. Ishioka, S. H. Friend, and R. Iggo. A rapid PCR fidelity assay. *Nucleic Acids Res.*, 22:3259–3260, 1994. [p. 173]

[240] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. *Science*, 239:pp. 487–491, 1988. [p. 174]

[241] B. Weiss, A. Jacquemin-Sablon, T. R. Live, G. C. Fareed, and C. C. Richardson. Enzymatic Breakage and Joining of Deoxyribonucleic Acid. *J. Biol. Chem.*, 243:4543–4555, 1968. [p. 174]

[242] M. Newman, T. Strzelecka, L. F. Dorner, I. Schildkraut, and A. K. Aggarwal. Structure of restriction-endonuclease BamHI and its relationship to *ECo*RI. *Nature*, 368:660–664, 1994. [p. 174]

[243] S. Shuman. Novel approach to molecular cloning and polynucleotide synthesis using vaccinia DNA topoisomerase. *J. Biol. Chem.*, 269:32678–84, 1994. [p. 174]

[244] M. J. Casadaban and S. N. Cohen. Analysis of gene control signals by DNA fusion and cloning in *Escherichia Coli. J. Mol. Biol.*, 138:179 – 207, 1980. [p. 175]

[245] S. Busby, D. Kotlarz, and H. Buc. Deletion mutagenesis of the Escherichia coli galactose operon promoter region. *J Mol Biol.*, 167:259–74, 1983. [p. 175]

[246] J. Newman, D. Egan, T. S. Walter, R. Meged, I. Berry, M. Ben Jelloul, J. L. Sussman, D. I. Stuart, and A. Perrakis. Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+ strategy. *Acta Crystallogr. D Biol. Crystallogr.*, 61:1426–1431, 2005. [p. 176, 187, 265]

[247] Roche Applied Science. *PCR Applications Manual (3rd Edition)*, 2006. [p. 176]

[248] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. Enzymatic Amplification of -Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science*, 230:pp. 1350–1354, 1985. [p. 176]

[249] K. B. Mullis and F. A. Faloona. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.*, 155:335 – 350, 1987. [p. 176]

[250] S. N. Ho, H. D. Hunt, R. M. Horton, J. K. Pullen, and L. R. Pease. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, 77:51 – 59, 1989. [p. 178]

[251] S.-H. Lin and J. C. Lee. Communications between the High-Affinity Cyclic Nucleotide Binding Sites in E. coli Cyclic AMP Receptor Protein:Effect of Single Site Mutations. *Biochemistry*, 41:11857–11867, 2002. [p. 187]

[252] T.-Y. Teng. Mounting of crystals for macromolecular crystallography in a free-standing thin film. *J. Appl. Crystallogr.*, 23:387–391, 1990. [p. 187]

[253] A. G. W. Leslie and H. R. Powell. Processing diffraction data with mosflm. In R. J. Read and J. L. Sussman, editors, *Evolving Methods for Macromolecular Crystallography*, volume 245 of *NATO Science Series*, pages 41–51. Springer Netherlands, 2007. [p. 187]

[254] N. . Collaborative Computational Project. The *CCP*4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, 50:760–763, 1994. [p. 187, 188, 201]

[255] Z. Dauter. Efficient use of synchrotron radiation for macromolecular diffraction data collection. *Prog. Biophys. Mol. Biol.*, 89:153 – 172, 2005. [p. 202]

[256] E. Lyman, J. Pfaendtner, and G. A. Voth. Systematic Multiscale Parameterization of Heterogeneous Elastic Network Models of Proteins. *Biophys. J.*, 95:4183–4192, 2008. [p. 215]

[257] C. Globisch, V. Krishnamani, M. Deserno, and C. Peter. Optimization of an Elastic Network Augmented Coarse Grained Model to Study CCMV Capsid Deformation. *PLoS ONE*, 8:e60582, 2013. [p. 215]

Appendix A

Molecular Dynamics configuration files

Configuration files and a sample script for performing MD in AMBER

A.1 Energy Minimisation

A.1.1 Minimising Solvent

```
CAP_2cAMP: initial minimisation solvent + ions
 &cntrl
  imin   = 1,       maxcyc = 10000,
  ncyc   = 5000,    ntb    = 1,
  ntr    = 1,       cut    = 10.0,
  ntwx   = 100
 /
Hold the Protein fixed
500.0
RES 1 403
END
END
```

A.1.2 Minimising Solute

```
CAP_2cAMP: initial minimisation whole system
 &cntrl
  imin   = 1,       maxcyc = 50000,
  ncyc   = 25000,   ntb    = 1,
  ntr    = 0,       cut    = 10.0
 /
```

A.2 Equilibration

A.2.1 Temperature Equilibration

```
CAP_2cAMP: heat equilibration
 &cntrl
  imin=0,           irest=0,
  nstlim=100000,    dt=0.002,
  ntc=2,            ntf=2,
  cut=10.0,         ntb=1,
  ntpr=500,         ntwx=5000,
  ntt=3,            gamma_ln=1.0,
  ntx=1,            ig=-1,
  tempi=0.0,        temp0=300.0,
  ntr=1,            ioutfm=1
 /
Keep CAP fixed with weak restraints
2.5
RES 1 403
END
END
```

A.2.2 Pressure Equilibration

```
CAP_2cAMP: density equilibration
 &cntrl
  imin=0,           irest=1,
  nstlim=25000,     dt=0.002,
  ntc=2,            ntf=2,
  ntx=5,            taup=1.0,
  cut=8.0,          ntb=2,
  ntpr=500,         ntwx=500,
  ntt=3,            gamma_ln=2.0,
  temp0=300.0,      ig=-1,
  ntr=1,            ioutfm=1,
  ntp=1
 /
Keep CAP fixed with weak restraints
10.0
```

```
RES 1 403
END
END
```

A.3 Production MD

```
CAP_2cAMP: 4000ps of production MD
 &cntrl
  imin = 0,         irest = 1,
  ntb = 2,          pres0 = 1.0,
  taup = 2.0,       iwrap=1,
  cut = 10.0,       ntr = 0,
  ntc = 2,          ntf = 2,
  temp0 = 300.0,    ntx = 5,
  ntt = 3,          gamma_ln = 1.0,
  ntp = 1,          ig=-1,
  nstlim = 2000000, dt = 0.002,
  ntpr = 5000,      ntwx = 5000,
  ntwr = 10000,     ioutfm=1
 /
```

