On the Effect of Binding on Ubiquitin Dynamics

Dissertation

for the award of the degree

Doctor rerum naturalium

of the Georg-August-Universität Göttingen

within the doctoral program Physics of Biological and Complex System (PBCS)

of the Georg-August University School of Science (GAUSS)

submitted by

Jan Henning Peters

from Lippstadt

Göttingen 2013

Thesis committee:

Prof. Dr. Bert de Groot (supervisor, reviewer)
Computational and Biomolecular Dynamics Group,
Max Planck Institute for Biophysical Chemisty

Prof. Dr. Christian Griesinger (reviewer)
Department of NMR-based Structural Biology,
Max Planck Institute for Biophysical Chemisty

Prof. Dr. Jörg Enderlein
Drittes Physikalisches Institut
Georg-August-Universität Göttingen

Vorveröeffentlichungen der Dissertation

Teilergebnisse dieser Arbeit wurden im folgenden Beitrag veröffentlicht:

Publikationen:

Peters, J. H. and de Groot, B. L. Ubiquitin Dynamics in Complexes Reveal Molecular Recognition Mechanisms Beyond Induced Fit and Conformational Selection. PLoS Computational Biology 8, e1002704 (2012).

Abstract

Protein-protein interactions play an important role in all biological processes. But the principles underlying these interactions are not yet fully understood. Ubiquitin is a small signalling protein that non-covalently interacts with and is recognised by a multitude of other proteins. I have conducted molecular dynamics simulations of ubiquitin in complex with 12 different binding partners and compared ensembles of bound and unbound ubiquitin to determine the influence of complex formation on the dynamic properties of this protein. Along the main mode of fluctuation of ubiquitin, binding in most cases reduces the conformational space available to ubiquitin to a characteristic subspace of that covered by unbound ubiquitin. This behaviour can be well explained using the model of conformational selection. For lower amplitude collective modes, a spectrum of zero to almost complete coverage of bound by unbound ensembles was observed. The significant differences between bound and unbound structures are exclusively situated at the binding interface. Overall, the findings correspond neither to a complete conformational selection nor induced fit scenario. Instead, I introduce a model of conformational restriction, extension and shift, which describes the full range of observed effects. The observation of characteristic restrictions of the main mode dynamics in complexes, lead me to develop the hypothesis that ubiquitin binding can be modulated by changing main mode behaviour, for example by mutation of residues in the hydrophobic core. Using a screening protocol based on non-equilibrium free energy simulations,eleven mutations of ubiquitin were identified that shift the equilibrium population of unbound ubiquitin along the main mode. I calculated the effect of these mutations on the free energy of binding to different binding partners that require ubiquitin to be a characteristic state, observing a significant shift in binding affinity.

Acknowledgements

First of all I would like to thank my Supervisor Bert de Groot, who has given me the opportunity to start my work and who guided and supported me patiently over the last four years to finish it. I would also like to the members of my thesis committee, Christian Griesinger and Jörg Enderlein who have guided in this project.

I am grateful to my colleagues, past and present, in the Theoretical and Computational Biophysics Department. I do believe that there are few places in the world where there is such a concentration of knowledge and passion about Molecular Dynamics simulations. I am grateful to Helmut Grubmüller for assembling this department and for tirelessly sharing his enthusiasm about the field. It was one of his talks at the university of Leipzig that first called my attention to the department.

I owe thanks to Ansgar Esztermann and Martin Fechner for their excellent efforts to keep the machines running that made my work possible. Special thanks goes to Eveline Heinemann who helped whenever I struggled with bureaucracy. She keeps the department running and orchestrates the construction of the most impressive Phd-heads of Göttingen.

Daniel Seeliger, Monika Sharma and especially Servaas Michielssens have helped greatly in their collaboration and their contributions towards the development of the free energy screening protocol and its application to unbound ubiquitin mutants. Without them, the research described in chapter 4 would still be in its beginnnings.

This work has almost exclusively been created using open source software. The simulations were performed and partly analysed using GROMACS, further analysis was scripted in python, the results plotted in gnuplot and

documented using LaTeX. The literature cited in this work has been organised using the open source system Zotero which is a plugin to Firefox. Finally, all these programs run on GNU/Linux. Each of these projects has been developed and supported by a community of volunteers that make it possible for me and so many others not only use but if need be fully access and at least to some degree understand high quality scientific software, and I am grateful for their efforts that have supplied me with such powerful tools.

Contents

CONTENTS

1

Introduction

1.1 Protein-protein interactions

The function of cells and in extension all biological systems is governed by the interactome (1), the network of interactions between its molecular constituents, including an a huge number of protein-protein interactions (2). This thesis will focus on non-covalent interactions, that is interactions that are not based on chemical bonds. These interactions can be permanent, as in the case of molecular assemblies like the haemoglobin or aquaporin tetramers or the ribosome or transient as in the antibody-antigen recognition (3) signalling pathways (4) or intracellular transport (5). Complex interaction networks have been identified using experimental methods like two-hybrid screening (6), but the reliability of these high-throughput methods seems to be limited (7, 8) and verification of their results is necessary. The ability of proteins to form non-covalent complexes is described by their *binding affinity* (described below). A multitude of experimental methods has been developed to measure the binding affinity. Often, they are based on indirect observations of the binding process, as in isothermal titration calorimetry that measures the heat taken up or released upon binding. Other approaches are based on differences of optical properties between bound and unbound proteins or observe fluorescence. More extensive treatment of the different methods including their specific strengths and limitations can be found in the literature (9, 10, 11). The *binding affinity* indicates if and how strong proteins interact, but not how. For this, structural information is needed. X-ray structures of protein complexes provide a wealth of high resolution structural information but reflect a static snapshot of the structure, leaving the mechanism of complex formation and dynamics in the complex unaddressed.

Nuclear magnetic resonance (NMR) spectroscopy provides some information of dynamics, but until recently, only very fast molecular motions in the picosecond to nanosecond range (using NMR relaxation measurements) and slow dynamics (slower than 10ms, observed by peak splitting) were observable, while observations in the biologically relevant microsecond timescale require new approaches (12). The experimental determination of complex structures faces specific problems in addition to those encountered in unbound structure determination, like managing to co-crystallise bound pro-

teins and potential low concentrations of complexes with low binding affinity. Compared to the growing number of experimentally determined structures of unbound proteins, there is only a small number of known structures of protein complexes.

Computational methods may be used to verify protein-protein interactions and find complex structures based on the knowledge of unbound protein structures, but these predictions both of binding affinities (13) and complex structures (14, 15) remain challenging and highly non-trivial tasks. Computational methods have been shown to produce high quality complex structures of proteins known to interact in a number of cases. However, they are not well-suited to distinguish binding from non-binding protein pairs or accurately predict binding affinities, as has been shown in cross-docking studies, where docking algorithms have been applied to pairs of proteins known not to interact as well as pairs known to interact (16, 17, 18, 19). Most of the methods consider the binding partners to be rigid bodies, so predictions get highly inaccurate if significant conformational changes are part of the binding process. Also, entropic contributions to binding free energy can only be estimated using heuristic models. With the increase in computational power, flexibility has been introduced in the computational methods, and shows promising results (20). And a better understanding of the role of dynamics in protein is likely to bring further advances in this field.

Understanding protein-protein interaction could allow many new approaches for drug development, be it the disruption of unwanted interaction or the use of proteins as drugs themselves. An extensive, recent overview about protein-protein interactions can be found in (21).

In the case of transient interactions, the strength of binding is an important factor. The interactions between the proteins have to be strong enough to enable interactions but weak enough to allow dissociation if necessary.

The strength of protein binding is characterised by the binding affinity or dissociation constant K_d, which is given by the concentrations of unbound binding partners ($[A]$ and $[B]$ in the case of a complex involving two proteins) and the concentration of the complex ($[AB]$) as

$$K_d = \frac{[A] \cdot [B]}{[AB]} \tag{1.1}$$

The binding affinity is a macroscopic quantity and as such not easily accessible to microscopic methods like molecular dynamics simulation. It is however related to the Gibbs free energy of dissociation (ΔG_d) as

$$\Delta G_d = -RT \ln \frac{K_d}{c_0} \tag{1.2}$$

where c_0 is the concentration that defines the standard state ($c_0 = 1\,\mathrm{mol\,L^{-1}}$) and R is the gas constant ($8.3144\,\mathrm{J\,K^{-1}\,mol^{-1}}$). Since this means

$$K_d = c_0 \cdot e^{-\frac{\Delta G_d}{RT}} \tag{1.3}$$

and considering equation 1.1 it is clear that significant concentrations of the complex

will only be observed if ΔG_d is positive (or the binding free energy $\Delta G_a = -\Delta G_d$ is negative).

1.1.1 Binding models
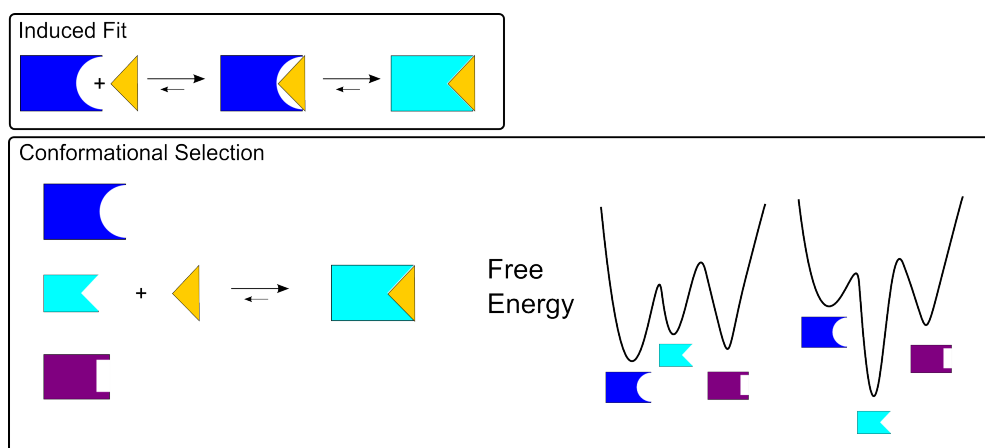
Figure 1.1: Illustration of the alternative binding models of induced fit and conformational selection. According to the induced fit model, the binding partners first form an encounter complex before the conformational changes necessary for the formation of the final complex are induced by the interactions. The induced fit model assumes that the binding partners show different conformations when unbound including those compatible with binding, which are then selected in the binding process as they are energetically favorable.

Different models have been proposed to explain the selective interaction observed between proteins. The lock and key model (22) already established a structural complementarity between the binding partners as possible explanation. However, in some cases, significant differences are observed between the structure of a protein in solution and the structure of the same protein in a complex. Currently, two different models are mainly used to explain these differences (figure 1.1). The induced fit model (23, 24) postulates that after the formation of a preliminary "encounter complex", the interaction between the binding partners induces conformational changes into the complex structures. The conformational selection model (25, 26, 27) (which is similar to the "dynamic fit" model proposed in 1964 (28)) takes into consideration the inherent flexibility of proteins. According to this model, unbound proteins can with a certain probability sample the same conformations as observed when bound. In this model, changes in the free energy landscape of the protein due to interactions in the complex shift the conformational density towards the complex structure upon binding. More recent studies (29, 30) have indicated that elements of both models play a role in protein binding with an initial conformational selection step followed by induced fit rearrangements (31).

1.2   Ubiquitin

An interesting model system to investigate the conformational effects of complex formation is ubiquitin with its binding partners. Ubiquitin is a globular 76 residue protein that plays an important role in biological pathways, as the ubiquitination (the covalent attachment of ubiquitin to a lysine side chain of a protein, also known as ubiquitylation) can among other functions control the degradation or regulate transport of this protein. Structurally (32, 33) ubiquitin mainly consists of a 5-stranded β-sheet, one short and one longer (3.5 turn) α-helix. An important surface feature is the solvent-exposed *hydrophobic patch* (Leu-8, Ile-44, Val-70), which is involved in most binding interfaces. The C-terminus of ubiquitin is exposed and flexible.

An introduction into the history of the discovery of ubiquitin and its functions can be found in (34).

The first function of ubiquitination to be understood and still its most widely known is its role in protein degradation (35, 36, 37). Other functions of ubiquitin besides its role in protein degradation have been found in protein synthesis (38), membrane protein endocytosis (39, 40), regulation of the cell cycle (41), DNA repair (42), transcriptional regulation (43) and stress responses (44, 45).

The process of ubiquitination (figure 1.2) has been extensively researched (46, 47). First, ubiquitin is activated by the ubiquitin-activating enzyme E1 at the expense of energy which is supplied by the degradation of Adenosine triphosphate (ATP) to Adenosine monophosphate (AMP) and inorganic phosphate. The activated ubiquitin then is transferred to an ubiquitin-carrier protein or ubiquitin-conjugating enzyme (E2) and finally attached to a substrate protein in a reaction catalysed by a ubiquitin-protein ligase (E3). While the E1 enzyme is a unique protein in most organisms, specificity of ubiquitination is provided by a variety of different E2 and especially E3 enzymes (of which there are estimated to be more than a thousand in mammalian cells (34)). As a result of this pathway, the carboxylic acid of the C-terminal glycine of ubiquitin is covalently bound to the epsilon amine of a lysine in the substrate. Ubiquitin itself contains seven lysines (residues 6,11,27,29,33,48,63), and additional ubiquitins can be attached by the same mechanism to each of these Lysine side-chains as well as to the N-terminus ("'Linear chains"' (48, 49, 50)). Different configurations of such polyubiquitin chains have been observed in nature. The best known case is a chain of at least four Lys-48-linked ubiquitins, which is necessary to mark a protein for protesomal degradation. Another common case is that of Lys-63-linked ubiquitin chains, but also mixed configurations (51) and even branched ubiquitin chains (52) have been observed. The possibility to form different types of polyubiquitin chains allows ubiquitin to be part of a number of different signalling pathways, with new ones still being discovered. These different ubiquitin signals have to be recognised by corresponding receptor proteins to correctly process the attached protein or affect a signal pathway (53). Finally, when the ubiquitinated protein has reached its destination or as part of a corrective or regulatory function (for example to extend the lifespan of certain proteins otherwise headed to the 26S proteasome), the ubiquitin or polyubiquitin chain has to be removed from the

protein. This is done by deubiquitinating enzymes (DUBs).
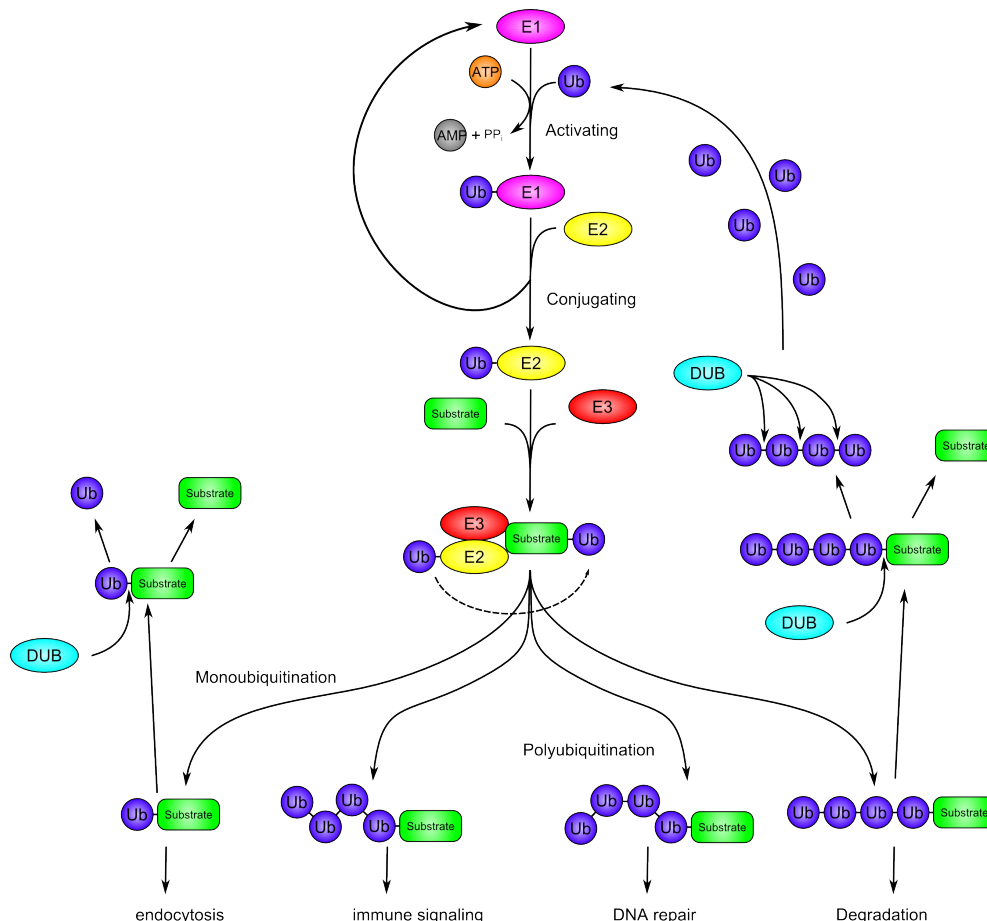
Figure 1.2: Schematic diagram of the ubiquitylation system.

As ubiquitin is involved in such a variety of different cell functions, it has been implicated as essential in a number of different diseases (54) including cancer (55), HIV (56, 57) (which "hijacks" the ubiquitin system to mark a protein for degradation that would otherwise hinder replication of the virus) and different neurodegenerative diseases like Huntington's (58) and Alzheimer's (59).

1.2.1 Ubiquitin binding domains

It is estimated that ubiquitin non-covalently interacts with more than 150 cellular proteins (60), including members of the E1/E2/E3 ubiquitination process, deubiquitinating enzymes or receptor enzymes. Within these, certain classes of structurally similar binding motives or ubiquitin binding domains (60, 61, 62) have been observed. The first domains to be discovered were the ubiquitin-interacting motif (UIM) and the

ubiquitin-associated (UBA) domain. Other significant UBDs include zinc finger motifs for example the polyubiquitin-associated zinc-finger (PAZ) and the Npl4 zinc finger (NZF) motif, the coupling of ubiquitin conjugation to endoplasmic reticulum degradation (CUE) domain, the Gga and TOM1 (GAT), the GRAM-like ubiquitin-binding in Eap45 (GLUE) domain and ubiquitin-conjugating enzyme variant (UEV) motif. A recent list of known UBDs can be found in (60), and new UBDs are still being discovered . Some of these domains preferentially bind specific polyubiquitin chains, in other cases, the receptor protein contains more than one UBD to recognise and bind to polyubiquitin chains.

1.2.2 Ubiquitin complex structures

In this function, ubiquitin is recognised by and interacts with a multitude of other proteins. For several of these complexes, the structures have been experimentally determined using either X-ray crystallography or nuclear magnetic resonance spectroscopy. As of February 2013, the Protein database (PDB (63)) contains 140 structures that contain at least one chain of wild type human ubiquitin. Many of these are complex structures. For this study, a selection of the available structures was made according to three main criteria. As the *pincer mode* of ubiquitin has already been implied to play an important role in binding (12), so care was taken to consider complexes containing ubiquitin in a variety of conformations along this mode. To reduce computational cost and in view of a possible collaboration involving the use of RDC measurements, complexes with small binding partners were prefered over ones involving big binding partners. Finally, high resolution X-ray structures with few missing atoms were prefered over lower quality structures to supply reliable starting structures for the simulation. As a result, eleven structures (table 1.1 and figure 1.3) selected from the PDB for simulation. An additional X-ray structure (*dsk2*) of ubiquitin in complex with the UBA domain of Dsk2p was supplied by Stefan Becker from the department of NMR-based Structural Biology of the Max Planck Institute for Biophysical Chemistry.

1.3 Previous work

In 2008, Lange et al. (12, 75) used residual dipolar couplings to derived an ensemble of ubiquitin in solution, which showed significantly more structural variety than ensembles of unbound ubiquitin that had previously been derived from NMR experiments (76, 77). They compared this ensemble with 46 crystal structures from the PDB, most of which were of ubiquitin in complex with a binding partner. As the backbone structure of each of the X-ray structures was within a 0.8 Å RMSD of the nearest structure from the NMR ensemble, they concluded that the conformational selection model was sufficient to explain the structural diversity of bound ubiquitin on the backbone level. They also identified those residues of ubiquitin usually involved in binding to be more flexible than those not involved in binding, so it can be assumed that this flexibility of ubiquitin is essential to its ability to bind a variety of binding partners. Finally, they identified
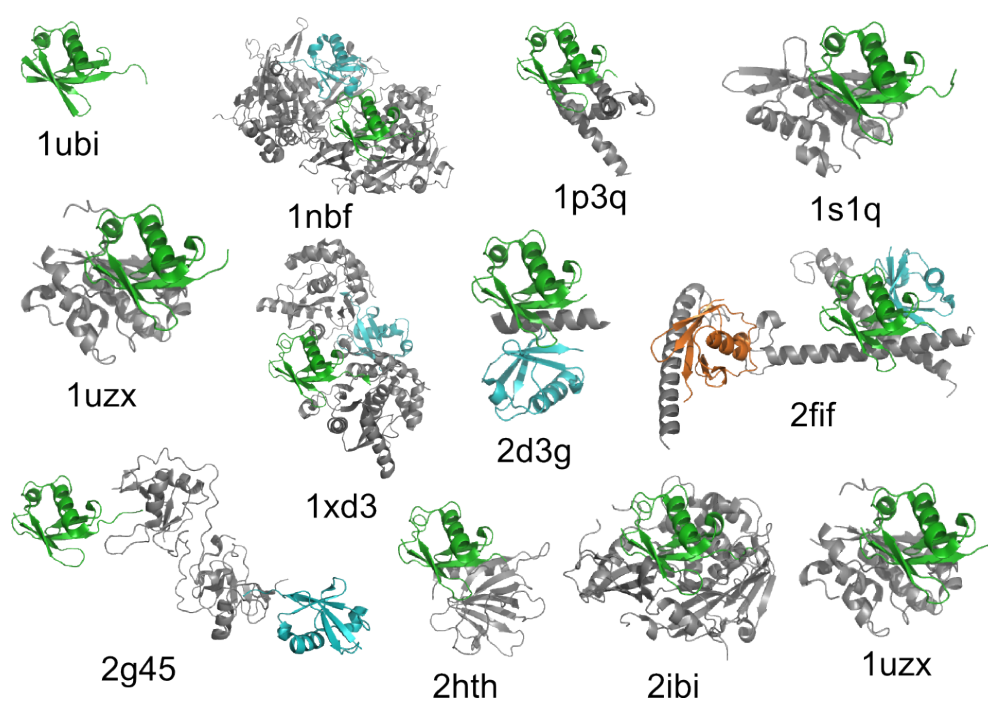
Figure 1.3: A collection of ubiquitin complex structures used in this study. The binding partners are drawn in grey, different ubiquitin chains in the same X-ray structure are coloured differently. Structures are labelled with the corresponding PDB code.

1. INTRODUCTION

Table 1.1: Structures used for simulation setup.

PDB code	binding partner	resolution (Å)	reference
1NBF	Ubiquitin carboxyl-terminal hydrolase 7 (HAUSP)	2.30	(64)
1P3Q	CUE domain of Vacuolar protein sorting associated protein (Vps9p)	1.70	(65)
1S1Q	Tumor susceptibility gene 101 protein (TSG101)	2.00	(66)
1UBI	none (unbound reference)	1.80	(33)
1UBQ	none (unbound reference)	1.80	(32)
1UZX	UEV domain of Vps23	1.85	(67)
1XD3	Ubiquitin Carboxyl-terminal esterase L3 (UCH-L3)	1.45	(68)
2D3G	UIM from hepatocyte growth factor-regulated tyrosine kinase substrate (Hrs-UIM)	1.70	(69)
2FIF	Rab5 GDP/GTP exchange factor	2.49	(70)
2G45	Ubiquitin carboxyl-terminal hydrolase 5	1.99	(71)
2HTH	Vacuolar protein sorting protein 36	2.70	(72)
2IBI	Ubiquitin carboxyl-terminal hydrolase 2	2.20	(73)
2OOB	E3 ubiquitin-protein ligase CBL-B	1.90	(74)
dsk2	UBA domain of Dsk2p	2.00	

the a collective mode of motion (the *pincer mode*) as the main mode of fluctuation responsible for binding.

Wlodarski and Zagrovic (30) performed a more thorough statistical analysis on the structural ensemble published by Lange et al. in comparison to a similar set of X-ray structures. They found that only a small number of structures from the NMR ensemble constitute the most similar to all bound X-ray structures, deducing that the unbound protein only rarely enters conformations compatible with binding. They also identified statistically significant differences between bound and unbound protein structures near the binding interface by calculating the average atomic rmsd between each bound structure and the most similar unbound structure considering only atoms up to a certain distance from the binding partner. They concluded that ubiquitin shows considerable "residual induced fit" which they consider to be "comparable in magnitude to conformational selection" observed in ubiquitin.

It has recently been shown (78) however, that the observed differences between the experimental bound structures and a molecular dynamics (MD) ensemble of unbound ubiquitin decrease with an increasing number of snapshots considered from the simulation ensemble indicating that indeed conformational selection largely suffices to explaining the conformational heterogeneity of ubiquitin in different complexes.

1.4 Aim and organisation of this thesis

This thesis aims to give a detailed description of the effect of non-covalent binding on the dynamics of ubiquitin using MD simulations.

The principles of MD simulations and the methods used in the analysis of the dynamics ensembles generated by MD are described in chapter 2. In chapter 3, these methods are applied to investigate and analyse the dynamic properties of a selection of ubiquitin complexes (listed in table 1.1). In particular, characteristic effects on the *pincer mode*, the main mode of fluctuation of unbound ubiquitin are found to be the main effect of binding. Similar effects can be reached by the mutation on ubiquitin residues as will be shown in chapter 4, where also the effect of such mutations on binding will be investigated. Finally, the results of this thesis will be discussed in chapter 5 including a short outlook on open questions and possible future investigations.

1. INTRODUCTION

2

Theory and Methods

2.1 Molecular dynamics simulations

Molecular dynamics (MD) simulations produce atomic resolution dynamic ensembles of molecular systems. In this study, these are used to investigate the dynamics of ubiquitin bound to various binding partners and compare them to the dynamics of unbound ubiquitin. The first part of the following chapter describes the principles and approximations of MD simulations, the second part introduces MD-based methods to determine free energy differences and finally the methods used to analyse and compare the ensembles are presented.

2.1.1 Approximations used in MD

2.1.1.1 Born-Oppenheimer approximation

The state of a molecular system evolves according to the time-dependent Schrödinger equation:

$$i\hbar \frac{\partial}{\partial t} \psi(\mathbf{r}, \mathbf{R}) = \mathcal{H} \psi(\mathbf{r}, \mathbf{R}) \tag{2.1}$$

where $\mathbf{r}$ and $\mathbf{R}$ are the positions of electrons and nuclei, ψ is the wave function and $\mathcal{H}$ the Hamiltonian operator. Evaluating this equation is not feasible for large molecules like proteins. For systems containing more than two independent particles, no analytical solution of the Schrödinger equation exists, and numerical approaches become computationally demanding for systems containing more than a relatively small number of atoms. On the other hand, in most cases the precise time evolution of the wave functions is not necessary to the understanding of biological functions so that a more coarse view of protein dynamics often is appropriate. A number of simplifications are being used to model molecular dynamics on an atomic scale while reducing the computational complexity. Electron dynamics happen on timescales significantly faster than those of nuclei, since the mass of electrons is several orders of magnitude smaller than that of nuclei. To simplify the Schrödinger equation, the Born Oppenheimer

approximation (79) can be used, which decouples the wave functions for electrons and nuclei:

$$\psi(\mathbf{r}, \mathbf{R}) = \psi_n(\mathbf{R}) \cdot \psi_{e;\mathbf{R}}(\mathbf{r}) \tag{2.2}$$

As electrons posses a mass which is several orders of magnitude smaller than that of the nuclei, there is a separation of timescales between fast electrons and slow nuclei. The electron wave function $\psi_{e;\mathbf{R}}(\mathbf{r})$ can be calculated under the assumption that the nuclei supply an electric potential that does not change in time. For the calculation of the nuclear wave function $\psi_n(\mathbf{R})$, the electrons can be assumed to adjust *infinitely fast* to the nuclear position.

2.1.1.2   Classical description of nuclear dynamics

Using the Born-Oppenheimer approximation, atomic electrons are considered to move with the corresponding nuclei. Consequently, we now consider the motion of the nuclei (as electronic motion is not separately considered, these are equivalent to atoms), which can be considered to behave like point particles that move according to Newton's equations of motion:

$$m_i \frac{\partial \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i, \qquad i = 1 \ldots N \tag{2.3}$$

where m_i and $\mathbf{r}_i$ are the mass and position of atom i, while $\mathbf{F}_i$ is the force acting on it. The forces are calculated as the gradient of the potential energy function $V(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ describing all inter-atomic interactions:

$$\mathbf{F}_i = \nabla V(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \tag{2.4}$$

The potential energy function is calculated from simple mathematical functions defined in the force-field (section 2.1.1.3). This treatment of atoms as classical particles is justified in most cases, as the deBroglie wavelength of the atoms under the conditions considered is considerably smaller than the distance of the atoms (the deBroglie wavelength of a carbon atom at $T = 300K$ is $\lambda_{dB} \approx 0.4\,\text{Å}$, while a C-C bond is $d \approx 1.5\,\text{Å}$ long). Another criterion is the treatment of bond vibrations - quantum mechanical calculations are necessary if the resonance frequency of a harmonic oscillator exceeds $k_B T/h$ which is approximately 200cm^{-1}. As can be seen in table 2.1, different bond vibrations, mainly those involving hydrogen, are close to this value (80). As will be seen in section 2.1.5, most of the fast bond vibrations are removed from the system by the use of constraints - while this is mainly done to increase the possible timestep in the simulation and hence improve simulation speed, it also has the advantage that the resulting constrained bonds correspond to the quantum-mechanical harmonic oscillator in its ground state better than a classical harmonic oscillator would.

Table 2.1: Typical vibrational frequencies in molecular systems, according to (80)

type of bond	type of vibration	wavenumber (cm^{-1})
C-H, O-H, N-H	stretch	3000-3500
C=C, C=O	stretch	1700-2000
HOH	bending	1600
C-C	stretch	1400-1600
H_2CX	sciss,rock	1000-1500
CCC	bending	800-1000
O-H...O	libration	400-700
O-H...O	stretch	50-200

2.1.1.3 Force Field

The interaction forces between the atoms in a molecular system are determined by quantum mechanical effects. As these can not be calculated with reasonable effort for systems consisting of $\approx N \geq 100$ atoms (which includes all systems considered in this work), these interactions have to be modelled in a way consistent with the classical approximation used in MD simulations. For this, empirical or semi-empirical force fields have been devised, whose parameters are determined from quantum mechanical calculations or adapted to reproduce experimental observables for simple model systems. Force field parametrisation is an ongoing field of research. Existing force fields like the gromos force field (81, 82), the charmm force field (83, 84) oplsaa (85) and different iterations of the amber force field (86, 87, 88) (which includes amber99sb, the force field used in most simulations in this study) are continuously adapted to fit new experimental observations (89) or to maintain plausible results for new, long timescale simulations (90).

Bonded interactions are depending on bond-lengths (2.6), bond angles (2.7), dihedral angles (2.8) and improper dihedrals (2.9). As each atom is only involved in a given number of covalent bonds which does not change over the course of the simulation, the computational effort necessary to calculate these bonded interactions increases linearly with the number of atoms in the system. The non-bonded interactions are approximated by a Lennard-Jones potential ((91), eqn. 2.10), a mathematically simple model of the quantum mechanical interaction between a pair of neutral atoms (which consist of the attractive van der Waals force between induced dipole moments and the repulsive exchange interaction which is based on the fact that two electrons cannot occupy the same quantum mechanical state) and the Coulomb interactions (2.11) between charged particles. These forces act between each pair of atoms in the system, increasing the computational effort necessary to compute them quadratically with the system size and hence severely limit the maximal size of the systems accessible to the method. The electrostatic interaction between two particles depends on their respective charges. In MD simulations, the particles represent atoms, most of them involved in covalent chemical bonds, resulting in shared and sometimes (as in the case of aromatic rings) delocalised charges. To model this, atoms are assigned effective charges based on the electric wave

function of the molecule they are part of, which is assumed to depend only on the immediate molecular surroundings (in the context of proteins, this is usually the amino-acid residue in which the corresponding atom is located). It is calculated for model systems using numerical quantum mechanical methods and considered to be constant over the course of the simulation, excluding polarisation effects as well as chemical reactions which require the formation of breaking of covalent bonds or the relocation of electrons. In practice, non-bonded interactions (especially the Lennard-Jones potential) are negligibly small for distances larger than $1.4\,\mathrm{nm}$. Optimisation methods like neighbour lists, cut-offs and shift functions (80) use this to limit the number of non-bonded interactions that have to be calculated each step and hence improve simulation performance.

$$V(r) = V_{bonded}(r) + V_{nonbonded}(r) \tag{2.5}$$

$$= \sum_{\text{bonds } i} \frac{k_i}{2} (b_i - b_{i,0})^2 \tag{2.6}$$

$$+ \sum_{\text{angles } i} \frac{f_i}{2} (\theta_i - \theta_{i,0})^2 \tag{2.7}$$

$$+ \sum_{\text{dihedrals } i} \frac{V_i}{2} [1 + \cos(n\psi)] \tag{2.8}$$

$$+ \sum_{\text{impropers } i} \kappa_i (\xi_i - \xi_{i,0})^2 \tag{2.9}$$

$$+ \sum_{\text{pairs } i,j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{d_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{d_{ij}} \right)^6 \right] \tag{2.10}$$

$$+ \sum_{\text{pairs } i,j} \frac{q_i q_j}{4\pi\epsilon_0 \epsilon_r r_{ij}} \tag{2.11}$$

2.1.2   Water model

The natural environment of proteins consists mainly of water. MD simulations in this study were conducted with the proteins surrounded by a large number of explicitly modelled water molecules (and ions corresponding to 150 mM of NaCl). While "implicit solvent" methods are being developed for MD, their results are generally less accurate than those of simulations performed with explicit water (92, 93, 94). As water makes up a significant part of the simulation box, a number of different models have been developed to simulate water both accurate and efficiently (95, 96, 97). The SPC/E water model (97) used here is a three site model which considers both bond length and bond angles to be constant. As constant charges are assigned to the atoms of the water molecule, polarisability is not modelled. To correct for this, the SPC/E model adds a
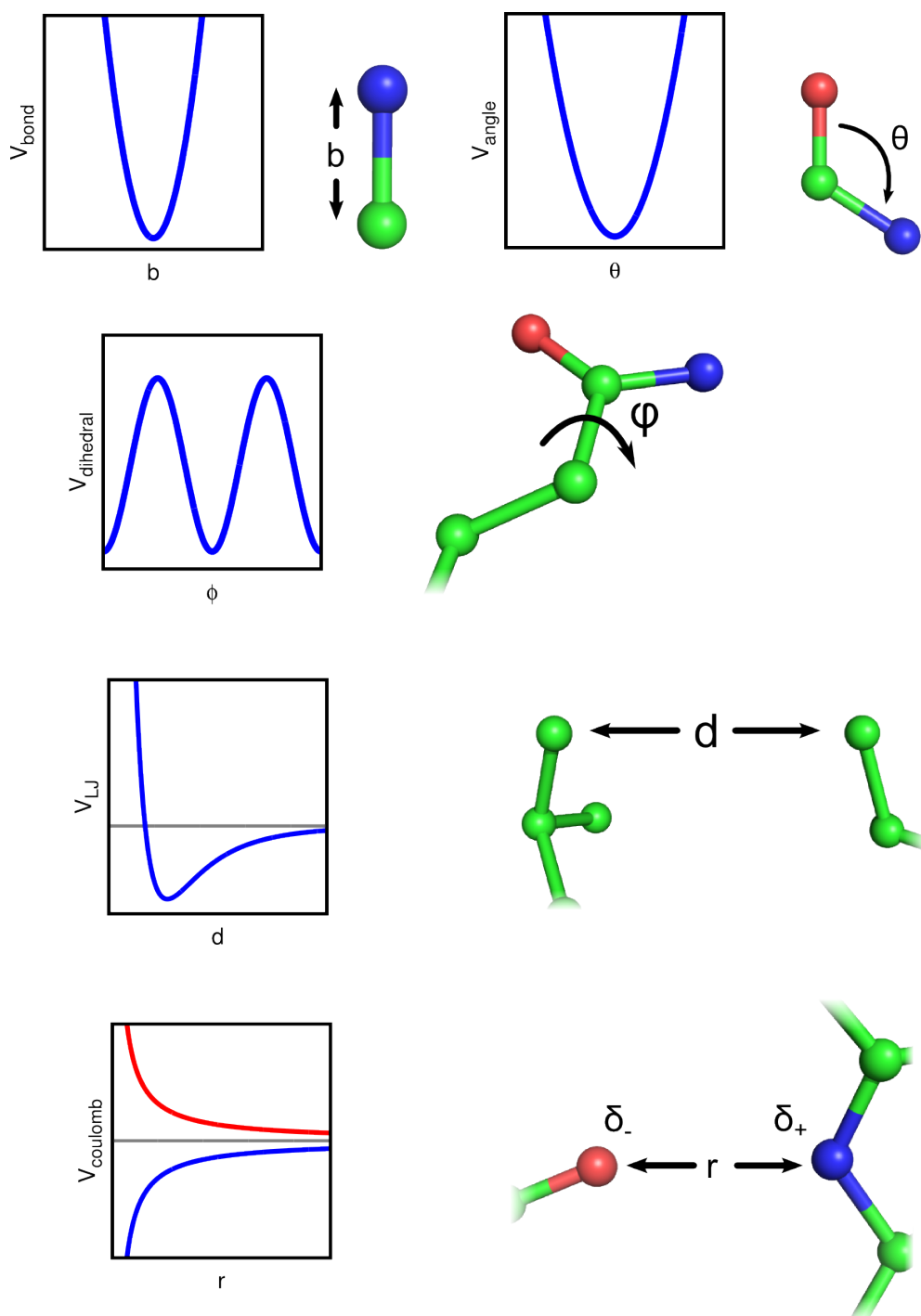
Figure 2.1: Different components of classical force fields.

correction term 2.12 to the potential energy function,

$$E_{pol} = \frac{1}{2} \sum_i \frac{\left(\mu - \mu^0\right)^2}{\alpha_i} \tag{2.12}$$

where $\mu = 2.35\text{D}$ is the dipole of the effectively polarised water molecule, $\mu^0 = 1.85\text{D}$ is the dipole moment of an isolated water molecule and $\alpha_i = 1.608 \times 10^{-40}\text{Fm}^2$ is an isotropic polarisability constant.

2.1.3 Integrating the equations of motion

Even in the classical approximation, the equations of motion for a system of $n \geq 3$ bodies cannot be solved analytically and numeric integration has to be used. A number of different numeric integration schemes exist. The one used in this work is the leapfrog algorithm (eqn 2.13, 2.14) (98) which uses a discrete timestep Δt:

$$\mathbf{v}_i\left(t + \frac{\Delta t}{2}\right) = \mathbf{v}_i\left(t - \frac{\Delta t}{2}\right) + \frac{\mathbf{F}_i(t)}{m_i}\Delta t \tag{2.13}$$

$$\mathbf{r}_i\left(t + \Delta t\right) = \mathbf{r}_i(t) + v_i\left(t + \frac{\Delta t}{2}\right)\Delta t \tag{2.14}$$

Both the velocity $\mathbf{v}_i$ of and the force $\mathbf{F}_i$ acting on each particle are assumed to be constant over the duration Δt, so it has to be chosen small enough to correctly integrate the fastest dynamics in the system. In molecular systems, these are bond oscillations which have periods on the order of 10fs.

As at least five integration steps should be performed per period to accurately model the dynamics using a leapfrog algorithm (99), this would require a timestep of $\Delta t \leq 2\text{fs}$. The time-step can be increased if steps are taken to remove fast oscillations from the system (see section 2.1.5).