A.4 Sample Script for MD

```
module purge
module load dot
module load amber/cuda/SPDP/gcc/12.0

PROTEIN=cap
VAR=2CAMP

#Execute Commands
pmemd.cuda -O -i ../../wat_min1.in -o ${PROTEIN}_${VAR}_min1.out -p ${PROTEIN}_${VAR}.prmtop -c
${PROTEIN}_${VAR}.inpcrd -r ${PROTEIN}_${VAR}_min1.rst -ref ${PROTEIN}_${VAR}.inpcrd
#
pmemd.cuda -O -i ../../wat_min2.in -o ${PROTEIN}_${VAR}_min2.out -p ${PROTEIN}_${VAR}.prmtop -c
${PROTEIN}_${VAR}_min1.rst -r ${PROTEIN}_${VAR}_min2.rst
#
pmemd.cuda -O -i ../../heat_nc.in -o ${PROTEIN}_${VAR}_heat.out -p ${PROTEIN}_${VAR}.prmtop -c
${PROTEIN}_${VAR}_min2.rst -r ${PROTEIN}_${VAR}_heat.rst -ref ${PROTEIN}_${VAR}_min2.rst -x
${PROTEIN}_${VAR}_heat.nc
#
pmemd.cuda -O -i ../../density_nc.in -o ${PROTEIN}_${VAR}_density.out -p
${PROTEIN}_${VAR}.prmtop -c ${PROTEIN}_${VAR}_heat.rst -r ${PROTEIN}_${VAR}_density.rst -ref
${PROTEIN}_${VAR}_heat.rst -x ${PROTEIN}_${VAR}_density.nc
#
pmemd.cuda -O -i ../../md-prod_nc.in -o ${PROTEIN}_${VAR}_mdprod1.out -p
${PROTEIN}_${VAR}.prmtop -c ${PROTEIN}_${VAR}_density.rst -r ${PROTEIN}_${VAR}_mdprod1.rst -x
${PROTEIN}_${VAR}_mdprod1.nc
```

Appendix B

Normal Mode Analysis configuration files

Configuration files for performing NMA in GROMACS.

B.1 Energy Minimisation

```
; STANDARD MD INPUT OPTIONS NMA MINIMISATION
; for use with GROMACS
define                   = -DFLEXIBLE
constraints              = none
integrator               = l-bfgs
tinit                    = 0
nsteps                   = 100000
nbfgscorr                = 50
emtol                    = .0005
emstep                   = 0.1
gen_vel                  = yes
gen-temp                 = 300
nstcomm                  =  1
; NEIGHBORSEARCHING PARAMETERS
; nblist update frequency
nstlist                  = 0
; ns algorithm (simple or grid)
ns-type                  = simple
; Periodic boundary conditions:
pbc                      = no
; nblist cut-off
rlist                    = 0
domain-decomposition     = no
; OPTIONS FOR ELECTROSTATICS AND VDW
; Method for doing electrostatics
coulombtype              = Cut-Off
```

```
rcoulomb-switch          = 0
rcoulomb                 = 0
; Dielectric constant (DC) for cut-off or DC of reaction field
epsilon-r                = 1
; Method for doing Van der Waals
vdw-type                 = Cut-off
; cut-off lengths
rvdw-switch              = 0
rvdw                     = 0
```

B.2 Normal Mode Analysis

```
; STANDARD MD INPUT OPTIONS NMA
; for use with GROMACS
define                   = -DFLEXIBLE
constraints              = none
integrator               = nm
tinit                    = 0
nsteps                   = 100000
nbfgscorr                = 50
emtol                    = .0005
emstep                   = 0.1
gen_vel                  = yes
gen-temp                 = 300
nstcomm                  =  1
; NEIGHBORSEARCHING PARAMETERS
; nblist update frequency
nstlist                  = 0
; ns algorithm (simple or grid)
ns-type                  = simple
; Periodic boundary conditions: xyz (default), no (vacuum)
; or full (infinite systems only)
pbc                      = no
; nblist cut-off
rlist                    = 0
domain-decomposition     = no
; OPTIONS FOR ELECTROSTATICS AND VDW
; Method for doing electrostatics
coulombtype              = Cut-Off
rcoulomb-switch          = 0
rcoulomb                 = 0
; Dielectric constant (DC) for cut-off or DC of reaction field
epsilon-r                = 1
; Method for doing Van der Waals
vdw-type                 = Cut-off
; cut-off lengths
rvdw-switch              = 0
rvdw                     = 0
```

Appendix C

Python Programming

A number of functions written in Python by the author; used to perform least-square fitting of coordinates (using the Kabsch algorithm [232]) and overlap calculations are included below. The calculation performed in the program at the end was used for determining the overlaps between PDB difference vectors and NMA and PCA modes as seen in section 6.3.8.

```
#!/usr/bin/python
import numpy as np, math
import string, os

# Residue weights for mass weighting
ResWeights=dict(zip('ALA ARG ASN ASP CYS GLU GLN GLY HIE HIS ILE LEU LYS MET PHE \
PRO SER THR TRP TYR VAL'.split(),[71.078, 156.19, 114.10, 115.08, 103.14, 129.11,
 128.13, 57.051, 137.14, 137.14, 113.16, 113.16, 128.17, 131.20, 147.17, 97.115, 87.077,
 101.10,186.21,163.17,99.131]))

# # # # # # # # # # # # # # # # # # # # # # # # #
# Read options and arguments from command line  #
# # # # # # # # # # # # # # # # # # # # # # # # #

from optparse import OptionParser
parser=OptionParser()

usage = "\
This program calculates the vectors between a series of PDB structures.\n\
It then compares these vectors to the eigenvectors from a Principal Component\n\
Analysis (PCA) or Normal Mode Analysis (NMA) (NMWIZ format) by calculating the\n\
overlap between these vectors.\n\n\
usage: %prog [option1] arg1 [option2] arg2 ... <PDB1> <PDB2>"
parser = OptionParser(usage)

parser.add_option("-n", "--nmwiz", action="store", metavar="FILE",
                        type="string", dest="nmwiz", help="NMWIZ file "
                        "containing coordinates and eigenvectors "
                        "to fit the PDBs to")
parser.add_option("-a", "--alignmask", action="store", metavar="MASK",
                        type="string", dest="alignmask", help="Residue mask"
                        " to align structures. Format = :1-100 (for residues 1-100)")
parser.add_option("-e", "--vecmask",action="store", metavar="MASK",
                        type="string", dest="vecmask", help="Residue mask "
                        "to select the residues used in calculating"
                        "the overlap of the eigenvectors. Format = "
                        ":1-100 (for residues 1 to 100)")
parser.add_option("-o", "--outtab",action="store", metavar="FILE",
                        default=’overlaps.tex’, type="string", dest="outtab",
                        help="Output file for table (in latex format)")
parser.add_option("-p", "--nmwout",action="store", metavar="FILE",
                        type="string", dest="nmwout", help="Output file "
                        "for nmwiz created from pdbs (uses nmwiz input"
                        " coordinates as output coordinates)")
parser.add_option("-x", "--vecstart",action="store", metavar="INDEX",
                        default=0, type="int", dest="vecstart",
                        help="First eigenvector to read from nmwiz file")
parser.add_option("-y", "--vecstop",action="store", metavar="INDEX",
                        default=6, type="int", dest="vecstop",
                        help="Last eigenvector to read from nmwiz file")

(options, pdbs) = parser.parse_args()

# Read nmwiz file into a Nx3 array (for coordinated)
def readnmwiz(infile):
    print "Reading NMWIZ file: "+infile
    with open(infile) as f:
        fileLines=f.readlines()
    evecs=[]
    evecIndex=[]
    evecScale=[]
    for line in fileLines:
        if line.startswith(’resids’):
            resids=line.split()[1:]
        elif line.startswith(’resnames’):
            resnames=line.split()[1:]
        elif line.startswith(’bfactors’):
            bfactors=line.split()[1:]
        elif line.startswith(’coordinates’):
            coords1D=line.split()[1:]
            coords=np.array([[float(coords1D[i]) for i in range(0,len(coords1D),3)],
            [float(coords1D[i]) for i in range(1,len(coords1D),3)],
            [float(coords1D[i]) for i in range(2,len(coords1D),3)]])

        elif line.startswith(’mode’):
            lineData=line.split()
            evecIndex.append(lineData[1])
            evecScale.append(lineData[2])
            evecs.append([[float(lineData[i]) for i in range(3,len(lineData),3)],
            [float(lineData[i]) for i in range(4,len(lineData),3)],
            [float(lineData[i]) for i in range(5,len(lineData),3)]])

    evecs=np.array(evecs)
    return coords,evecs,evecIndex,evecScale,[int(i) for i in resids],resnames,bfactors

# Extract data from a PDB file
def readpdb(inf,CA=False,opt2=''):
    print "Reading PDB: "+inf
    with open(inf) as f:
        fileLines = f.readlines()

    atomtype=[]
    altloc=[]
    residno=[]
    coords=[]
    atomno=[]
    resname=[]
    chid=[]
    altres=False

    if CA:
        for line in fileLines:
            if line[0:6]=='ATOM  ':
                if line[12:16].rstrip(' ').lstrip(' ')=='CA':
                    if line[16]!=' ' and int(line[22:26])==residno[-1]:
                        altres=True
                        continue
                    atomno.append(int(line[6:11]))
                    atomtype.append(line[12:16])
                    altloc.append(line[16])
                    resname.append(line[17:20])
                    chid.append(line[21])
                    residno.append(int(line[22:26]))
                    coords.append([float(line[30:38]),float(line[38:46]),float(line[46:54])])

    else:
        for line in fileLines:
            if line[0:6]=='ATOM  ':
                if line[16]!=' ' and int(line[22:26])==residno[-1]:
                    altres=True
                    continue
                atomno.append(int(line[6:11]))
                atomtype.append(line[12:16])
                resname.append(line[17:20])
                chid.append(line[21])
                residno.append(int(line[22:26]))
                coords.append([float(line[30:38]),float(line[38:46]),float(line[46:54])])
                #~ Bfac.append(float(line[60:66]))
                #~ element.append(line[76:78].lstrip(' '))


    if altres:
        print 'Alternate configurations found in PDB! Using first seen!'
    return atomno,atomtype,resname,chid,residno,np.transpose(np.array(coords))

# Calculates the center of geometry of a coordinate set
def returnCOG(coords):
    COG=[]
    for dim in coords:
        COG.append(np.mean(dim))

    return np.array(COG)

#Calculates the center of mass of a coordinate set
def returnCOM(coords,masses):
    COM=[]
    for dim in coords:
        COM.append(1/(np.sum(masses))*np.sum(masses*dim))

    return np.array(COM)

#Centers a coordinate set to the origin
def centerOrigin(coords,masses=np.array([])):
    if not masses.any():
        center=returnCOG(coords)
    else:
        center=returnCOM(coords,masses)

    for dim in range(len(coords)):
        for x in range(len(coords[dim])):
            coords[dim][x]=coords[dim][x]-center[dim]
    return coords

#Centers a coordinate set to the origin (based on mask)
def centerOriginMask(coords,maskedcoords,maskedmasses=np.array([])):
    if not maskedmasses.any():
        center=returnCOG(maskedcoords)
    else:
        center=returnCOM(maskedcoords,maskedmasses)

    for dim in range(len(coords)):
        coords[dim,:]-=center[dim]
        maskedcoords[dim,:]-=center[dim]
    return coords,maskedcoords

# Calculates the (rotational) covariance matrix between matA and matB
def getCovariance(matA,matB):
    return matA.dot(np.transpose(matB))
```

```
#Checks sign of a value
def checkSign(value):
    if value<0:
        return -1
    elif value>0:
        return 1

# Corrects the rotation matrix if coords are not right-handed by
# creating a matrix to multiply single value decomposition vectors by
def correctRotMat(u,v):
    d=checkSign(np.linalg.det(np.transpose(v).dot(np.transpose(u))))
    return np.array([[1,0,0,],[0,1,0],[0,0,d]])

# Calculates the rotation matrix to rotate matB to align with matA
#(and S, which gives RMSD value if selected) using singular value decomposition
def calcRotMat(matA,matB,getS=False):
    cov=getCovariance(matA,matB)
    u,s,v=np.linalg.svd(cov)
    flip=correctRotMat(u,v)
    s = np.dot(s,flip)
    rotMat=(np.dot(u,flip)).dot(v)

    if not getS:
        return rotMat
    else:
        return rotMat, s

# Rotate a matrix by the rotation matrix: rotMat
def rotateMat(mat,rotMat):
    return rotMat.dot(mat)

# Overall routine to run the RMS fit (rotating matB)
def RMSfit(matA,matB,retRMS=False):
    E0 = np.sum( np.sum(matA * matA,axis=0),axis=0)\
    + np.sum( np.sum(matB * matB,axis=0),axis=0)

    rotmat,s=calcRotMat(matA,matB,True)
    RMSD = E0 -(2.0*sum(s))
    RMSD = np.sqrt(abs(RMSD/len(matB[0])))

    if not retRMS:
        return rotateMat(matB,rotmat)
    else:
        return rotateMat(matB,rotmat),RMSD

#Center and Rotate matrix B to matrix A according to mask
def centerRMSfitmask(matA,matB,mask=[[0,401]],RM=False,masses=np.array([])):
    matAmask=[[],[],[]]
    matBmask=[[],[],[]]
    for mi in mask:
        matAmask=[matAmask[c]+list(matA[c][mi[0]:mi[1]]) for c in range(len(matA))]
        matBmask=[matBmask[c]+list(matB[c][mi[0]:mi[1]]) for c in range(len(matB))]

    matAmask=centerOrigin(np.array(matAmask),masses)
    matBmask=centerOrigin(np.array(matBmask),masses)

    #Get initial residuals
    E0 = np.sum( np.sum(matAmask * matAmask,axis=0),axis=0)\
    + np.sum( np.sum(matBmask * matBmask,axis=0),axis=0)

    #Return rotation matrix and s to calculate errors
    rotmat,s=calcRotMat(matAmask,matBmask,getS=True)

    RMSD = E0 -(2.0*sum(s))
    RMSD = np.sqrt(abs(RMSD/len(matBmask[0])))

    if not RM:
        return rotateMat(centerOrigin(matB),rotmat),RMSD
    else:
        return rotateMat(centerOrigin(matB),rotmat),rotmat, RMSD

# Calculates the overlap of two vectors given as np arrays (order doesn't matter)
def Overlap(Vec1,Vec2):
    return abs(np.dot(Vec1,Vec2))/math.sqrt(Vec1.dot(Vec1)*Vec2.dot(Vec2))

# Read the residue mask as given on the command line
def readResMask(mask):
    return [int(i) for i in mask.lstrip(":").split("-")]

# Determine the starting and stop indices for the coordinates based on the residue mask
def translateMask(mask,resids):
    maskbounds=readResMask(mask)
    start = [i for i, x in enumerate(resids) if x == maskbounds[0]]
    stop = [i for i, x in enumerate(resids) if x == maskbounds[1]]
    if len(start)!=len(stop):
        print "Invalid Alignment Mask failed for: "+pdbs[i]
    return start,stop

# Make the masked coordinates
def makeMaskedCoords(coords,start,stop):
    rco=coords[:,start[0]:stop[0]]
    for i in range(1,len(start)):
        rco=np.append(rco,coords[:,start[i]:stop[i]],axis=1)
    return rco

# Make the masked coordinates in 1D
def makeMasked1d(data,start,stop):
    maskeddata=data[start[0]:stop[0]]
    for i in range(1,len(start)):
        maskeddata=np.append(maskeddata,data[start[i]:stop[i]])
    return maskeddata

# Read all the PDBs and return useful information
# (residue numbers, residue names and coordinates)
def readAllPDBs(pdbs):
    pdbresids,pdbresnames,pdbcoords=[],[],[]
    for pdb in pdbs:
        atomno,atomtype,resname,chid,residno,coords= readpdb(pdb,CA=True)
        pdbresids.append(residno)
        pdbresnames.append(resname)
        pdbcoords.append(centerOrigin(coords))
        checkPDBres(residno,chid)
    return pdbresids,pdbresnames,pdbcoords

# Check that there are no gaps in the PDB numbering
def checkPDBres(resids,chid):
    ch='-'
    for i in range(len(resids)):
        if ch!=chid[i]:
            ch=chid[i]
            rid=resids[i]
        if rid!=resids[i]:
            rid=resids[i]
            print 'Check resid: '+str(resids[i])+'! It may be missing, so program may crash!'
        rid+=1

# Make difference vectors between all PDB pairs
def makePDBmodes(vMPDBs,vMcoords):
    vectors=[]
    runcoords=[ rotateMat(x,calcRotMat(x,vMcoords)) for x in vMPDBs ]
    for i in range(len(runcoords)):
        for j in range(i+1,len(runcoords)):
            vectors.append(runcoords[i]-runcoords[j])
    return vectors