2.1.4 Boundary conditions

2.1.4.1 Periodic boundary conditions

As simulations are performed in an aqueous environment, an increase of the simulation system goes with an increase of the atoms to be simulated and consequently of the computational effort nec-
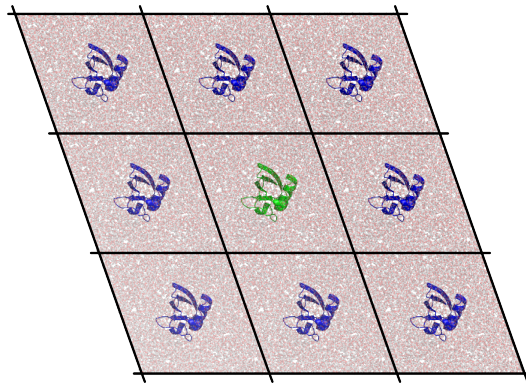

Figure 2.2: Periodic boundary conditions. The original system (green protein) is surrounded by its periodic images (blue protein). For better visibility, only one layer of the lattice is shown.

16

essary to simulate each time-step. Be-
cause of this, it is desirable to use a simulation box which is as small as possible while
modelling the environment as realistically as possible. However, when the part of the
system (usually the protein) is too close to the boundary of the system, boundary effects
could distort its behaviour. Using periodic boundary conditions (figure 2.2), boundary
effects are avoided. By transforming coordinates "periodic images" of the simulation
box are created. Particles leaving the simulation box reenter it at the opposite side and
particles near the boundary interact across the periodic boundary.

While periodic boundary conditions avoid effects caused by "artificial" boundaries
of the simulation system, they too can cause artefacts. Care has to be taken to avoid
the interaction of a particle with two copies of the same other particle or even with
itself, for example by modifying interaction potentials to vanish for distances greater
equal half the box length. This can be achieved in multiple ways including truncation,
switch or shift (100) functions (101), while electrostatic interactions can be handled
using the PME method (section 2.1.4.2).

A simulation box with periodic boundary conditions can have a variety of shapes,
as long as it is space filling. In the most general case, to minimise the volume of the
system (and hence the number of water molecules that need to be simulated) while
maximising the distance between the protein and it's periodic image, a space filling
shape best approximating a sphere should be chosen - for this, the dodecahedron (which
however is represented as a triclinic box in the computer (102)) has the optimal shape.
For non-globular proteins, non-symmetric (e.g. rectilinear) simulation boxes can be
more optimal, but in these cases care has to be taken to restrict the rotation of the
protein.

2.1.4.2   Efficient handling of electrostatics - PME

As mentioned in section 2.1.1.3, the computationally most expensive interactions in the
molecular system are the electrostatic ones, which are described by the coulomb potential. While bonded interactions are limited to a relatively small, unchanging number
of bonds and the Lennard-Jones potential quickly approaches zero (being proportional
to r^{-6}), the coulomb potential (which is proportional to r^{-1}) cannot simply be ignored
even at longer distances. An efficient handling of electrostatic interactions is Particle
Mesh Ewald (PME) summation (103, 104). It is based on the fact that the electrostatic
potential can be separated into two terms

$$V_{\text{Coulomb}}(\mathbf{r}) = V_{\text{Coulomb,sr}}(\mathbf{r}) + V_{\text{Coulomb,lr}}(\mathbf{r}) \tag{2.15}$$

where one part (the short ranged part $V_{\text{Coulomb,sr}}(\mathbf{r})$) quickly converges in real
space, while the other (the long ranged $V_{\text{Coulomb,lr}}(\mathbf{r})$) quickly converges in reciprocal (or Fourier) space. This way, the computational cost of calculating the electrostatic
interactions of N atoms is reduced from $O(N^2)$ to $O(N \cdot \log N)$.

2.1.4.3 Temperature and Pressure coupling

In its original form, MD simulates a constant number of particles in a constant volume and (barring numerical integration errors) obeys conservation of energy as all potentials are conservative. The simulated ensemble is a NVE or *microcanonical* ensemble. A simulation of a constant number of particles, constant Volume and constant Temperature (NVT), also called a *canonical* ensemble, more closely resembles the natural state of the protein. A number of methods (so called thermostats (105, 106, 107, 108)) have been developed to keep the simulation system at a constant temperature. In general, they adjust the velocities of the particles in the system so that the kinetic energy K correspond to

$$K = \frac{3N}{2} k_B T \tag{2.16}$$

where N is the number of atoms and T is the set temperature. It can also be necessary to simulate at constant pressure instead of constant volume (NPT ensemble). A number of different pressure coupling algorithms (105, 109, 110, 111) adaptively change the simulation box size to obtain a constant pressure in the system.

2.1.5 Optimisation methods

Biologically relevant dynamics of proteins usually are collective, correlated motions of significant parts of the protein over relatively long ($\geq 1ns$) timescales. As mentioned before (section 2.1.3), the integration timestep, and hence the computational effort necessary to compute a given time span of protein dynamics is determined by the fastest motions in the system, which usually are small scale, uncorrelated bond oscillations (99). The fastest oscillations in MD (in order of increasing period according to (112)) are bond stretch vibrations (10-30fs) and bond angle vibrations involving hydrogen (13-32fs). Different methods have been developed to "average out" these fast vibrations, allowing for a longer time-step and hence more efficient simulations.

2.1.5.1 SETTLE and Lincs

The fastest motions in a macromolecular system are bond vibrations that have a high frequency but low amplitude and are generally not functionally relevant. In explicit solvent simulations, water molecules usually make up the biggest portion of the simulated system (contributing more than 10 times as many atoms than the protein itself). Hence, an efficient simulation of water significantly improves the performance of MD simulations. The SETTLE algorithm (113) makes it possible to assume rigid bonds and angles in small molecules like water. It enforces constant bond lengths by moving atoms parallel to the bond direction. For larger molecules, LINCS (LINear Constraint Solver (114)) is an efficient algorithm to accomplish the constant bond length. The assumption of constant bond lengths has the additional advantage that it reduces errors due quantum mechanical effects which would be neglected in the classical approximation of these oscillations.

2.1.5.2 Virtual sites

By removing fast bond length vibrations using SETTLE and LINCS, a simulation time step of $\Delta t = 2$fs can be used. This time step is necessary to correctly sample high-frequency bond angle vibrations involving hydrogen atoms. Feenstra et al. ((112)) have developed a method to remove these degrees of freedom altogether. Hydrogen atoms are turned into mass-less "dummy atoms", whose positions are geometrically constructed from the positions of three nearby heavy atoms. Forces acting on the hydrogen are distributed over these heavy atoms. Also, to keep the mass of the system constant, the hydrogen-mass is added to the heavy atom it is bound to. This way, the timestep can be increased to $\Delta t = 4$fs or even $\Delta t = 5$fs

2.2 Free Energy Calculations

The Gibbs free energy (G) at is given by

$$G = H - TS \tag{2.17}$$

where H is the system's enthalpy, T is the temperature and S is the entropy of the system. It is a thermodynamic potential that describes the work that can be obtained from a process at constant pressure and temperature (in contrast to the Helmholtz Free energy which describes the same property at constant volume and temperature). Protein behaviour and reactions are governed by free energy differences. Protein folding and complex formation occur along free energy gradients. In principle, if the free energy difference between two states is given by ΔG, the probability ρ_1 to find the system (in thermodynamic equilibrium) in state 1 compared to the probability ρ_2 to find it in state 2 at temperature T is given by

$$\frac{\rho_1}{\rho_2} = e^{-\Delta G / k_B T} \tag{2.18}$$

obviously, if ΔG is significantly greater than $k_B T$ (or smaller than $-k_B T$), the system will almost certainly be found in one of the two states. In Chapter 4, this principle will be used to determine if mutations of ubiquitin selectively destabilise complexes with selected binding partners and hence can be used to selectively influence the binding behaviour.

Like the entropy, which it depends on, free energy is an ensemble property, and cannot simply be calculated as an ensemble average like the enthalpy. This turns the calculation of the *absolute* free energy of a system using MD simulations into a very challenging task, as it would require the exhaustive exploration of the phase space available to the system. As this exploration is limited by the sampling time, the results of such calculations are usually very inaccurate. However, the calculation of *relative* free energy differences between two states of the system, while still challenging, usually results in more accurate results.

2.2.1 Thermodynamic Cycles

The determination of binding affinities is of high interest and has several applications including docking and drug design. This has led to a number of different approaches to calculate or estimate binding free energies, including both empiric (knowledge-based) and physics-based methods (14). In this work, we will focus on physics-based methods, especially free energy perturbations, which are based on MD simulations. Free energy perturbation methods calculate the work needed to "move" a system from one state to another, not necessarily through physically meaningful paths. This is expressed by a change from the Hamiltonian of the system in one state ($\mathcal{H}_1$) to that describing the other state ($\mathcal{H}_2$):

$$\mathcal{H}_\lambda = \lambda \cdot \mathcal{H}_1 + (1 - \lambda) \cdot \mathcal{H}_2, \quad \lambda \in [0, 1] \tag{2.19}$$

To determine the free energy difference ΔG due to the difference between state 1 and state 2, different methods have been developed including Thermodynamic Integration (TI, section 2.2.3) and the Crooks Gaussian intersection method (section 2.2.5), which has been used in chapter 4. Sufficient sampling is of critical importance in all free energy methods, and the computational difficulty of reaching convergence increases dramatically with the magnitude of the perturbation - the difference between the two Hamiltonians $\mathcal{H}_1$ and $\mathcal{H}_2$.

Binding free energies result from a large perturbation - in one state, both binding partners are free in solution, in the other they are in the complex. This renders binding free energy calculations intangible by perturbation methods. A transition from the protein wild type to the mutated protein, while physically impossible in reality, requires a perturbation which is much smaller and can actually be performed in a simulation by removing some atoms from the system ("turning off" all of their interactions) and inserting other ("turning on" their interactions). As we are interested in the difference between the binding free energies of the wild type and the mutant, we can make use of the fact that the free energy is a state property which is the same independently of how the state was reached. Hence, the free energy differences along a closed circle of reactions (like the one depicted in figure 2.3) will always add up to be zero.

Using this fact, we can calculate the difference $\Delta\Delta G$ between the binding free energies of wild type and mutant from the difference of the free energy differences of mutation for the bound and unbound protein as

$$\Delta\Delta G = \Delta G_1 - \Delta G_2 = \Delta G_3 - \Delta G_4 \tag{2.20}$$

2.2.2 Free Energy Perturbation

Free energy perturbation (FEP)(115) can be used to calculate the free energy difference between a systems in state 1 described by the Hamiltonian $\mathcal{H}_1$ and state 2 described
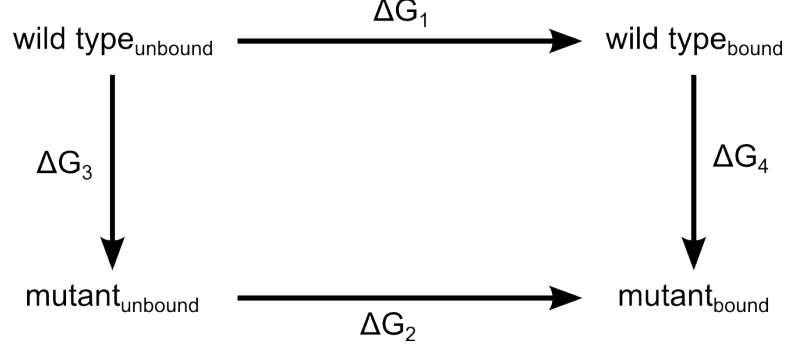
Figure 2.3: Example of a thermodynamic cycle. To determine whether a mutation changes the binding affinity of a protein to another one, the difference between the binding free energy of the wild type (ΔG_1) and that of the mutated protein (ΔG_2) needs to be determined. However, the values (ΔG_3) and (ΔG_4) are more accessible to free energy perturbation methods. Using the conservation of energy in this closed cycle, we can conclude $\Delta\Delta G = \Delta G_1 - \Delta G_2 = \Delta G_3 - \Delta G_4$.

by $\mathcal{H}_2$ using

$$\Delta G = -\frac{1}{\beta} ln \left\langle e^{-\beta(\mathcal{H}_2 - \mathcal{H}_1)} \right\rangle_1 \qquad (2.21)$$

where $\langle\rangle_1$ denotes an ensemble average of the system in state 1. While FEP has been used to calculate the free energy differences resulting from amino acid substitutions, they require excessive sampling, especially for large perturbations (e.g. including the addition or removal of atoms)(116, 117, 118).

2.2.3 Thermodynamic Integration

Thermodynamic Integration (TI) (119) uses the generalised force $\partial\mathcal{H}/\partial\lambda$ to calculate the free energy difference via

$$\Delta G = \int_0^1 \frac{\partial\mathcal{H}}{\partial\lambda} d\lambda \qquad (2.22)$$

where the coupling parameter λ is either changed slowly over time (slow-growth thermodynamic integration - SGTI) or set to a number of discrete values between 0 and 1 in which case the integral in equation 2.22 turns into a sum

$$\Delta G = \int_0^1 \frac{\partial\mathcal{H}}{\partial\lambda} d\lambda \approx \sum_{i=0}^N \left\langle \frac{\partial\mathcal{H}}{\partial\lambda} \right\rangle_{\lambda_i} \Delta\lambda \qquad (2.23)$$

both in the case of this discrete thermodynamics integration (DTI) and SGTI, it is assumed that the system is at (or very close to) equilibrium over the cause of the simulation. To achieve this, the change of λ has to be very slow, resulting in high

computational costs.

2.2.4 Jarzynski's equation and Crooks' fluctuation theorem

According to Jarzynski(120, 121), the free energy difference ΔG between two states can be calculated from a series of short, non-equilibrium transitions as

$$e^{-\beta \Delta G} = \left\langle e^{-\beta W_\tau} \right\rangle \qquad (2.24)$$

where $\langle \rangle$ denotes the average over an ensemble of transitions which were started from an equilibrated canonical ensemble. W_τ is the work

$$W_\tau = \int_0^1 \frac{\partial \mathcal{H}_\lambda}{\partial \lambda} d\lambda \qquad (2.25)$$

integrated over the transition. Fundamentally, the work W_τ consists of the free energy difference and the dissipation of energy due to the fact that the transition was not performed in equilibrium (i.e. W_τ is closer to ΔG, the slower the transition is performed up to the case where the system is in equilibrium all the time, in which $W_\tau = \Delta G$ which is the case assumed in thermodynamic integration). Hence, this approach is limited by the fact that transitions with a high weight occur only with a low probability, making it necessary to perform a high number of transition simulations. As a result, the overall computational cost is similar to that of equilibrium methods like SGTI and DTI (122, 123).

By performing transitions both from state 1 to state 2 (*forward*) and from state 2 to state 1 (*reverse*), this problem can be circumvented. Crooks' fluctuation theorem (124) states that the distributions P_f (for forward transitions) and P_r (for reverse transitions) are related to the free energy difference by

$$\frac{P_f(W)}{P_r(-W)} = e^{\beta(W-\Delta G)} \qquad (2.26)$$

2.2.5 Gaussian Intersection method

From equation 2.26 it follows that the free energy difference between two states is $\Delta G = W_i$ for the work for which the transition distributions intersect ($P_f(W_i) = P_r(-W_i)$). Unfortunately, the population of the distributions at the intersection point usually is small, resulting in large statistical errors. Fortunately, these work distributions can be approximated by a Gaussian distribution (125), resulting in the Crooks Gaussian Intersection (CGI) method (126). After fitting forward and backward transitions to two Gaussian distributions

$$P_f(W) = \frac{1}{\sigma_f \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{W-\bar{W}_f}{\sigma_f}\right)^2} \tag{2.27}$$

$$P_r(W) = \frac{1}{\sigma_r \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{W-\bar{W}_r}{\sigma_r}\right)^2} \tag{2.28}$$

the intersection point (and hence the free energy difference) is given by

$$\Delta G_{CGI} = \frac{\frac{W_f}{\sigma_f^2} - \frac{W_r}{\sigma_r^2} \pm \sqrt{\frac{1}{\sigma_f^2 \sigma_r^2}(W_f + W_r)^2 + 2\left(\frac{1}{\sigma_f^2} - \frac{1}{\sigma_r^2}\right)\ln\frac{\sigma_r}{\sigma_f}}}{\frac{1}{\sigma_f^2} - \frac{1}{\sigma_r^2}} \tag{2.29}$$

2.2.6 Umbrella sampling

Protein dynamics are often described as the movement of the protein on its "energy landscape". The potential of mean force (PMF) is a free energy with respect to a generalised coordinate ξ (or a set of such variables) that depends on the coordinates of the system (127). It was introduced by Kirkwood in 1935 (119) and is defined as

$$W(\xi) = W(\xi_0) - k_B T \ln\left[\frac{\langle \rho(\xi) \rangle}{\langle \rho(\xi_0) \rangle}\right] \tag{2.30}$$

where $\rho(\xi)$ is the probability distribution of finding the system at position ξ in equilibrium. The potential is defined up to a constant and is set to an arbitrary value $W(\xi_0)$ at and arbitrary position ξ_0.

For a converged equilibrium ensemble, the PMF can directly be calculated from the distribution of structures along the generalised coordinate according to equation 2.30, but in many cases, especially high energy barriers are present, this convergence is hard to reach in standard MD simulations. Different methods have been developed to overcome this problem and improve sampling in such energetically unfavourable states. One of these approaches is Umbrella sampling (128). To ensure good sampling along the whole range of ξ, a set of N_w simulations are performed, each with an additional "umbrella potential"

$$w_i(\xi) = K_i/2(\xi - \xi_i^c)^2 \tag{2.31}$$

which restrains the system to positions $\xi_i^c (i = 1 \ldots N_w)$ with a force constant K_i. The position of the system in the reaction coordinate is recorded and an umbrella histogram $h_i(\xi)$ is calculated for each simulation. To calculate a free energy profile for the system along ξ, the influence of the umbrella potential $w_i(\xi)$ has to be considered. A well established method to calculate the PMF from these biased histograms is the weighted histogram analysis method (WHAM) (129, 130), which calculates the unbiased

distribution $P(\xi)$ using

$$P(\xi) = \frac{\sum\limits_{i=1}^{N_w} g_i^{-1} h_i(\xi)}{\sum\limits_{j=1}^{N_w} n_j g_j^{-1} \exp\left[-\beta(w_j(\xi) - f_j\right]} \tag{2.32}$$

and

$$\exp(-\beta f_j) = \int \mathrm{d}\xi \exp\left[-\beta w_j(\xi)\right] P(\xi) \tag{2.33}$$

β denotes the inverse temperature $1/k_B T$ with the Boltzmann factor k_B and the temperature T. n_j is the total number of points in the corresponding histogram h_j and the statistical inefficiency $g_i = 1 + 2\tau_i$ with τ_i, the integrated autocorrelation time of the corresponding umbrella window.

2.3 Trajectory analysis

During MD simulations, the positions of particles of interest (usually at least the atoms of the protein) are saved at regular time intervals. This collection of structures is called a trajectory and describes the time-evolution of the system. The amount of data generated this way is huge, and trajectories contain significant noise that is not directly relevant to biological function. Relevant dynamics are usually realised by slow, large scale, collective motions in the protein, while most motions occurring in the protein (and hence generated by MD simulations) are fast, small scale local vibrations. To extract relevant information from the noisy collection of data in MD trajectories, a number of statistical methods has to be employed, which will be described in the following sections.

2.3.1 Principal Component Analysis

The conformation of a molecular system consisting of N atoms (with $N > 100$ in most biologically relevant systems) can be described by $3N$ coordinates. Hence, a MD trajectory of such a system is a trajectory in $3N$-dimensional space, with each coordinate containing only little relevant information. Biologically relevant motions usually involve the collective motion of whole parts of the protein. Principal component analysis (131, 132, 133) is a well established statistical method that can be used to identify such collective modes of motion in proteins. Mathematically, PCA uses an orthogonal transformation to transform a set of variables (in this case the $3N$ Cartesian coordinates) into a (possibly smaller) set of linearly uncorrelated generalised coordinates. This can be accomplished by diagonalisation of the covariance matrix

$$\mathbf{C} = \left\langle \left(\mathbf{r_t} - \langle \mathbf{r} \rangle\right) \left(\mathbf{r_t} - \langle \mathbf{r} \rangle\right)^T \right\rangle \tag{2.34}$$

where $\mathbf{r_1}, \mathbf{r_2} \dots \mathbf{r_T}$ denote the T points of the trajectory (in this case the structures in the trajectory, but the ensemble can be more general). PCA consists of finding the orthonormal transformation matrix $\mathbf{R}$

$$\mathbf{R}^T \mathbf{C} \mathbf{R} = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_{3N}) \tag{2.35}$$

$\operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_{3N})$ is a diagonal matrix with the eigenvalues λ_i on the diagonal. The order of the λ_i is ambiguous and is chosen so that $\lambda_1 \geq \lambda_2 \geq \dots \lambda_{3N}$. Structures can be projected on the principal modes (or eigenvectors) to give the principal components p_i:

$$\mathbf{p} = \mathbf{R}^T (\mathbf{r} - \langle \mathbf{r} \rangle) \tag{2.36}$$

The eigenvectors are usually sorted by the magnitude of the corresponding eigenvalues. The eigenvector with the highest corresponding eigenvalue (the first eigenvector) corresponds to the new coordinate that shows the highest variance in the data, the second eigenvector shows the second highest variance and so on. By construction, the eigenvalues represent the variance of the projections to the corresponding eigenvectors. When applied to protein trajectories, it is found that the eigenvalues decrease rapidly (see figure 2.5), indicating that the majority of the motion of the protein is represented by a small number of eigenvectors (133).

Figure 2.4: Simple example of the results of a PCA: The two-dimensional point cloud has a strong correlation between the x and the y coordinate (in this example it is given by $(x_1, x_1 + x_2)$, where x_1 and x_2 both are Gaussian distributions). A PCA rotates the coordinate system so that the coordinates are noncorrelated.

2.3.2 Functional mode analysis

The first few eigenvectors found by a PCA of MD trajectories describe large scale correlated motions of the protein. In many cases, these large scale motions are biologically relevant. However, not all important collective modes directly correspond to PCA eigenvectors. Functional mode analysis (FMA) (134, 135) is a method aimed at finding collective modes that best correlate with a function f which is assumed to depend on the structure $\mathbf{x}$ of the system under consideration. In mathematical terms, FMA tries to minimise the residuals ϵ in the regression problem

$$\mathbf{f} = \mathbf{X}\beta + \epsilon \tag{2.37}$$

where $\mathbf{f}$ is the vector containing the functional values, $\mathbf{X}$ is the Cartesian coordi-
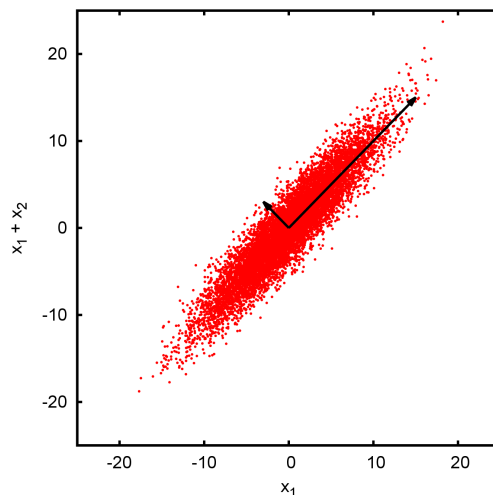
nate matrix containing the corresponding structures and β is a one-dimensional set of coefficients, that represents a collective mode of fluctuation that best captures the fluctuations in f.

2.3.3 Partial least squares regression and its application in discrimination analysis

Partial least squares regression (PLS) can be used to find a linear model to calculate an external parameter from protein structures as for example in FMA. By defining a label of which structures belongs to which class (in this case -1 denoting structures from unbound ensembles and $+1$ denoting structures from bound ones) as this external parameter, PLS can be used to calculate a model which describes differences between these two classes of structures provided such a difference exists. The resulting linear model yields a difference vector similar to a PCA eigenvector.

Figure 2.5: Example of an "eigenvalue spectrum".

If a structural difference between the classes exist, the projection of structures onto this difference vector will make it possible to assign a structure to one or the other class. If it is not possible to completely distinguish structures belonging to the two different classes, the model will still produce the best possible distinction, allowing quantification of the remaining overlap between bound and unbound ensembles. For this, both ensembles are projected onto the difference vector and histograms of the projections are calculated (fig. 3.6).

The PLS-DA algorithm used in this study produces a model that maximises the difference of the projection of two structures from different classes (bound vs. unbound) while minimising the difference between structures from the same class. Consequently, if more than one structural mode can be used to distinguish the two classes, the resulting model will not necessarily represent both of them, especially if one would result in stronger variation within the classes. While the method can be used to determine whether or not a full distinction between bound and unbound ensembles can be found, additional steps are necessary to fully characterise the structural differences. For this, PLS-DA was performed on sub-groups of atoms (i.e. the backbone as well as each residue including side-chain individually) after fitting of the ensemble on the backbone atoms.

Helland's Algorithm (136) as implemented by Denham (137) was used to perform the partial least squares discrimination analysis (PLS-DA) on the simulation ensembles. PLS performs a regression on a basis that is optimised to correlate with the external
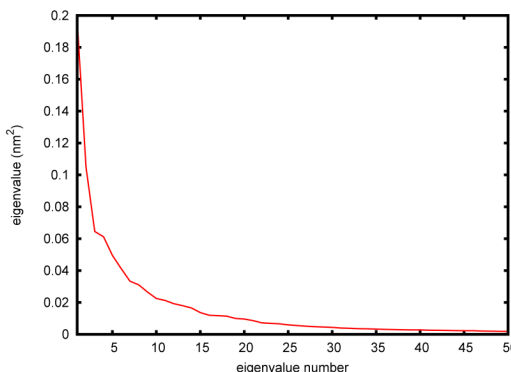
parameter. Choosing a high dimensional basis generally improves the quality of the model on the training data but can decrease its predictive power due to overfitting. For this, the combined structures of the bound and unbound ensemble were divided into a model building set (containing half of both ensembles) and a test set (containing the other half of each ensemble). Comparing model quality for both training and test set (figure 2.6) shows both correlations to reach a plateau for > 5 dimensions and no overfitting effects, so a ten dimensional basis was used in all PLS-DA calculations.
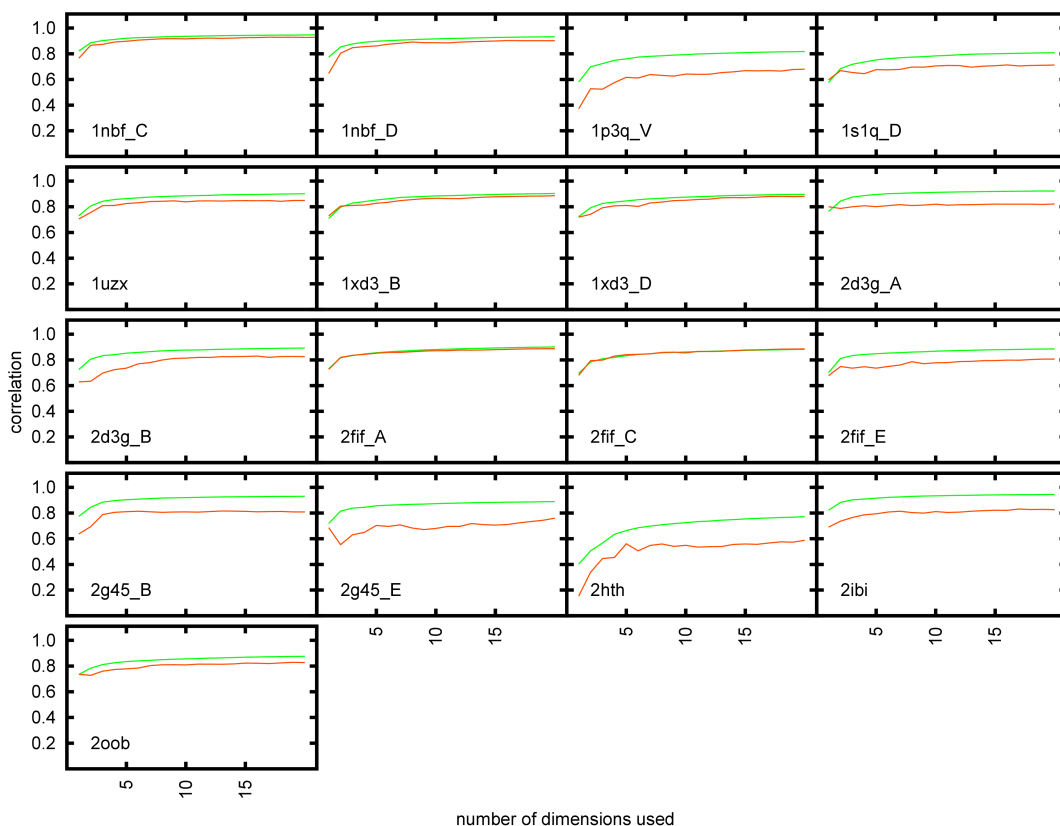
Figure 2.6: Cross correlation test of PLS-DA models. Correlation between target and model for training (green) and test (orange) set for PLS-DA between unbound and bound ensembles based on backbone atoms of residues 1-70 evaluated for different basis dimensionality.

For comparison, the projections of structures from both ensembles onto the difference vector found by PLS-DA were sorted into the same set of 100 bins spanning their combined range. The overlap of one ensemble by the other is defined as the normalised sum of the products of the number of structures for each bin. Coverage of one ensemble by another is defined as the fraction of structures from the first ensemble in bins containing a minimum number (50) of structures from the other ensemble.

The stationary bootstrap algorithm 2.3.6 was used to estimate the uncertainty of

overlaps and coverage.

2.3.4 Clustering

Clustering is a well established machine learning method which aims to automatically partition a large number N of multi-dimensional points into a small number k of classes. One of the best known clustering algorithms is k-means clustering (138). In k-means, each cluster is defined by its mean point and each data point is assigned to the cluster whose mean is nearest to the point. While finding clusters that perfectly fulfil this description is computationally expensive (NP-complete), iterative algorithms to find locally optimal solutions are commonly used (139).
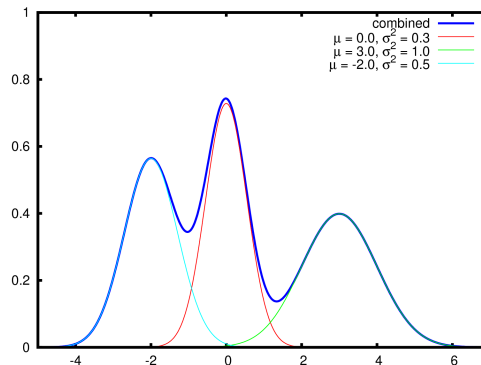
The clusters generated by k-means are usually of similar size, which is not always the best partition of the data. A more sophisticated approach is the use of Gaussian Mixture models (GMM) optimised by the Expectation-maximisation Algorithm (140). Here, the clusters are defined as multivariate Gaussian distributions

$$p(x_i) = \sum_{j=1}^{k} \frac{1}{\sigma_{i,j}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_{i,j}}{\sigma_{i,j}}\right)^2} \qquad i = 1 \ldots N \qquad (2.38)$$

where $i = 1 \ldots N$ are the dimensions of the data points and $j = 1 \ldots k$ are the k different gaussian distributions (which correspond to k "clusters") each with a mean value $\mu_{i,j}$ and a standard deviation $\sigma_{i,j}$ for each dimension. As the Gaussian distributions tend to overlap (figure 2.7), each data point can be assigned a set of probabilities as to which of the underlying distributions of the GMM it belongs to. Usually, this probability is near zero for all but one of the distributions.

Figure 2.7: Gaussian Mixture Model. The resulting model and the three underlying Gaussian distributions are shown in one dimension.

2.3.5 Comparison of structural ensembles

A necessary condition for the conformational selection binding model (section 1.1.1), and one which is often considered as a strong indication of the applicability of this model, is the existence of conformations of the protein that are observed in the complex in the absence of a binding partner (12). In this work, ensembles of the protein both in complex and free in solution have been considered, so the question of whether *the* bound conformation can be observed in the unbound state is replaced by the more complicated task to compare bound and unbound ensembles. Due to the high dimensionality, the probability of finding the same conformation in two simulation ensembles - even if both

simulate the same system under the same conditions just with slightly different starting conditions - is astronomically small.

2.3.5.1 Histogram comparison methods

After reduction of the dimensionality (e.g. by projection on a PLS-DA difference mode or PCA eigenmode), the histograms of the ensembles are compared. This can be done using different similarity measures, two of which have been used here, the overlap

$$o(\mathbf{h_A}, \mathbf{h_B}) = \sum_i \frac{h_{A,i} \cdot h_{B,i}}{N_A N_B} \qquad (2.39)$$

where $h_{A,i}$ and $h_{B,i}$ are the counts of the i-th histogram bin for the corresponding ensemble. N_A and N_B are normalisation factors

$$N_{A/B} = \sqrt{\sum_i (h_{A/B,i} \cdot h_{A/B,i})} \qquad (2.40)$$

The overlap of two histograms is 1 if both histograms have the same (up to a constant factor) count in each bin, i.e. are identical.

It is zero if they are completely distinct, i.e. if each bin that has a non-zero count in one histogram is empty in the other and vice versa.

Another measure is the *coverage* of one histogram by the other

Figure 2.8: Histogram comparisons. In graph **A**, the histograms are completely distinct, both overlap and coverage are zero. In graph **B**, the red histogram is completely covered by the blue one, resulting in a coverage of 1.0. Both the coverage of the blue histogram by the red one and the overlap of the two histograms is below 1.0.

$$o_t(\mathbf{h_A}, \mathbf{h_B}) = \sum_i \frac{h_{A,i} \cdot H(h_{B,i} - t)}{\sum_i h_{A,i}} \qquad (2.41)$$

where $H(x)$ is the Heaviside function, which is 1 for any argument greater than 0 and 0 everywhere else. t is a threshold which can be set to $t > 0$ to remove the influence of low-populated bins that can be the result of noise. In contrast to the overlap, the coverage is not commutative. It is equal to 1 if all bins containing a non-zero count for histogram A contain a count above the threshold for histogram B. It is 0 if the overlap of the histograms is zero.

2.3.6   Bootstrap uncertainty estimation

Bootstrap methods (141) are used to determine the uncertainty of statistical measurements calculated from finite samples. To determine the variation of these values on the data sample used for calculation, new ensembles are constructed by choosing random subsets of the original data and performing the same analysis on them. Classical bootstrapping methods assume that the elements of the original sample are uncorrelated. This assumption is not fulfilled for samples based on MD trajectories. An alternative approach, the stationary bootstrapping method (142) was used to determine uncertainties in these cases (e.g. in section 3.2.2). In this approach, not individual data points but series of points are selected from the original data set.

3

Ubiquitin dynamics in complexes[1]

Thus far most studies have focused on static snapshots of ubiquitin complexes in comparison to structures from solution ensemble of unbound ubiquitin. Here, based on several experimental structures of ubiquitin in different complexes (64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74) depicted in figure 1.3, we have performed and analysed MD simulations of ubiquitin interacting with different binding partners, thereby taking into account the flexibility the proteins display in the bound state. It has been shown (144) that MD simulations of unbound ubiquitin agree quantitatively with solution NMR data.

While structural information is available for several protein complexes (a few examples are listed in table 1.1), fundamental properties of complexes (first and foremost the binding affinity) cannot be deduced from just one structure, but are based on ensemble properties. I employed molecular dynamics (MD) simulations (see section 2.1) to investigate how binding influences the dynamic behaviour of ubiquitin. In particular, I identified differences between bound and unbound ubiquitin ensembles, characterising both restrictions and extensions of the accessible conformational space.

3.1   Simulation setup

From the Protein Data Bank (PDB,(63)), eleven structures of ubiquitin in complex with a binding partner and two structures of unbound ubiquitin were selected (see table 1.1 for PDB codes and references). In addition, one X-ray structure of an additional complex (*dsk2*) was supplied Stefan Becker from the department of NMR-based Structural Biology of the Max Planck Institute for Biophysical Chemistry. To avoid unspecific interactions, structures containing more than one complex were separated before simulation. Simulations were performed using GROMACS 4 (145). In accordance with recent evaluations of simulation setups ((144) and (146)) the ffamber port (147) of

---

[1]The results presented in this chapter have already been published by the same author in (143).

the amber99sb force field (88), particle-mesh Ewald electrostatics (103, 104) were employed with fourth order interpolation, a maximum grid spacing of 0.12nm and a cutoff of 0.9nm. Water was modelled using the SPC/E water model (97). A twin-range van der Waals cut-off (0.9/1.4 nm) was used. Both protein and solvent where separately held at a temperature of 300K using the v-rescale algorithm (108) ($\tau_T = 0.1$ps) and pressure coupled at 1 bar using the Berendsen algorithm (105) ($\tau_p = 1$ps). A 4fs time step was achieved by using Lincs bond constraints (114), SETTLE (113) constraints on water and virtual sites (112). After a steepest descent energy minimisation and a 1ns equilibration using position restraints on the protein, 10 production runs of 100ns each were performed for each ensemble, using random starting velocities. Simulation snapshots were taken every 10ps for analysis. For each simulation of bound ubiquitin, an unbound control simulation from the same starting structure of ubiquitin was performed without the binding partner. To allow for relaxation of structural differences, the first 10ns of these unbound simulations were not included in the analysis. An unbound reference ensemble was created from simulation trajectories based on the unbound X-ray structures 1UBI and 1UBQ and these unbound control trajectories. Ensembles based on similar structures (i.e. from starting structures from the same PDB entry) were not used in comparisons with either bound or control ensembles.

3.2 Results

In total, Eighteen structures of ubiquitin in complex with twelve different binding partners were simulated both in the presence of the binding partner (bound) and in its absence (control). Additional simulations starting from two X-ray structures without binding partner (1UBI (33) and 1UBQ (32)) were conducted for comparison.

3.2.1 Conformational overlap and restriction observed in the main modes of ubiquitin backbone dynamics

To investigate the effect of binding on the backbone dynamics of ubiquitin, a principal component analysis (PCA) of the backbone atoms of residues 1-70 of the ubiquitin chain was performed. The first eigenvector corresponds to the *pincer mode* (figure 3.2, previously described in (12)), that has direct influence on the geometry of the "hydrophobic patch", a group of three hydrophobic residues (Leu 8, Ile 44 and Val 70) that are involved in most binding interfaces of ubiquitin with other proteins. A 1μs simulation of unbound ubiquitin (figure 3.3 "1ubi") spans a conformational space similar to that covered by a large number of known experimental structures from both X-ray and NMR experiments (figure 3.1).

Like the unbound simulation ensemble, also simulations of bound ubiquitin show considerable conformational variety and in fact show a conformational entropy similar to unbound simulations (figure 3.4, estimated according to (148)).

However, while the dynamics of bound ubiquitin ensembles are considerable, specific restrictions can be observed in most of the 12 complexes when considering the main
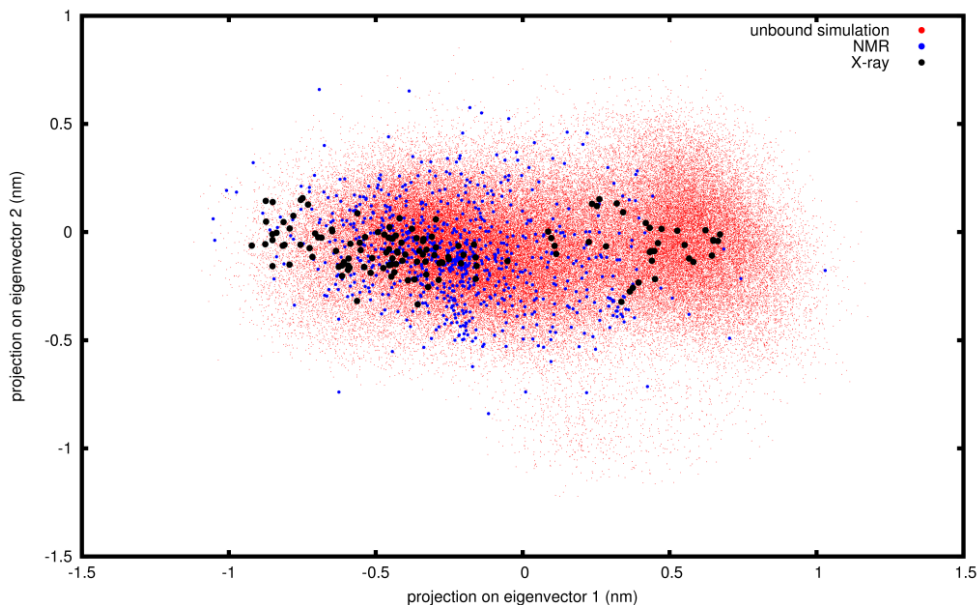
Figure 3.1: Simulation ensembles cover the same conformational space as known experimental structures. PCA projection of unbound MD simulation (starting structure 1ubi, red and a collection of experimental X-ray (black, 139 structures from 63 different PDB entries) and NMR (blue, 783 structures from 35 different PDB entries) structures.

backbone dynamic modes (figure 3.3). All bound trajectories sample a subspace of that spanned by the unbound trajectory. The first two eigenvectors displayed here cover about 30% of the total variance, and are the only ones for which significant differences between bound and unbound ensembles could be observed.