# Make a matrix of overlaps for all PDB difference vectors against the nmwiz vectors
def makeOverlapsMat(vMPDBs,nmwMaskedVecs,nmwMaskedCoords,retnmwdata=False):
    Overlaps=[]
    for i in range(len(vMPDBs)):
        Overlaps.append([])
        rm=calcRotMat(vMPDBs[i],nmwMaskedCoords)
        nmwMV=[rotateMat(Vec,rm) for Vec in nmwMaskedVecs]
        for j in range(i+1,len(vMPDBs)):
            Overlaps[i].append([])
            rm=calcRotMat(vMPDBs[i],vMPDBs[j])
            runcoords=rotateMat(vMPDBs[j],rm)
            for k in range(options.vecstart-1,options.vecstop):
                Overlaps[i][j-i-1].append(Overlap(vMPDBs[i].flatten()-runcoords.flatten(),
                                                                      nmwMV[k].flatten()))
    if retnmwdata:
        return Overlaps,nmwMV
    else:
        return Overlaps

# Write the latex overlaps table
def writeOverlapTable(outname,Overlaps,full=False):
    with open(outname,'w') as of:
        of.write("\\documentclass{article}\n"
                 "\\usepackage[table]{xcolor}\n"
                 "\\usepackage{longtable}\n"
                 "\\usepackage{underscore}\n"
                 "\\begin{document}\n"
                 "\\begin{longtable}{|l|l|"+len(Overlaps[0][0])*"l|"+"}\n"
                 "\\hline\n"
                 "PDB1 & PDB2 & "
                 +" & ".join([str(x) for x in range(1,len(Overlaps[0][0])+1)])
                 +"\\\\ \n"
                 "\\hline\n")
        loop1=range(len(Overlaps))

        for i in loop1:
            loop2=range(len(Overlaps[i]))

            for j in loop2:
                print Overlaps[i][j]
                if Overlaps[i][j][0]=='-':
                    continue
                if full:
                    pdb2=j
                else:
                    pdb2=i+j+1
                of.write(string.ascii_uppercase[i]+' & '+string.ascii_uppercase[pdb2])
                for k in range(len(Overlaps[i][j])):
                    of.write(' & ')
                    if Overlaps[i][j][k]>=0.5:
                        of.write('\\cellcolor{blue!25} ')
                    of.write("%.3f" % Overlaps[i][j][k])
                of.write('\\\\')
        of.write("\\hline\n"
                 "\\caption{PDBs used are:"+
                 ", ".join([string.ascii_uppercase[x]+": "+pdbs[x] for x in range(len(pdbs))])
                 +"}\n"
                 "\\end{longtable}\n"
                 "\\end{document}")
```

```
# Write an nmwiz file (with minimum necessary information)
def Qwritenmwiz(outfilename,resids,resnames,coords,vectors):
    string='nmwiz_load '+outfilename+'\n'
    string+=makenmwizline1d([ 'CA' for x in range(len(resnames))],'atomnames')
    string+=makenmwizline1d(resnames,'resnames')
    string+=makenmwizline1d(resids,'resids')
    string+=makenmwizlinecoord(coords)
    for i in range(len(vectors)):
        string+=Qmakenmwizlinemode(vectors[i],str(i+1))
    with open(outfilename,'w') as f:
        f.write(string)

# Write a line of nmwiz data
def makenmwizline1d(data,title):
    string=title+' '
    string+=' '.join([str(x) for x in list(data)])
    string+='\n'
    return string

# Write the coordinate line for nmwiz file
def makenmwizlinecoord(data):
    string='coordinates '
    string+=' '.join(["%.4f" % i for i in data.flatten('F')])
    string+='\n'
    return string

# Write a mode line for nmwiz file
def Qmakenmwizlinemode(data,modeno):
    string='mode '+modeno+' '
    string+=' '.join(["%.4f" % i  for i in data.flatten('F')])
    string+='\n'
    return string

# # # # # # # # # # # # # # # # # # # # # # #
# Begin Program script                      #
# # # # # # # # # # # # # # # # # # # # # # #

# Readnmwiz
(nmwcoords,nmwevecs,nmwevecIndex,nmwevecScale,
nmwresids,nmwresnames,nmwbfactors)=readnmwiz(options.nmwiz)
masses=np.array([ResWeights[item] for item in nmwresnames])

#read PDBs
pdbresids,pdbresnames,pdbcoords=readAllPDBs(pdbs)

# Centers nma/pca coordinates (nmwcoords) and all the PDB coordinates to the origin (masked)
# Then it Least square fits the PDB coordinates to the nmwcoords (masked)
# The masked difference vectors between the PDBs is determined (vMPDBs)
if options.alignmask and options.vecmask:
    start,stop=translateMask(options.alignmask,nmwresids)
    nmwMaskedCoords=makeMaskedCoords(nmwcoords,start,stop)
    vstart,vstop=translateMask(options.vecmask,nmwresids)
    # Get evecs from nmwiz according to mask
    nmwMaskedVecs=[makeMaskedCoords(x,vstart,vstop) for x in nmwevecs]
    maskedMasses=makeMasked1d(masses,start,stop)
    # Center nmwiz coords and align masked coords
    nmwcoords,nmwMaskedCoords\
    =centerOriginMask(nmwcoords,nmwMaskedCoords,maskedMasses)

    nmwVMaskCoords=makeMaskedCoords(nmwcoords,vstart,vstop)
    vMPDBs=[]
    for i in range(len(pdbcoords)):
        start,stop=translateMask(options.alignmask,pdbresids[i])
        runcoords=makeMaskedCoords(pdbcoords[i],start,stop)
        maskedMasses=makeMasked1d(masses,start,stop)
        pdbcoords[i],runcoords==centerOriginMask(pdbcoords[i],runcoords,maskedMasses)
        rotmat,s=calcRotMat(nmwMaskedCoords,runcoords,getS=True)
        pdbcoords[i]=rotateMat(pdbcoords[i],rotmat)

        vstart,vstop=translateMask(options.vecmask,pdbresids[i])
        vMPDBs.append(makeMaskedCoords(pdbcoords[i],vstart,vstop))

    maskedResids=makeMasked1d(pdbresids[-1],vstart,vstop)
    maskedResnames=makeMasked1d(pdbresnames[-1],vstart,vstop)
else:
    print "At the moment align and vector masks needs to be assigned! EXITING"
    quit()

# Get the overlaps between PDB difference vectors and nmwiz coordinates
Overlaps, NMWMV=makeOverlapsMat(vMPDBs,nmwMaskedVecs,nmwVMaskCoords,True)
writeOverlapTable(options.outtab,Overlaps,full=False)

# Compile latex table to make PDF
os.system(’pdflatex ’+options.outtab)
if options.nmwout:
    Qwritenmwiz(options.nmwout,maskedResids,maskedResnames,vMPDBs[0],
                makePDBmodes(vMPDBs,nmwVMaskCoords))
    Qwritenmwiz(options.nmwiz.rstrip(’.nmd’)+’_c.nmd’,maskedResids,
                maskedResnames,vMPDBs[0],NMWMV)
```