In all but one of the bound ensembles, the free energy profile along the "pincer mode" appears to have changed to shift the equilibrium towards either side of the conformational range (figure 3.3). While in most cases the shift is partial and most of the conformational space still is sampled (albeit with a lower probability on one side), some trajectories can be described as purely "open" (the ensembles based on the PDB structures 1xd3 and 2fif) or "closed" (ensembles based on PDB structures 1nbf and 2ibi). Besides the obvious exception of the ensemble 1ubi based on an unbound ubiquitin structure, only one ensemble of bound ubiquitin (2hth) shows a distribution very similar to the unbound reference ensemble and therefore does not indicate restriction of the ubiquitin dynamics in the complex.

Figure 3.5 shows a possible explanation for the restriction in both the open and closed states in two of the complexes. Ubiquitin bound to HAUSP (the binding partner in complex 1nbf) resides in a cavity that restricts its conformation in the closed state. In the open conformation, clashes would occur between residues Leu-8 and Thr-9 of ubiquitin and Ser-353 and Met-407 of HAUSP. In the complex of ubiquitin and UCH-L3 (complex 1xd3), residues Leu-8 and Thr-9 reside in a cavity of UCH-L3 when ubiquitin
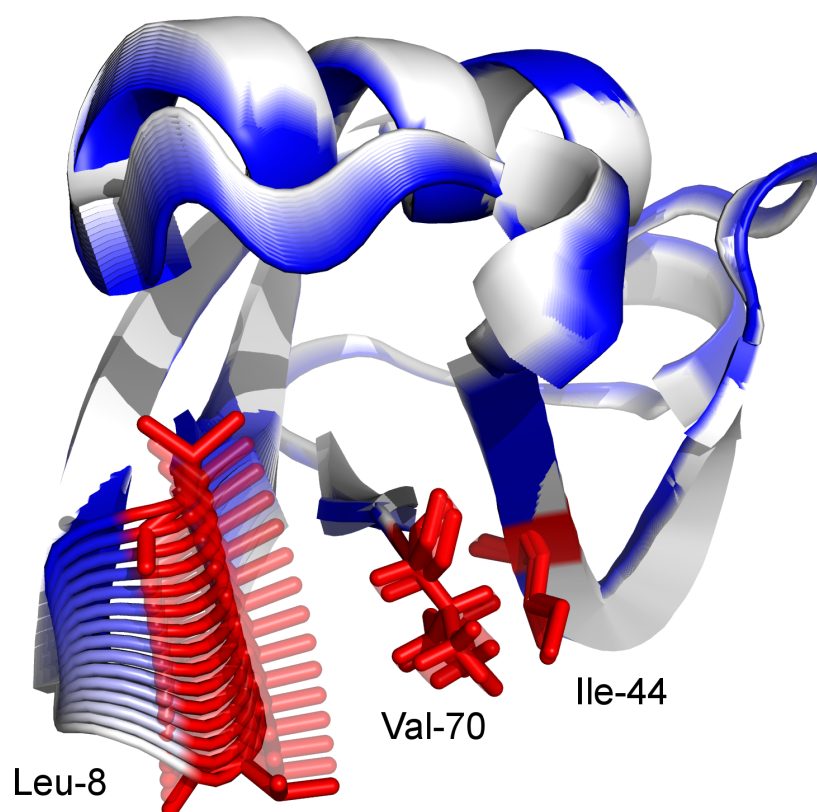
Figure 3.2: Visualisation of the first PCA eigenvector. It corresponds to *pincer mode* already described in (12). The residues of the hydrophobic patch (Leu8, Ile44 and Val70) are marked in red.
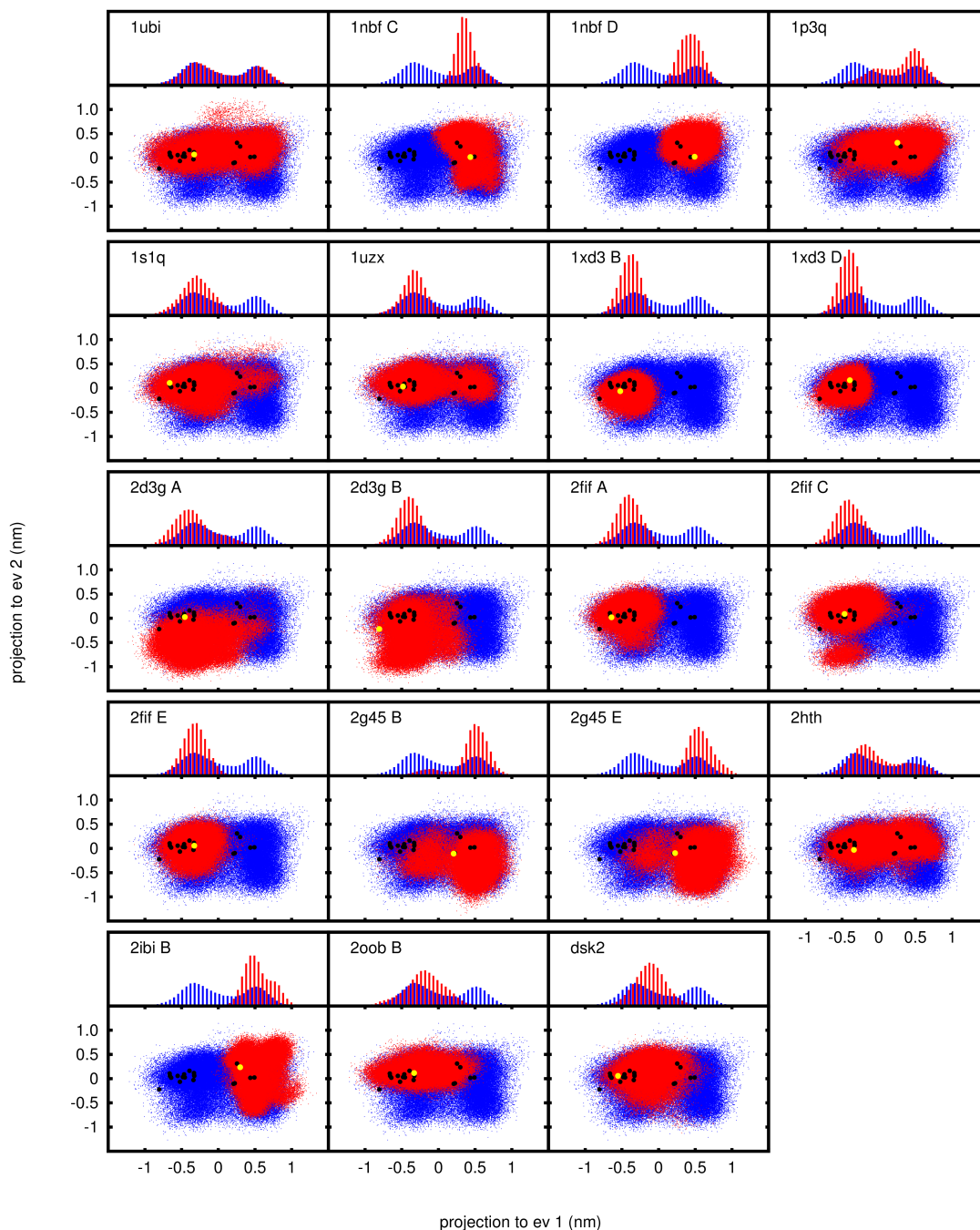
Figure 3.3: PCA results. Projection to the first two PCA-eigenvectors based on the backbone of residues 1-70 of all simulated ensembles. For comparison, the combination of unbound reference ensembles is also plotted in blue. The original X-ray structures are marked in yellow, all others in black. Histograms for the projection on the first eigenvectors are plotted above the corresponding plots. PDB codes for the starting structures of the simulations are in the upper left corner. Capital letters denote the chain identifier.

Figure 3.4: Bound ensembles show significant structural dynamics. Conformational entropy observed in unbound (blue) and bound (red) simulation ensembles estimated according to the Schlitter formula (148).

is in the open conformation. In the closed conformation, a clash between these residues and Leu-220 of the binding partner would occur which precludes these conformations.

The C-terminal tail of ubiquitin, comprising residues 71-76, shows high flexibility in the unbound and most of the bound ensembles to a degree that some parts of it are fully resolved only in four of the eleven experimental structures used for simulation setup (PDB codes 1nbf, 1s1q, 1ubi and 2g45) with three experimental structures (PDB codes 1uzx, 1xd3 and 2ibi) missing only the last residue. Four of these structures (1nbf, 1xd3, 2g45 and 2ibi) are the only ones in this study that show a significantly stronger restriction of dynamics if the C-terminal residues are included in the analysis. Besides this, the dynamic behaviour of the ubiquitin tail seems to be rather unstructured. Hence, like in other studies (12, 78) we focus on the analysis of ubiquitin dynamics to residues 1-70 as we have done in the PCA and will do in the following analysis, where inclusion of the C-terminal residues also does not qualitatively change the results while significantly increasing estimated uncertainties.
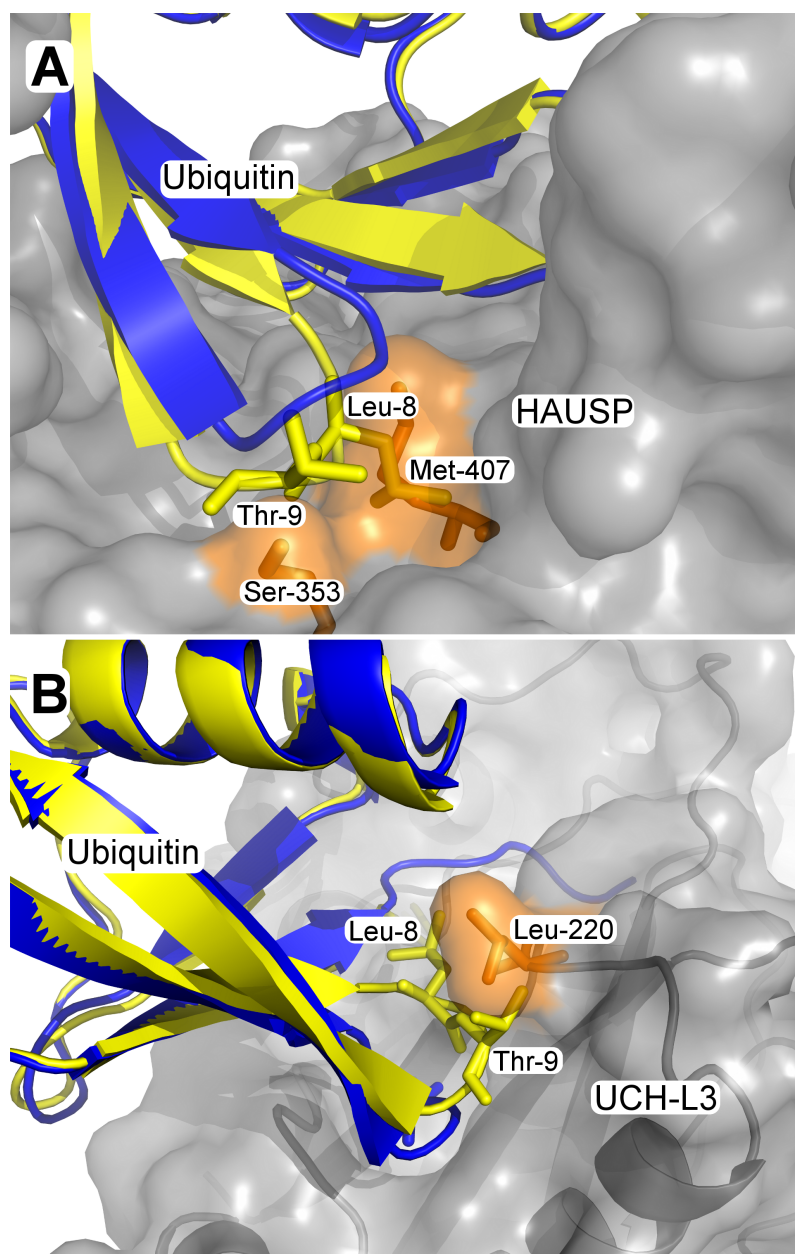
3.2.2 Differences between bound and unbound conformational ensembles as observed using Partial Least Squares Discrimination Analysis (PLS-DA)

The principal component analysis indicates conformational overlap between bound and unbound ensembles on the level of the dominant collective backbone degrees of freedom. However, PCA as a method is not aimed at discrimination, especially if the amplitude of the differences is small compared to the variation within the ensembles. It is well possible that differences between the ensembles on a more local level are not detected by PCA. To determine differences between multidimensional ensembles, partial least

Figure 3.5: Steric clashes restricting *pincer mode* dynamics in the complex. Detail from the X-ray structure (A) 1nbf (ubiquitin bound to HAUSP) and (B) 1xd3 (ubiquitin bound to UCH-L3). For each structure, the compatible ubiquitin structure is shown in blue, while an incompatible structure that has been fitted to the same position is shown in yellow. Clashes with the binding partner are marked in orange.

squares discrimination analysis (PLS-DA, cf. Materials and Methods) has been found to be more effective than PCA (149).

Indeed, using this method, models can be found to almost completely distinguish some of the bound ensembles from the unbound reference ensemble The magnitude of these differences is however significantly smaller than that of the main fluctuation modes of ubiquitin (compare length-scales in figures 3.3 and 3.6).
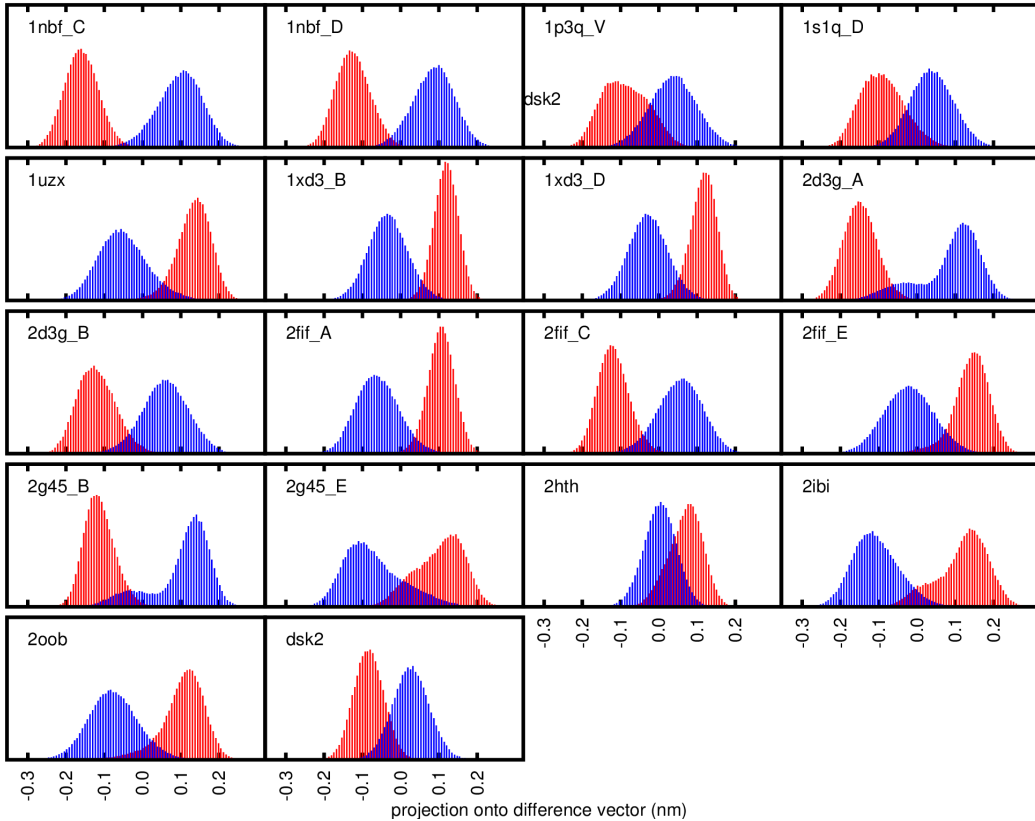

Figure 3.6: PLS-DA results on backbone atoms of residues 1-70. Different bound ensembles (red) and the unbound reference ensemble (blue) have been projected onto the difference vector between these ensembles as determined by PLS-DA.

PLS-DA distinguishes between ensembles both on a global as well as on a local level. Even the systematic difference between two ensembles in e.g. a single side chain rotamer will result in a zero overlap.

While both bound and unbound control ensembles are fully covered by the unbound reference ensemble along the main mode of ubiquitin dynamics (figure 3.7 A), the coverage of the bound ensembles after PLS-DA on the backbone atoms of residues 1-70 (figure 3.7 B) is found to be significantly lower. When also considering all non-hydrogen side chain atoms (figure 3.7 C), several bound ensembles are no longer covered by the unbound reference ensemble. To validate the significance of the observed differences, the

same method has been applied to calculate the coverage of unbound control ensembles by the unbound reference ensemble. It was found to be significantly higher (i.e. almost full), as expected.

The observed differences correlate well (r = -0.92) with number of ubiquitin atoms involved in binding (i.e. with a distance of less than 0.6nm from the binding partner, figure 3.7 D). Hence a more extensive binding interface correlates with more significant differences to the unbound state.

3.2.3 Local conformational differences on the residue level can be observed using PLS-DA

To localise differences between bound and unbound ensembles, individual PLS-DA calculations were performed on the conformations of each residue (including side chains) of ubiquitin separately after superimposing the backbone of the whole chain.

Only a small number of residues for each complex ensemble show an overlap with the unbound reference ensemble which is significantly below 1.0 and none of them shows an overlap below 0.2, Most of the unbound control ensembles show almost complete (1.0) overlap with the reference ensemble. The observed differences due to binding interactions are local, as all of the residues found to change their conformation are located near the binding partner (figure 3.8).

Again, in none of the cases, a complete distinction between bound and unbound ensembles could be found. Even for the residue displaying the smallest overlap between bound and unbound ensembles (residue His68 in ensemble 1nbf chain C) a small fraction of bound structures can be found in the same conformational region as the unbound ones (figure 3.9).
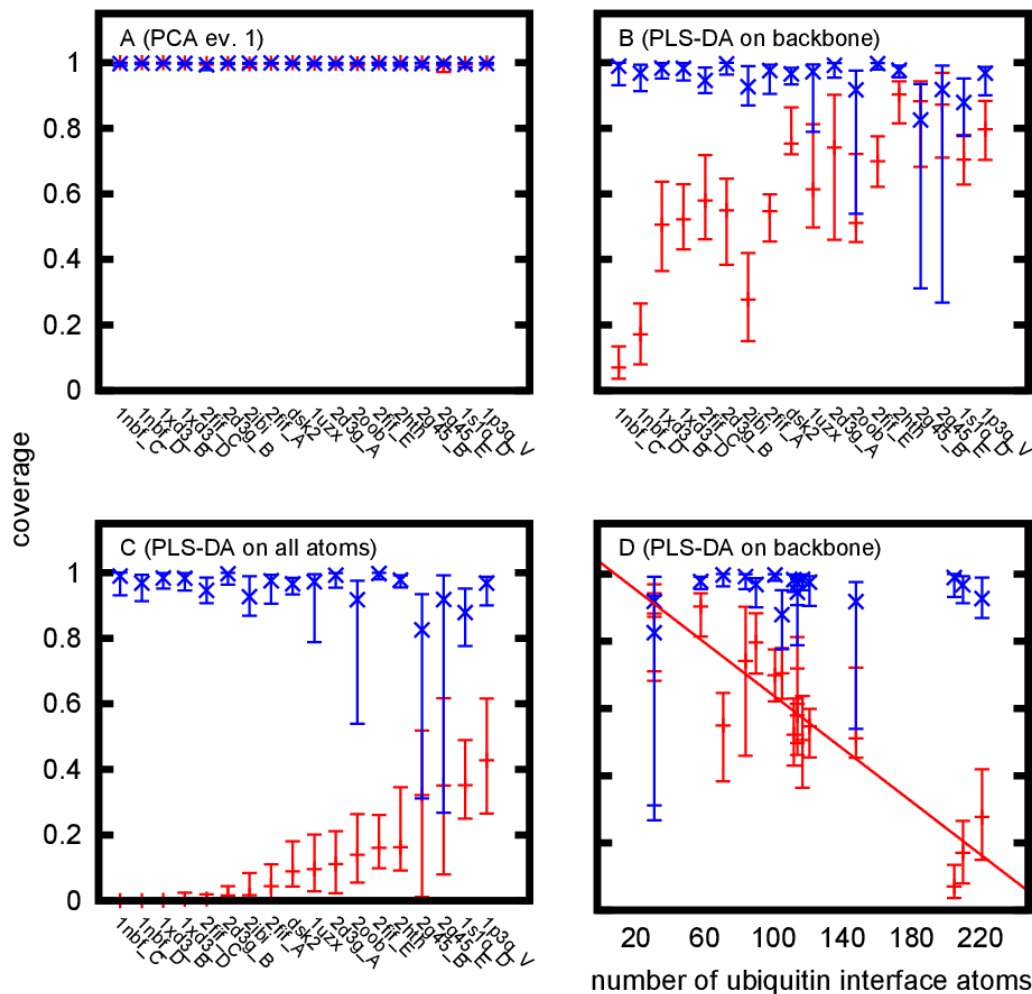
Figure 3.7: Coverage of different ensembles by the unbound reference ensemble. The histogram-coverage of bound ensembles (red) compared to coverage of unbound control ensembles (blue) after projection of the structures onto the first PCA-eigenvector (figure 3.3) of backbone atoms of residues 1-70 (A), the PLS-DA difference vector of backbone atoms of residues 1-70 (B and D), and the PLS-DA difference vector of all non-hydrogen atoms of residues 1-70 (C). Ensembles in A-C have been sorted according to the coverages displayed in C. Uncertainties have been determined using the stationary bootstrap method.
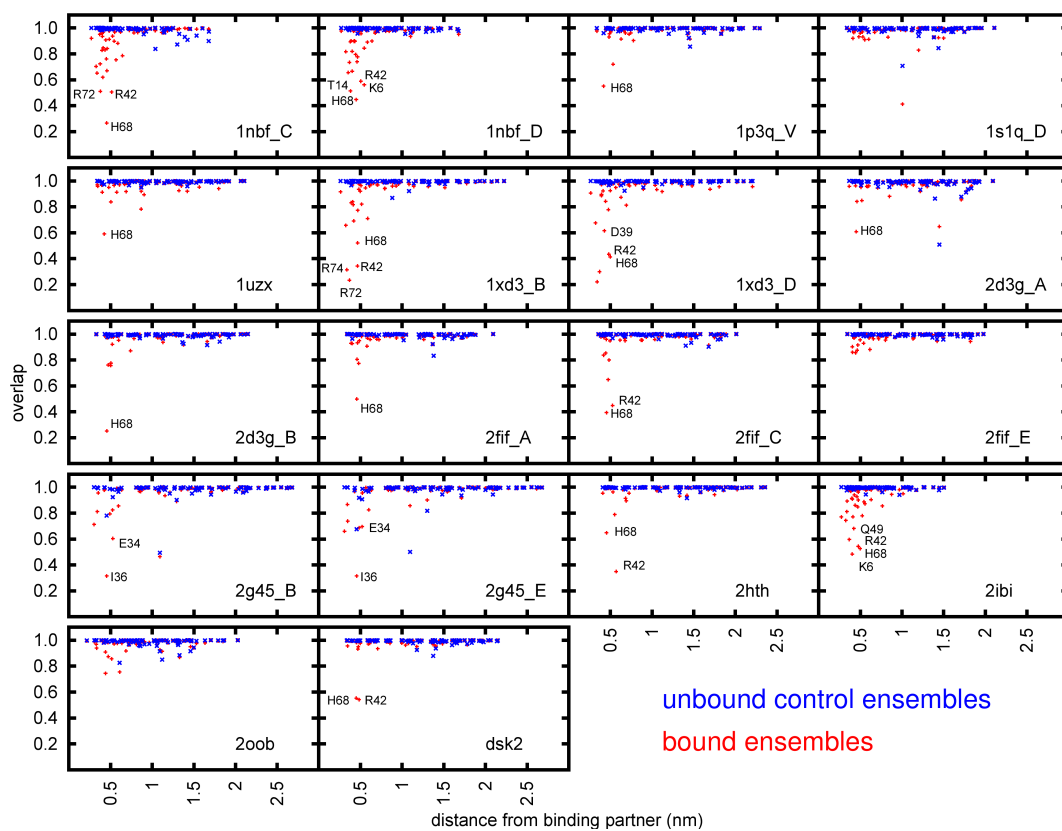
Figure 3.8: Overlap between bound (red) and unbound control (blue) ensembles. Overlap has been calculated with the unbound reference ensemble after projection to the difference vector found by PLS-DA on single residues after superimposing to the backbone and plotted versus distance from the binding partner. Residues displaying a significant difference in the bound ensemble are labelled.

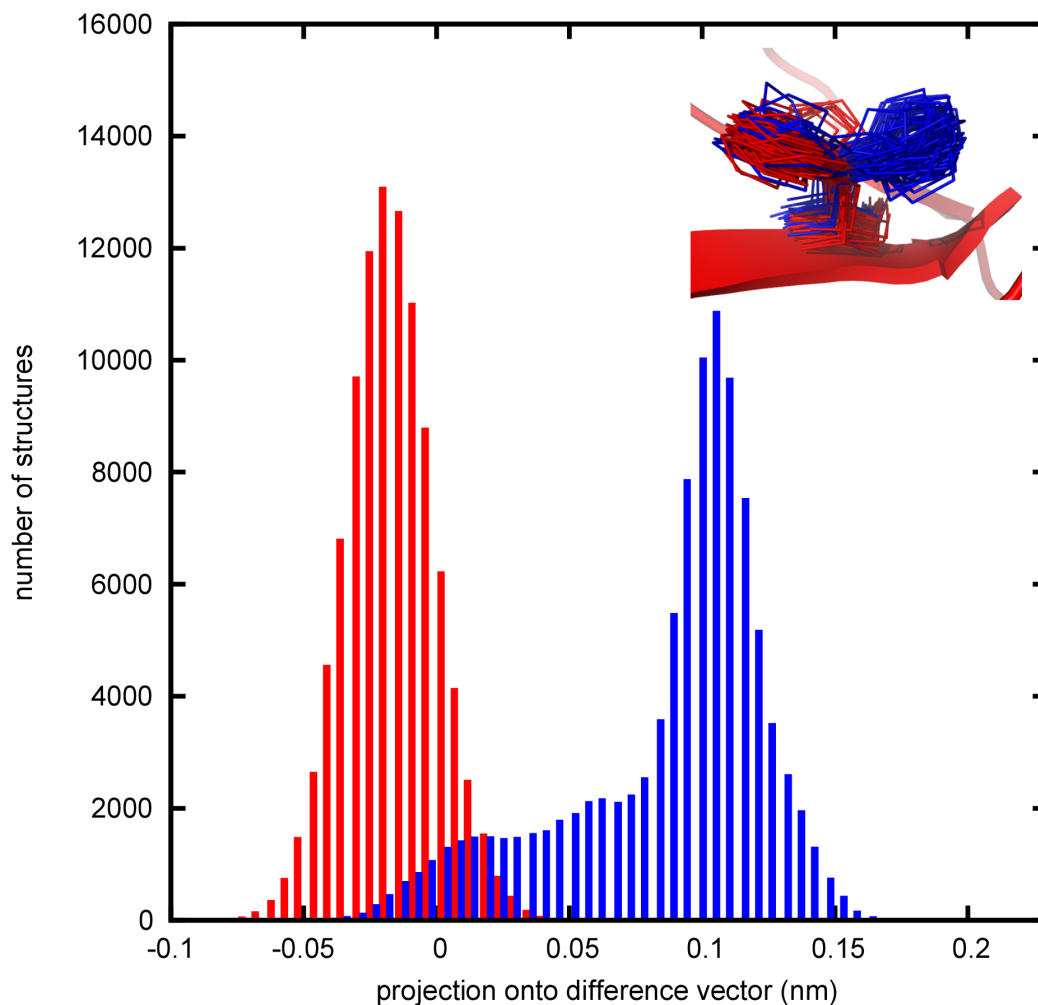**Figure 3.9: Example of a difference found in PLS-DA and its structural origin.**
Histogram of the projection of bound (red) and unbound (blue) ensemble onto the difference vector found by PLS-DA for residue His68 of ensemble 1nbf chain C. Out of all 11 complexes studied, this residue shows the smallest overlap between bound and unbound ensembles. The inset shows the corresponding structures from the simulation ensembles.

4

Influence of core mutations of ubiquitin on dynamic modes.

The results in this chapter have been reached in collaboration with Monika Sharma, Daniel Seeliger and Servaas Michielssens who have developed the protocol and performed the calculations described in section 4.2.

In the previous chapter it was shown that the *pincer more* of ubiquitin plays a crucial role in its ability to form a multitude of complexes. Different ubiquitin binding partners require ubiquitin to specifically be in one of the two main states along this mode. This leads to the idea that restricting the conformational preference of ubiquitin to one of these states should alter its binding specificity, which would allow to alter the binding behaviour by tuning the *dynamics* rather than the structure of the *interface*. In the following chapter, I describe a protocol to identify mutations of ubiquitin that shift its conformational preference along the pincer mode. It was used to identify several candidate mutations of *core residues* of ubiquitin that prefer either the open or the closed state along the pincer mode. Finally the hypothesis that these mutations specifically influence the binding behaviour of ubiquitin was tested by calculating changes in binding affinity of ubiquitin to several binding partners.

4.1   Experimental evidence of selective influence of ubiquitin core mutations on binding behaviour

Ubiquitin is a strongly conserved protein over a wide variety of organisms (150), indicating that changes in its sequence are likely to disrupt its function. The fact that mutation of surface residues can influence the binding behaviour of a protein is well established and often used to identify binding sites (for example in (151)). But it has also been observed that mutations of core residues of ubiquitin can selectively change its binding behaviour (152, 153).

Haririnia et al. (152) identified mutations (L69S, L67S) that *selectively* influence the binding behaviour even though the mutated residues are not part of the binding

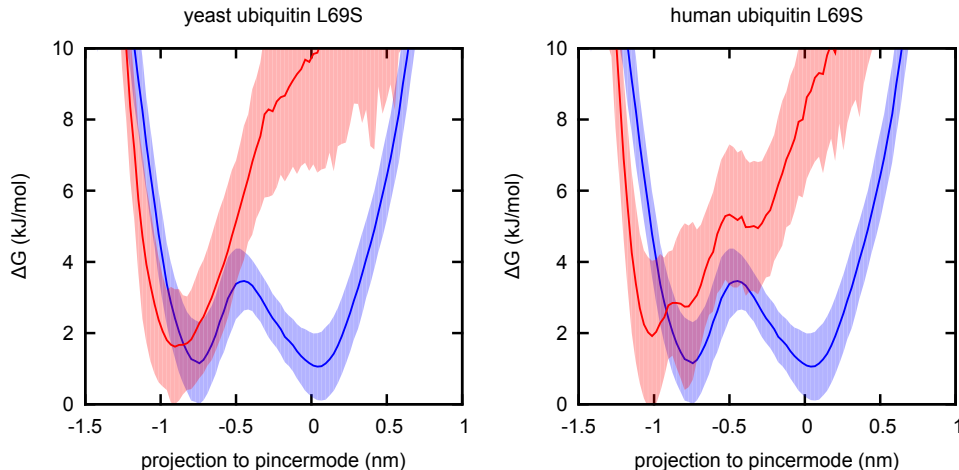# 4. INFLUENCE OF CORE MUTATIONS OF UBIQUITIN ON DYNAMIC MODES.

Figure 4.1: Free energy profiles for L69S mutants of yeast ubiquitin (red,left graph) and human ubiquitin (red, right graph) compared to that of wild type unbound ubiquitin (blue). The profiles where calculated using the Boltzman factor from simulation ensembles consisting of ten 100ns simualations. The yeast ubiquitin simulations are based on a set of NMR structures published by Haririnia et al. (PDB access code 2JWZ), while the human ubiquitin ensemble is based on structures from a wild type simulation that have been modified using the WHATIF software.

interface. Simulation of L69S mutations on both yeast and human ubiquitin show a shift to the closed conformation along the *pincer mode* (figure 4.1).

As seen in section 3.2.1, complex formation with specific binding partners can also significantly shift the population of the ubiquitin ensemble along the *pincer mode*. The binding partners investigated by Haririnia et al. were Rpn10 and S5a (both of which did not bind the mutated ubiquitin) as well as Rad23 and hHR24A (which did interact with the mutants). In terms of ubiquitin binding motifs, this means that the mutant does not bind to the ubiquitin-interacting motif (UIM) but does still interact with the Ubiquitin-associated domain (UBA).

In chapter 3, complexes have been investigated both including a UIM (2d3g) and a UBA (2oob) binding motif. It has been seen that ubiquitin bound to a UIM ubiquitin binding motif (as in the complex 2d3g) strongly prefers open conformations. This mismatch between the L69S mutant mainly occupying the closed state of the *pincer mode* and the ubiquitin bound to an UIM motif in the 2d3g ensemble preferring the open state could explain the inability of this mutant to bind another binding partner containing the same binding motif (it should be noted that 2oob, the complex ensemble containing a UBA motif, does also prefer the open state, and L67S, the other mutant found to show the same selectivity as L69S does not show preference for one of the states).

In a more recent study, Zhang et al. (153) engineered a number of ubiquitin
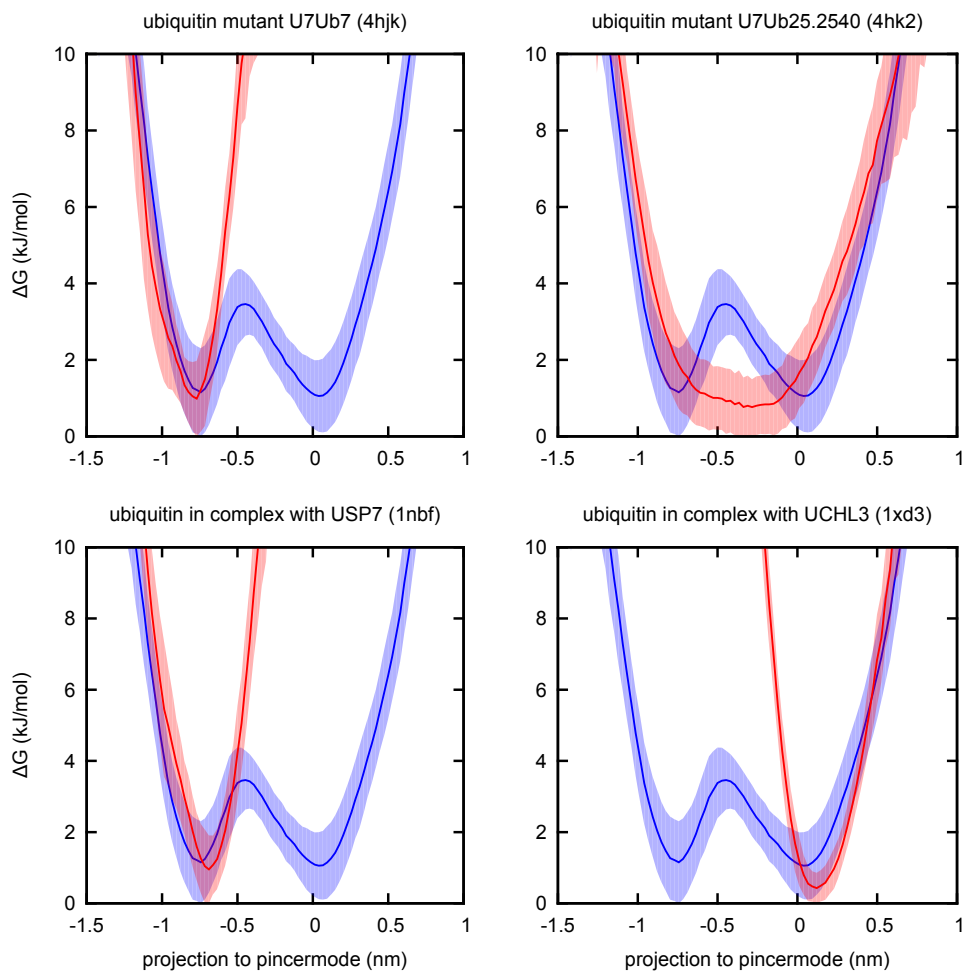
Figure 4.2: Free energy profiles for selective mutants from the work of Zhang et al. The mutants "U7Ub7" and "U7Ub25.2540" (upper row, red) have been optimised for increased binding affinity to USP7 (lower left, red) while they do not bind UCHL3 (lower right, red). The profiles have been calculated using the Boltzman equation from histograms of 10x100ns (4hjk), 4x10x100ns (4hk2) or 2x10x100ns (1nbf,1xd3). For comparison, the free energy profile of unbound ubiquitin has been plotted in each panel (blue).

4. INFLUENCE OF CORE MUTATIONS OF UBIQUITIN ON DYNAMIC MODES.

mutants specifically to increase the binding affinity to ubiquitin-specific protease 7 (USP7, which is the ubiquitin binding partner in structure 1NBF). They found that the mutations that best improved binding affinity to USP7 also significantly decreased binding affinity to several other ubiquitin binding proteins including the ubiquitin carboxyl-terminal esterase L3 (UCHL3, which is the binding partner to ubiquitin in structure 1XD3). The majority of mutants found by Zhang et al. contained cystein residues in positions 7 (sometimes 8) and 69, which formed a disulfide bond.

A MD simulation based on a crystal structure of one of these mutants (figure 4.2) shows that the mutation restricts ubiquitin to the "closed" state of the *pincer mode*, which would explain the preference for USP7, which requires ubiquitin to be in the closed state over UCHL3 which requires the open state (both depicted in figure 4.2). Another selective ubiquitin mutant found by Zhang et al. lacking the disulfide bond but including mutations of surface residues to optimise binding affinity does not show the same preference for the closed state.

Here, we aim to identify ubiquitin mutations that specifically influence binding behaviour without altering the binding interface. For this, we developed a fast growth thermodynamic integration/crooks Gaussian intersection (FGTI/CGI)-based protocol to calculate the difference in the change of free energy due to mutation for different states of a protein. Using this protocol, we performed an extensive scan of ubiquitin core mutants and identified a number of candidates for which one of the two states of the *pincer mode* is energetically more favourable than the other. Finally, we tested the influence of these mutations on the binding of ubiquitin to binding partners that have

Figure 4.3: **Fast growth thermodynamic integration and Crooks Gaussian intersection simulation protocol.** Two equilibrium ensembles are generated (here by MD simulations), one for the wild type (wt) protein, one for the mutated protein. From these ensembles, snapshots are selected and used to set up short (100ps) FGTI simulations, in which the wt is turned into the mutant and vice versa.

been shown in chapter 3 to clearly restrict ubiquitin to one of the two states, again using FGTI/CGI.

4.2 Determination of changes in the stability of open or closed states in ubiquitin by different mutations

In this chapter, the *pincer mode* is described as the first eigenvector of a PCA based on the 213 backbone atoms of residues 1-71 of 61 experimentally determined ubiquitin structures taken from the Protein Database PDB.