Appendix D

Plasmid Maps

D.1 pCR$^{\text{TM}}$-Blunt II-TOPO$^{®}$ map

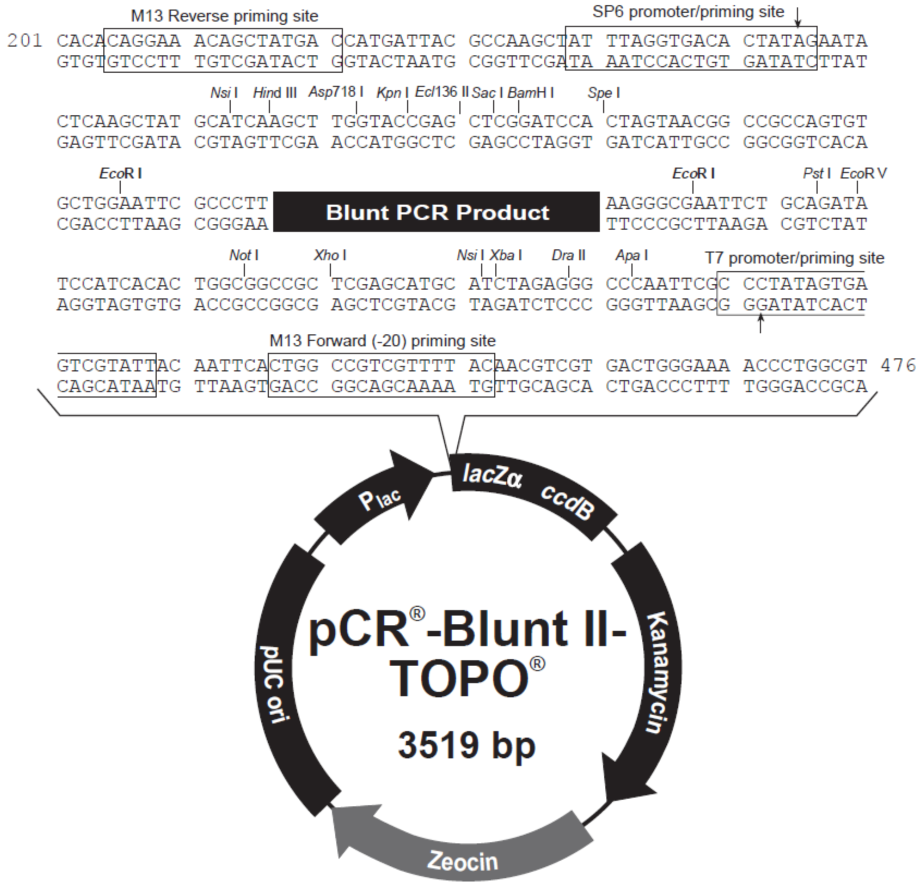

M13 Reverse priming site
SP6 promoter/priming site
201 CACACAGGAA ACAGCTATGA CCATGATTAC GCCAAGCTAT TTAGGTGACA CTATAGAATA
GTGTGTCCTT TGTCGATACT GGTACTAATG CGGTTCGATA AATCCACTGT GATATCTTAT
Nsi I
Hind III
Asp718 I
Kpn I
Ecl136 II
Sac I
BamH I
Spe I
CTCAAGCTAT GCATCAAGCT TGGTACCGAG CTCGGATCCA CTAGTAACGG CCGCCAGTGT
GAGTTCGATA CGTAGTTCGA ACCATGGCTC GAGCCTAGGT GATCATTGCC GGCGGTCACA
EcoR I
EcoR I
Pst I
EcoR V
GCTGGAATTC GCCCTT
CGACCTTAAG CGGGAA
Blunt PCR Product
AAGGGCGAATTCT GCAGATA
TTCCCGCTTAAGA CGTCTAT
Not I
Xho I
Nsi I
Xba I
Dra II
Apa I
T7 promoter/priming site
TCCATCACAC TGGCGGCCGC TCGAGCATGC ATCTAGAGGG CCCAATTCGC CCTATAGTGA
AGGTAGTGTG ACCGCCGGCG AGCTCGTACG TAGATCTCCC GGGTTAAGCG GGATATCACT
M13 Forward (-20) priming site
GTCGTATTAC AATTCACTGG CCGTCGTTTT ACAACGTCGT GACTGGGAAA ACCCTGGCGT 476
CAGCATAATG TTAAGTGACC GGCAGCAAAA TGTTGCAGCA CTGACCCTTT TGGGACCGCA
P lac
lacZα
ccdB
Kanamycin
Zeocin
pUC ori
pCR®-Blunt II-TOPO®
3519 bp

Figure D.1: The map for plasmid pCR$^{\text{TM}}$-Blunt II-TOPO$^{®}$

D.2 pREP4 map

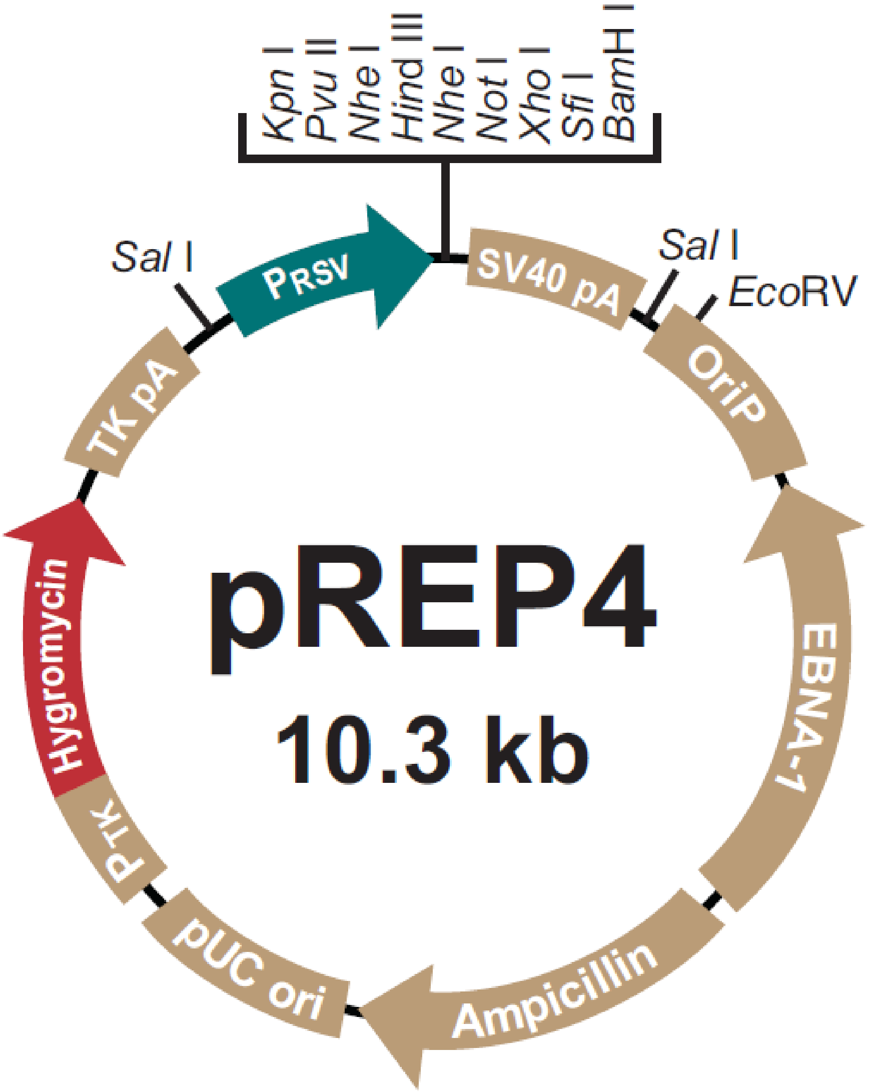


Figure D.2: The map for plasmid pREP4

D.3 pQE30 map

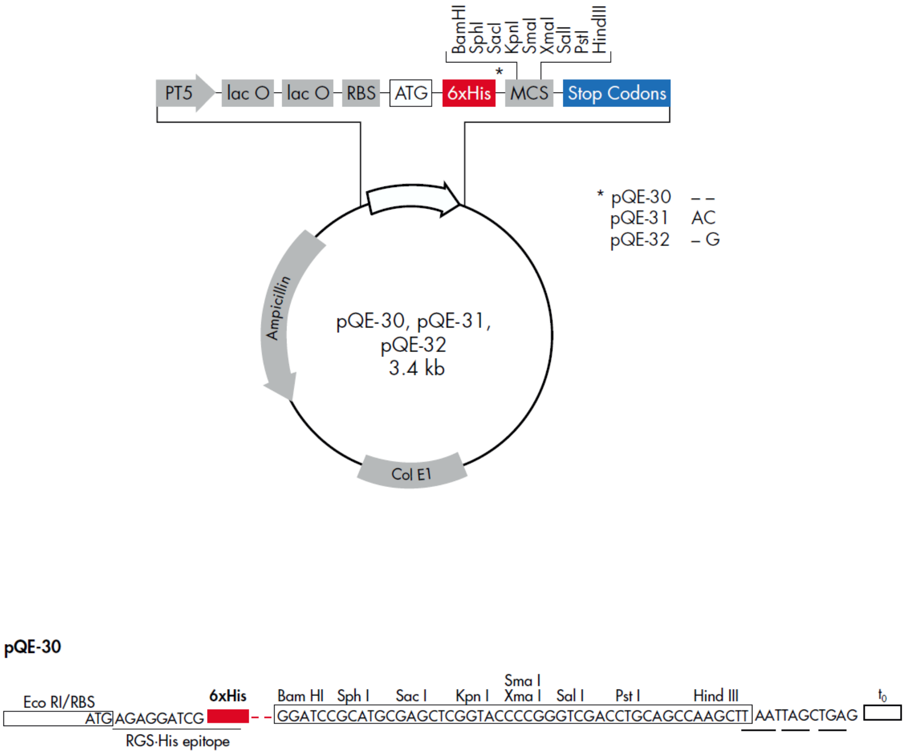


Figure D.3: The map for the plasmid pQE30, the insert was ligated between the BamHI and HindIII restriction sites.