4.2.1 Restriction to open and closed states

Using the Boltzman formula 2.18, the free energy profile of unbound ubiquitin (ensemble "1ubi" from chapter 3) was calculated from a histogram of projections of ensemble-structure to this eigenvector. It possesses two minima, one corresponding to an open state, the other to a closed. Between these two states is a barrier of approximately $2\,\mathrm{kJ\,mol^{-1}}$ (figure 4.4). As this is less than the thermal energy at $300\,\mathrm{K}$ which is (approximated for an ideal gas) $3/2 \cdot N_A \cdot kT \approx 3.74\,\mathrm{kJ\,mol^{-1}}$, a simulation of the protein will frequently cross this barrier. This is undesirable in the CGI simulations, as we're trying to analise the effect of the mutation in the open and closed states individually. To restrict the protein to either the *open* or the *closed* state, a modified version of the Essential Dynamics code of GROMACS (154, 155, 156) was used to add semi-harmonic potentials (equation 4.1 or a "mirrored" version) along the collective *pincer mode* coordinate ξ (figure 4.4).

$$V(\xi) = \left\{ \begin{array}{ll} V_0 \cdot (\xi - \xi_0)^2 & \xi > \xi_0 \\ 0 & \xi \leq \xi_0 \end{array} \right. \tag{4.1}$$

While this potential is necessary to prevent the simulation from leaving its assigned state, the influence of the potential on the simulation should otherwise be as small as possible. To achieve this, a high force constant $V_0 = 1000\,\mathrm{kJ\,mol^{-1}\,nm^{-1}}$ was selected. The resulting high force acting on the system if it enters the non-zero part of the potential will reduce the time the simulation spends in this region.

4.2.2 Screening of core residue mutants using FGTI/CGI
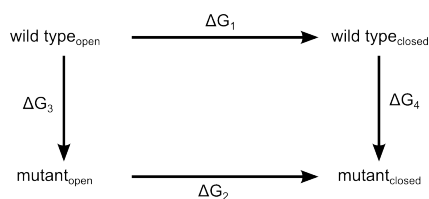
Figure 4.5: Thermodynamic cycle used to test stabilization of one over the other state of ubiquitin by a mutation.

14 hydrophobic core residues (table 4.1) of ubiquitin were chosen and mutated to the hydrophobic residues Alanine (A), Valine (V), Isoleucine (I), Leucine (L), Methionine (M), Phenylalanine (F), Tyrosine (Y) and Cysteine (C) as well as Serine (S) and Threonine (T), resulting in a total of 112 mutants.

To determine the free energy difference (ΔG) caused by a mutation, fast growth thermodynamic integration (FGTI) evaluated using the Crooks Gaussian intersection (CGI) method described in section 2.2.5 was used. For this, equi-

Figure 4.4: Potentials used to restrict ubiquitin to "closed" (A) or "open" conformations. The free energy profile of unbound ubiquitin (calculated from the "1ubi" simulation ensemble in chapter 3) is plotted in black, t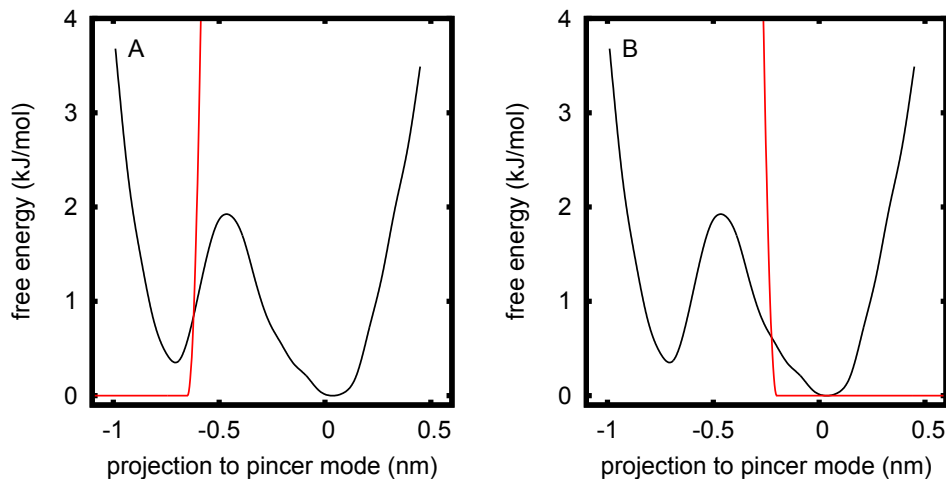he semi-harmonical potential to prevent the simulation to enter the open (A) or closed (B) state is plotted in red. Due to the high force constant, the non-zero part of the potential is almost vertical.

Table 4.1: Ubiquitin code residues mutated.

Residue number	Residue type	Short form
3	Isoleucine	I3
5	Valine	V5
13	Isoleucine	I13
15	Leucine	L15
17	Valine	V17
23	Isoleucine	I23
26	Valine	V26
30	Isoleucine	I30
36	Isoleucine	I36
43	Leucine	L43
56	Leucine	L56
61	Isoleucine	I61
67	Leucine	L67
69	Leucine	L69

librium ensembles of the system in both states, in this case wild type protein and mutant, are calculated using MD simulations similar to those described in chapter 3.

From the ensemble of unbound ubiquitin (based on the X-ray structure 1UBI) sim-

ulated in chapter 3, ten "open" and ten "closed" structures were selected and simulated using the software WHATIF (157). Each of these mutated structures was simulated for 10ns at equilibrium conditions, the ten open structures with the external potential restricting them to the open state and ten closed ones with the potential restricting them to the closed state. From these equilibrium ensembles, 100 snapshots were extracted from each the open and the closed ensemble for 100ps FGTI simulations using tools described in (158). The $\partial H/\partial \lambda$ profiles calculated in these simulations are then integrated to calculate transition energies.

Finally, transition energies are sorted into histograms and approximated by Gaussian distributions (figure 4.3). The intersection point of the Gaussian distributions for transformations in both directions (wild type to mutant and mutant to wild type) is the best approximation for the free energy difference between the two states.

FGTI/CGO can be also be used to calculate the change in thermal stability of a mutation (158), that is the change in the folding free energy of the protein due to the mutation. For this, the free energy change of mutation calculated in the folded protein is compared to that of the unfolded protein, which can be estimated from the free energy change of mutation of a GXG (Glycine-X-Glycine) tripeptide. Besides shifting the conformational population of ubiquitin to either the open or the closed state, a mutation candidate resulting from this screening should also not significantly destabilise the protein compared to the thermal stability of the wild type ($\Delta G_{unfold} = 23.6\,\mathrm{kJ\,mol^{-1}}$ according to (152)).

4.2.3 Zero-Cycles

Zero cycles are a common approach to validate methods calculating free energy differences. They are a simplified version of the thermodynamic cycles described in section 2.2.1 and like them are based on the fact that the free energy is a state property. The simplest zero cycle contains three states (figure 4.6). If the free energy differences between the three states are calculated and summed up in a way that a closed cycle is described, the sum of the free energy differences should be zero. Three zero cycles were calculated both with restrictions to the open and the closed state. With one exception, the free energy difference along a closed path of mutations was close to zero, as expected (table 4.2), and all values are within 1kcal ($4.184\,\mathrm{kJ\,mol^{-1}}$) of zero.
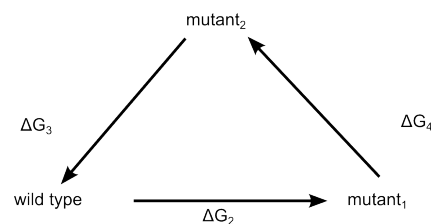

Figure 4.6: Zero cycle used to validate the free energy calculation protocol.

4.2.4 Results of the conformational screening

Of the 112 mutations tested, only in 15 cases (table 4.3 including the already known L69S mutant) a difference between the free energy differences calculated in the open

4. INFLUENCE OF CORE MUTATIONS OF UBIQUITIN ON DYNAMIC MODES.

Table 4.2: Zero cycles on unbound constrained ubiquitin with restritions to the open/closed state. All energies in kJ/mol.

Wild type residue	mutations	$\Delta\Delta G_{open}$	$\Delta\Delta G_{closed}$
V5	V→L→I→V	-0.1 ± 1.0	1.6 ± 0.8
I13	I→Y→F→I	-1.3 ± 1.2	-3.9 ± 1.2
L15	L→I→V→L	0.9 ± 0.8	-0.7 ± 0.8

and closed states was above 1kcal, and hence significant considering the accuracy of the method (figure 4.7, lower half). In addition, the free energy difference $\Delta G_{peptide}$ calculated on the GXG tripeptide were compared to the result for the preferred state to test if the mutation destabilises the protein (figure 4.7, upper half).

Table 4.3: Ubiquitin mutants showing significant difference between open and closed state. The free energy difference $\Delta G_{closed} - \Delta G_{open}$ between the mutation restricted to the open and the closed conformation indicates which of the two states is more stable. If the difference is negative, the mutants are predicted to prefer the closed conformation, if it is positive, the mutant are predicted to prefer the open conformation. The free energy differences $\Delta\Delta G_{closed}$ and $\Delta\Delta G_{open}$ are the change of binding free energy calculated for the mutant restricted to the closed and the open conformation. If both of these values are high (and positive), the mutant is significantly less stable than the wild type and might not fold correctly. For the 15 mutants in this list, umbrella sampling simulations were performed to confirm the predicted preference for one of the states. All energies in are kJ/mol.

| mutations	$\Delta G_{closed} - \Delta G_{open}$	$\Delta\Delta G_{closed}$	$\Delta\Delta G_{open}$	confirmed
I30Y	-13.1 ± 4.1	7.1 ± 2.3	20.2 ± 2.6	no
V17Y	-6.5 ± 3.9	18.8 ± 2.2	25.3 ± 2.4	no
L69T	-6.2 ± 2.1	15.4 ± 1.6	21.6 ± 1.4	yes
I30T	-5.4 ± 1.6	-0.8 ± 1.0	4.6 ± 1.0	no
L69S	-5.3 ± 2.2	18.6 ± 1.6	23.9 ± 1.4	yes
L56Y	-5.1 ± 2.5	19.4 ± 1.7	24.6 ± 1.7	no
V5L	-4.9 ± 1.3	1.2 ± 0.9	6.1 ± 1.1	yes
I36L	-4.9 ± 0.9	-0.7 ± 0.6	4.2 ± 0.8	yes
V5C	-4.4 ± 1.0	1.9 ± 0.8	6.3 ± 0.8	yes
I36F	4.5 ± 1.3	4.7 ± 0.9	0.2 ± 1.1	yes
I36S	5.2 ± 1.3	11.4 ± 1.0	6.1 ± 1.0	yes
I13F	5.5 ± 1.4	2.8 ± 1.2	-2.6 ± 1.1	yes
I36C	5.6 ± 1.0	4.9 ± 0.8	-0.7 ± 0.9	yes
I36Y	5.9 ± 1.7	3.0 ± 1.2	-2.9 ± 1.2	yes
I36A	8.6 ± 1.0	10.7 ± 1.4	2.1 ± 1.5	yes
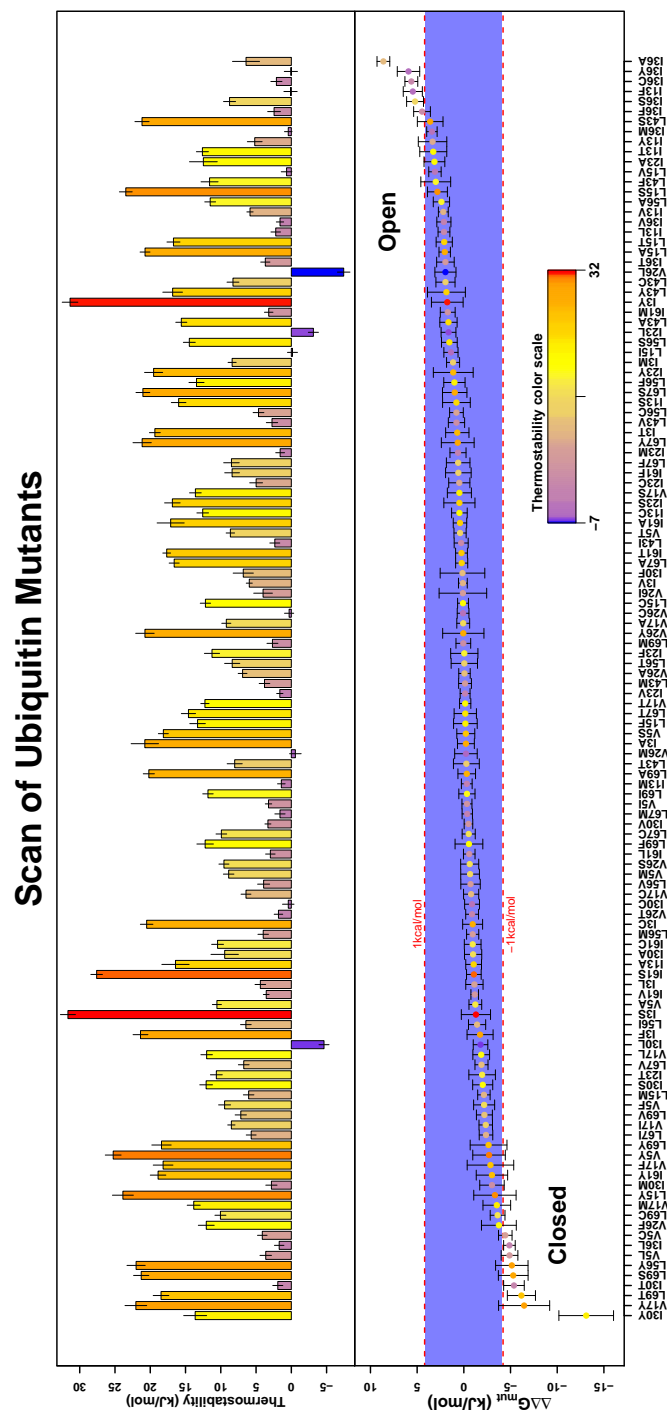
Figure 4.7: Changes in stability for ubiquitin mutants restricted to the open or closed state respectively. Plot provided by Servaas Michielssens.

4.2.5 Free energy profiles using umbrella sampling

While the FGTI/CGI protocol described above allows a wide screening of mutations at (comparatively) low computational cost, an independent validation of free energy profile would be desirable. For this, umbrella sampling simulations (section 2.2.6) were performed on the 15 most promising mutations (listed in table 4.3). 20 umbrella potentials were defined along the *pincer mode* and eleven restricted simulations of 15ns each where performed in each window, the last 10ns of which where used for analysis. The resulting histograms where analysed usig WHAM (129, 130) and a bootstrap (142) method was used to estimate the uncertainty along the profile. As can be seen in figure 4.8, the profiles fit the FGTI/CGI result in most cases. For some mutants, particularly for I30Y, V17Y, I30T and L56Y the calculated profile does not fit the preference of the closed state that has been found by FGTI/CGI.
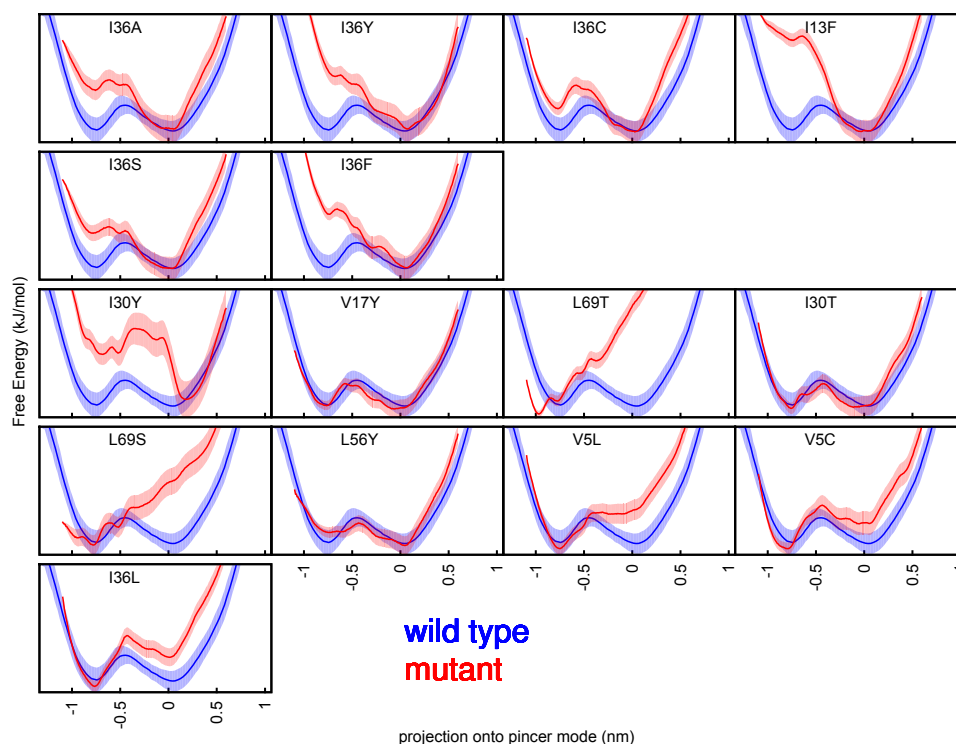

Figure 4.8: Free energy profiles for different ubiquitin mutants. Calculated using umbrella sampling.

4.3 Calculating binding free energy changes due to mutation of ubiquitin

FGTI/CGI-screening (section 4.2.4) and validation using umbrella sampling (section 4.2.5) have resulted in a list of eleven mutations, five (L69T, L69S, V5L, I36L and V5C) that shift the conformational population of unbound ubiquitin to the closed state and six (I36F, I36S, I13F, I36C, I36Y and I36A) that shift it to the open state. To validate if the binding behaviour of ubqiutin can indeed be changed by these shifts, the change in binding free energy due to some of these mutations in four characteristic complexes has been calculated in section 4.3.3.

For two reasons, the equilibrium simulations used for complexes have been chosen to be significantly longer (5x100ns instead of 10x10ns). First, the simulation system including the binding is bigger than that only consisting of ubiquitin, hence it can be assumed that more time is needed to sample the conformational space. Secondly, the goals and hence requirements of the two calculations is different. The calculations on unbound ubiquitin were a screening process that was supposed to identify valid candidates from a set of 112 candidates. Promising mutations were then validated using umbrella sampling. The calculations described in the following sections are performed on a smaller set of systems (six mutations in four complexes resulting in 24 combinations) to validate the influence on the binding behaviour. A longer equilibration, expected to lead to more accrate energy estimates, was also chosen since a further validation using umbrella sampling, as carried out for isolated ubiquitin, is not feasible for ubiquitin complexes.

4.3.1 Zero cycles of mutations in ubiquitin complexes

As already done for unbound ubiquitin in section 4.2.3, the reliability of FGTI/CGI results for ubiquitin complexes was tested using zero-cylcles.

Table 4.4: Zero cycles on ubiquitin complexes. All energies in kJ/mol.

Complex	Wild type residue	mutations	$\Delta\Delta G$
1xd3	K6	K→A→R→K	-2.5 ± 2.0
1xd3	I30	I→T→Y→I	-0.2 ± 2.1
1nbf	L69	L→S→T→L	4.1 ± 1.8
2g45	L69	L→S→T→L	-1.4 ± 1.7

Two of the four zero-cycles listed in table 4.4 are near zero within their uncertainty, and all are in the range of ± 1kcal ($4.184 \, \mathrm{kJ \, mol^{-1}}$). This suggests that that chosen protocol indeed provides a similar accuracy as obtained for isolated ubiquitin.