Appendix E

PACT Premier$^{\text{TM}}$ HT-96

Tube No.	Buffer/Salt	Buffer/Salt	pH	Precipitant
A1	0.1 M SPG buffer	None	4.0	25 % w/v PEG 1500
A2	0.1 M SPG buffer	None	5.0	25 % w/v PEG 1500
A3	0.1 M SPG buffer	None	6.0	25 % w/v PEG 1500
A4	0.1 M SPG buffer	None	7.0	25 % w/v PEG 1500
A5	0.1 M SPG buffer	None	8.0	25 % w/v PEG 1500
A6	0.1 M SPG buffer	None	9.0	25 % w/v PEG 1500
A7	0.2 M sodium chloride	0.1 M sodium acetate	5.0	20 % w/v PEG 6000
A8	0.2 M ammonium chloride	0.1 M sodium acetate	5.0	20 % w/v PEG 6000
A9	0.2 M lithium chloride	0.1 M sodium acetate	5.0	20 % w/v PEG 6000
A10	0.2 M magnesium chloride	0.1 M sodium acetate	5.0	20 % w/v PEG 6000
A11	0.2 M calcium chloride	0.1 M sodium acetate	5.0	20 % w/v PEG 6000
A12	0.01 M zinc chloride	0.1 M sodium acetate	5.0	20 % w/v PEG 6000
B1	0.1 M MIB buffer	None	4.0	25 % w/v PEG 1500
B2	0.1 M MIB buffer	None	5.0	25 % w/v PEG 1500
B3	0.1 M MIB buffer	None	6.0	25 % w/v PEG 1500
B4	0.1 M MIB buffer	None	7.0	25 % w/v PEG 1500
B5	0.1 M MIB buffer	None	8.0	25 % w/v PEG 1500
B6	0.1 M MIB buffer	None	9.0	25 % w/v PEG 1500
B7	0.2 M sodium chloride	0.1 M MES	6.0	20 % w/v PEG 6000
B8	0.2 M ammonium chloride	0.1 M MES	6.0	20 % w/v PEG 6000
B9	0.2 M lithium chloride	0.1 M MES	6.0	20 % w/v PEG 6000
B10	0.2 M magnesium chloride	0.1 M MES	6.0	20 % w/v PEG 6000
B11	0.2 M calcium chloride	0.1 M MES	6.0	20 % w/v PEG 6000
B12	0.01 M zinc chloride	0.1 M MES	6.0	20 % w/v PEG 6000
C1	0.1 M PCTP buffer	None	4.0	25 % w/v PEG 1500
C2	0.1 M PCTP buffer	None	5.0	25 % w/v PEG 1500
C3	0.1 M PCTP buffer	None	6.0	25 % w/v PEG 1500
C4	0.1 M PCTP buffer	None	7.0	25 % w/v PEG 1500
C5	0.1 M PCTP buffer	None	8.0	25 % w/v PEG 1500
C6	0.1 M PCTP buffer	None	9.0	25 % w/v PEG 1500
C7	0.2 M sodium chloride	0.1 M HEPES	7.0	20 % w/v PEG 6000
C8	0.2 M ammonium chloride	0.1 M HEPES	7.0	20 % w/v PEG 6000
C9	0.2 M lithium chloride	0.1 M HEPES	7.0	20 % w/v PEG 6000
C10	0.2 M magnesium chloride	0.1 M HEPES	7.0	20 % w/v PEG 6000
C11	0.2 M calcium chloride	0.1 M HEPES	7.0	20 % w/v PEG 6000
C12	0.01 M zinc chloride	0.1 M HEPES	7.0	20 % w/v PEG 6000
D1	0.1 M MMT buffer	None	4.0	25 % w/v PEG 1500
D2	0.1 M MMT buffer	None	5.0	25 % w/v PEG 1500
D3	0.1 M MMT buffer	None	6.0	25 % w/v PEG 1500
D4	0.1 M MMT buffer	None	7.0	25 % w/v PEG 1500
D5	0.1 M MMT buffer	None	8.0	25 % w/v PEG 1500
D6	0.1 M MMT buffer	None	9.0	25 % w/v PEG 1500
D7	0.2 M sodium chloride	0.1 M Tris	8.0	20 % w/v PEG 6000
D8	0.2 M ammonium chloride	0.1 M Tris	8.0	20 % w/v PEG 6000
D9	0.2 M lithium chloride	0.1 M Tris	8.0	20 % w/v PEG 6000
D10	0.2 M magnesium chloride	0.1 M Tris	8.0	20 % w/v PEG 6000
D11	0.2 M calcium chloride	0.1 M Tris	8.0	20 % w/v PEG 6000
D12	0.002 M zinc chloride	0.1 M Tris	8.0	20 % w/v PEG 6000

Tube No.	Buffer/Salt	Buffer/Salt	pH	Precipitant
E1	0.2 M sodium fluoride	None		20 % w/v PEG 3350
E2	0.2 M sodium bromide	None		20 % w/v PEG 3350
E3	0.2 M sodium iodide	None		20 % w/v PEG 3350
E4	0.2 M potassium thiocyanate	None		20 % w/v PEG 3350
E5	0.2 M sodium nitrate	None		20 % w/v PEG 3350
E6	0.2 M sodium formate	None		20 % w/v PEG 3350
E7	0.2 M sodium acetate	None		20 % w/v PEG 3350
E8	0.2 M sodium sulfate	None		20 % w/v PEG 3350
E9	0.2 M potassium/sodium tartrate	None		20 % w/v PEG 3350
E10	0.02 M sodium/potassium phosphate	None		20 % w/v PEG 3350
E11	0.2 M sodium citrate	None		20 % w/v PEG 3350
E12	0.2 M sodium malonate	None		20 % w/v PEG 3350
F1	0.2 M sodium fluoride	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F2	0.2 M sodium bromide	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F3	0.2 M sodium iodide	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F4	0.2 M potassium thiocyanate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F5	0.2 M sodium nitrate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F6	0.2 M sodium formate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F7	0.2 M sodium acetate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F8	0.2 M sodium sulfate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F9	0.2 M potassium/sodium tartrate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F10	0.02 M sodium/potassium phosphate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F11	0.2 M sodium citrate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
F12	0.2 M sodium malonate	0.1 M Bis Tris propane	6.5	20 % w/v PEG 3350
G1	0.2 M sodium fluoride	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G2	0.2 M sodium bromide	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G3	0.2 M sodium iodide	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G4	0.2 M potassium thiocyanate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G5	0.2 M sodium nitrate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G6	0.2 M sodium formate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G7	0.2 M sodium acetate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G8	0.2 M sodium sulfate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G9	0.2 M potassium/sodium tartrate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G10	0.02 M sodium/potassium phosphate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G11	0.2 M sodium citrate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
G12	0.2 M sodium malonate	0.1 M Bis Tris propane	7.5	20 % w/v PEG 3350
H1	0.2 M sodium fluoride	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H2	0.2 M sodium bromide	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H3	0.2 M sodium iodide	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H4	0.2 M potassium thiocyanate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H5	0.2 M sodium nitrate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H6	0.2 M sodium formate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H7	0.2 M sodium acetate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H8	0.2 M sodium sulfate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H9	0.2 M potassium/sodium tartrate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H10	0.02 M sodium/potassium phosphate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H11	0.2 M sodium citrate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350
H12	0.2 M sodium malonate	0.1 M Bis Tris propane	8.5	20 % w/v PEG 3350

Table E.1: Crystallisation conditions in the PACT premier^{TM} HT-96 screen.
Abbreviations: **HEPES;** N-(2-hydroxyethyl)-piperazine-N'-2-ethanesulfonic acid, **MES;** 2-(N-morpholino)ethanesulfonic acid, **PEG;** Polyethylene glycol, **Tris;** 2-Amino-2-(hydroxymethyl)propane-1,3-diol, **SPG buffer;** Succinic Acid, Phosphate, Glycine, **MIB buffer;** Malonic acid, Imidazole, Boric acid, **PCTP buffer;** Propionic acid, Cacodylate, Bis-tris propane, **MMT buffer;** Malic acid, MES, Tris. [246]