4.3.2 Testing calculated changes in binding free energy against experimental values

Wilkinson et al. (151) have measured the changes in binding free energy for different ubiquitin mutants to UCH-L3, which is the protein binding ubiquitin in the 1XD3 structure. To test the feasibility of the used protocol to calculate affinity changes due to mutation, calculations of these values were performed. The results are shown in figure 4.9.
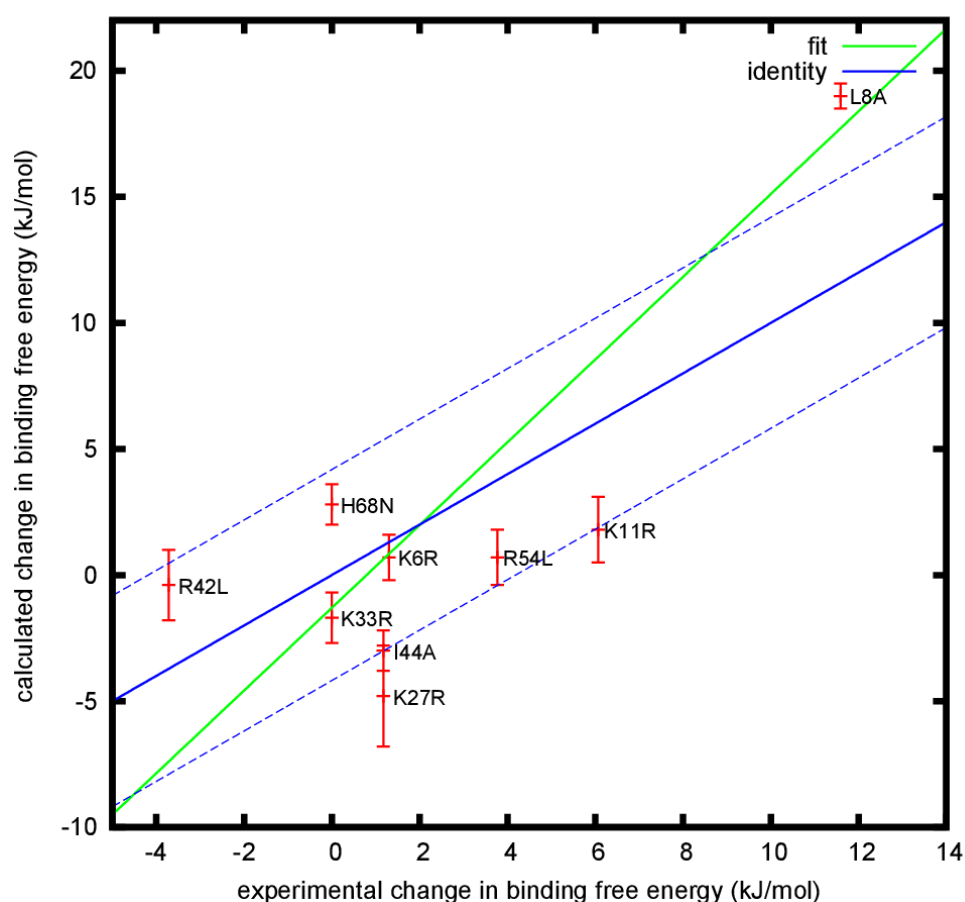
Figure 4.9: Comparison of experiental and calculated binding free energy changes. The binding free energy differences for ubiquitin binding with UCH-L3 has been measured for different mutations (151). Here, these values are compared with those calculated using the alchemical free energy method. The green line is a fit of the data points. The solid blue line is the identity. If the calculations would perfectly reproduce the experimental values, all points would lie on this line. The broken blue lines delimit a deviaton of ±1kcal/mol from experimental values.

Overall, the correlation coefficient is 0.70. The correlation coefficient for "conservative mutants" (i.e. K6R, L8A, K11R, K27R, K33R and I44A which conserve the charge

of the residue) is 0.93.

4.3.3 Calculating the effect of mutations known to prefer one pincer-mode state on binding behaviour

To test the hypothesis that the mutants identified in section 4.2 indeed affect binding selectivity by favouring a specific state in the *pincer mode*, the change in binding free energy was calculated for four complexes. Two complexes with ubiquitin predominantly in the open conformation and two with ubiquitin predominantly in the closed conformation were chosen. Six ubiquitin mutations were performed on each of these complexes, two preferring the open and four preferring the closed conformation according to 4.2. For each complex, five structures equally distributed along the *pincer mode* were selected from the equilibrium simulation (with wild type ubiquitin). Each of these structures was mutated using the software WHATIF (157) and simulated for 110ns using the GROMACS 4.5 simulation package (145). The first 10ns are not used in subsequent steps to allow the protein structure to relax into the ground state of the mutated ubiquitin. Of the remaining 100ns per trajectory, 50 snapshots (every 2ns) containing coordinates of all atoms of the simulation system (including solvent water and ions) were extracted and used to set up Fast Growth Thermal Integration (FGTI) simulation in which the mutated residue was changed back to the wild type residue according to a setup adapted from (158). Similarly, 300 structures were selected from the equilibrium simulation of wild type ubiquitin in the complex and FGTI simulations were performed in which the wild type protein was turned into the mutant.

Table 4.5: **Binding free energy changes due to ubiquitin mutation.**

		closed complexes		open complexes	
		1nbf	2g45	1xd3	dsk2
closed mutants	V5L	-2.6 ± 0.6	-4.5 ± 0.6	-2.1 ± 0.6	-0.6 ± 0.6
	I36L	2.2 ± 0.6	3.7 ± 0.7	4.3 ± 0.6	-1.5 ± 0.6
	L69S	6.2 ± 1.3	2.1 ± 1.3	17.2 ± 1.0	4.2 ± 1.2
	L69T	5.0 ± 1.8	7.0 ± 1.7	18.4 ± 1.5	9.9 ± 1.6
open mutants	I13F	10.3 ± 1.1	4.8 ± 1.1	1.4 ± 1.0	-1.7 ± 1.0
	I36A	-1.0 ± 0.7	9.9 ± 0.7	0.9 ± 0.6	0.8 ± 0.6

Assuming that the pincer mode is indeed the determining factor both in the binding as well as in the selected mutants, it would be expected that binding affinity would decrease if the preferred state of complex and mutant are not compatible, while binding should improve or not be affected if complex and mutant prefer the same state. As can be seen in figure 4.10, most binding free energy changes for mutations preferring the same state as the complexes they have been introduced are within ± 1.0kcal/mol of zero. Most of the mutations that prefer a pincer mode conformation that is different from the complex cause a $\Delta\Delta G_{binding}$ which is above 1.0kcal/mol. This indicates a

destabilisation of the complex and a decrease of binding affinity. The average change in binding free energy is $\Delta\Delta G_{binding,compatible} = 1.8$ kjoule/mol for the mutants that are compatible with the complex and $\Delta\Delta G_{binding,incompatible} = 6.3$ kJ mol^{-1} for mutants that are not compatible. The difference is statistically significant with a 95% confidence level according to a two-sample t-test performed on the data. However, in four cases (1nbf-I36A, 1xd3-V5L, dsk2-V5L and dsk2-I36L), the free energy of binding is hardly affected by mutation at all. In the case of the V5L mutation, the $\Delta\Delta G_{binding}$ values calculated here would indicate that all complexes are stabilised, and those that require the closed state, which should be preferred by V5L are actually more stable than the ones requiring the open conformation (as would have been expected), even though the difference is actually very small.

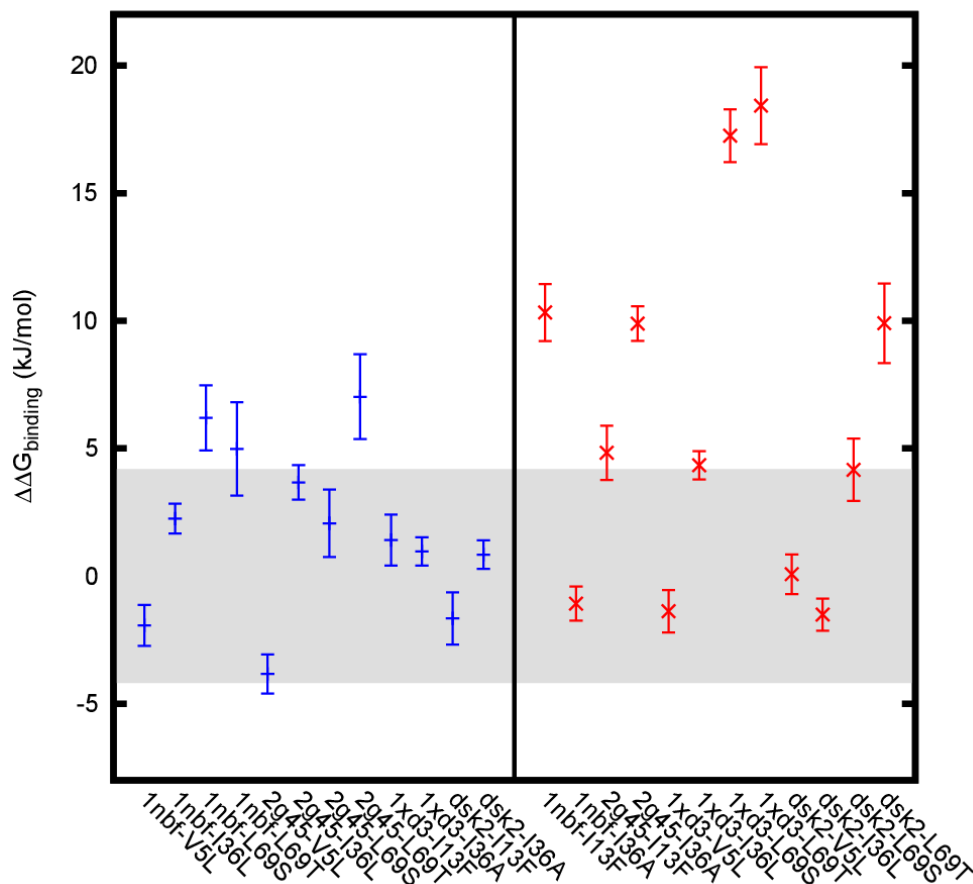**Figure 4.10:** $\Delta\Delta G_{binding}$ **(in kJ/mol) for different mutations in different complexes.** Mutations that prefer the same pincer mode state as the complex are plotted in blue on the left side, those preferring a different conformation are plotted in red on the right side. The grey area marks a range of ± 1.0kcal/mol, values within this range cannot be assumed to be significantly different from zero.

A look at the potential of mean force along the *pincer mode* (figure 4.11) shows that the dynamics of mutants in the complex are usually similar to those of the wild type in the same complex. It remains to be determined if this is a result of binding having a stronger influence on the conformational behaviour of ubiquitin than mutation. The equilibrium simulations used to estimate the PMFs in figure 4.11 were all started from conformations taken from the corresponding bound wild type ensembles, so a bias due to the starting structures cannot be excluded at this time. It should be noted, however, that the starting structures for the unbound mutant simulations were also picked from the unbound wild type ensemble and hence contained a similar starting structure bias that the mutant simulations seem to have overcome quickly.
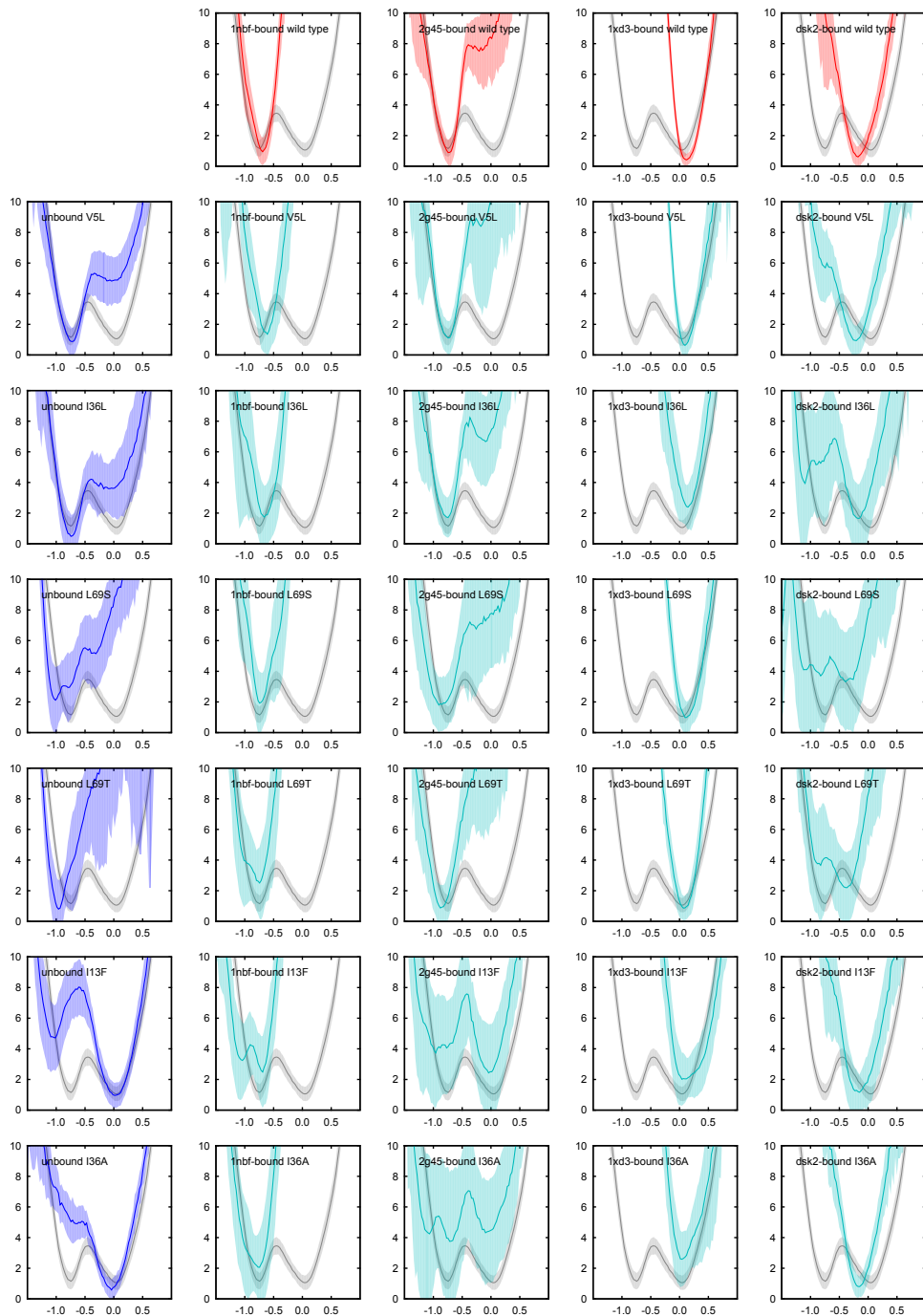
Figure 4.11: Comparisons of potentials of mean force for bound wild type ubiquitin, unbound and bound mutated ubiquitin. PMFs estimated from wild type ubiquitin simulations in the complex are shown in the first row in red, PMFs calculated from umbrella sampling simulations of unbound mutants are plotted in the first column in blue. It can be seen that the PMFs estimated from simulations of mutant ubiquitin in the complex (in cyan) are in most cases similar to those of the wild type in the complex, indicating that complex formation has a stronger influence on ubiquitin dynamics than single point mutations.

5

Discussion

5.1 Influence of binding on ubiquitin dynamics

We compared ensembles of ubiquitin structures from molecular dynamics (MD) simulations with and without binding partners aimed at a detailed investigation of the conformational effects of protein binding.

The main collective mode of fluctuation found in unbound ubiquitin is the *pincer mode* which strongly influences the shape of the binding surface (fig. 3.2). It has been indicated (12) that the flexibility of this mode is essential for ubiquitin to interact with a large number of different binding partners. Indeed, this mode is characteristically affected by binding as all but one of the bound ensembles show a significant shift or restriction of conformational density, while still the whole range of flexibility of unbound ubiquitin is required to accommodate all observed bound states. As all bound ensembles are completely covered by the unbound ensemble along the *pincer mode*, the conformational selection model is applicable for this aspect of binding.

Employing the partial least squares distinction analysis method that specifically aims at identifying differences between ensembles, low amplitude difference modes between bound and unbound ubiquitin ensembles were identified.

The observation of the unbound protein displaying the bound state conformation is often considered indicative of conformational selection ((12, 27, 78, 159)). I observed a significant fraction of the unbound ubiquitin ensemble showing a strong similarity (especially in the main *pincer mode*) to the conformations of bound ubiquitin. This is consistent with a conformational selection binding scenario, while the differences between bound and unbound ensembles on the local level indicate residual induced fit effects as have been introduced in recent binding models (29, 30, 31).

It is still possible that a portion of the binding events occurs according to an induced fit scenario. An alternative classification of the binding process is based on the inclusion of binding kinetics (160, 161, 162). As I have concentrated our analysis on the comparison of bound and unbound states rather than on association and dissociation events, a kinetic approach is beyond the scope of this thesis.

An aspect not considered in recently discussed binding models (29, 30, 31) is the

dynamic nature of bound proteins. Earlier work (163) already indicated that binding does not necessarily decrease the conformational entropy of proteins. I have also found that the dynamics of the bound ubiquitin ensembles are on a similar scale as those of unbound ubiquitin (fig. 3.3, 3.4).

In general, two effects of binding on the conformational space of the protein can be expected (fig. 5.1). Conformations accessible to the unbound protein can be prohibited by interactions (fig. 3.5) with the binding partner (conformational restriction) while conformations that were energetically unfavourable to the unbound protein can become accessible due to favourable interactions with the binding partner (conformational extension).
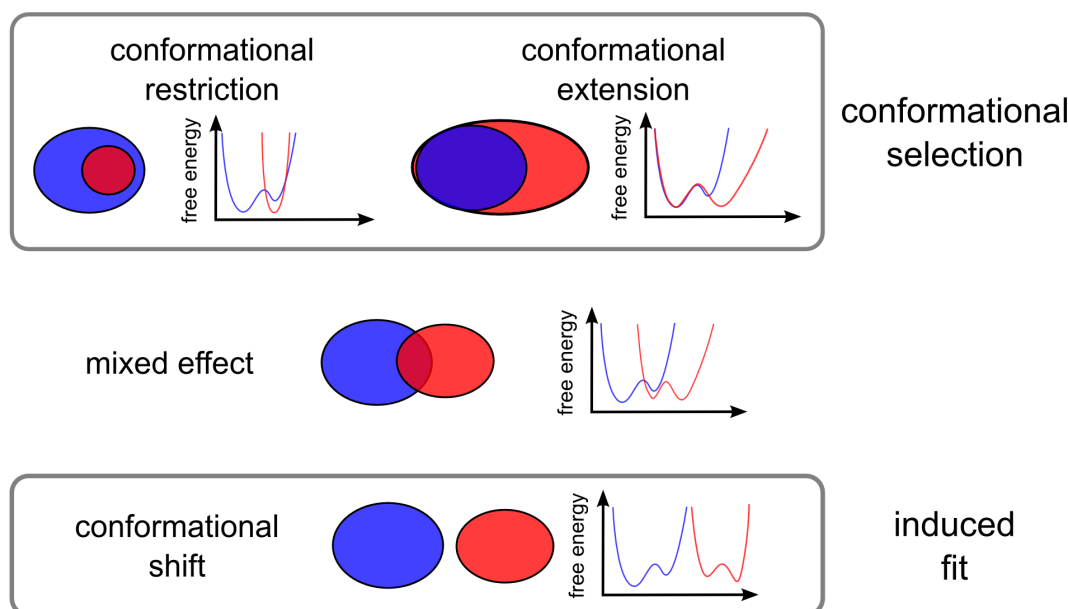

Figure 5.1: Schematic description of the proposed binding models. The blue ensemble would be that of the unbound protein, the red that of the bound. A sketch of possible free energy profiles fitting the corresponding models is given on the right.

These two effects are not mutually exclusive and indeed in most cases I observe a combination of both effects in the binding behaviour of ubiquitin. In the most extreme cases, all conformations accessible to the unbound protein are restricted, with all theconformations in the complex being the effect of conformational extension. This "conformational shift" corresponds best to the induced fit binding model.

In the case of conformational extension, changes of the energy landscape due to binding allow the protein to access conformations that are energetically unfavourable in the absence of the binding partner. While not generally considered, conformational extension is well compatible with the conformational selection model of binding, as the binding process itself can well take place in the overlap between the bound and unbound states.

Most complexes considered in this study can be described by the scenario of conformational extension combined with conformational restriction, showing a significant overlap between bound and unbound ensembles. Interestingly, also for those complex with near-zero overall overlap, substantial overlap is found between the bound and unbound states on the level of individual residues. Hence, for these complexes, each residue samples states in the unbound state that are found in the bound state, but the probability to find all contact residues in a complex compatible state *simultaneously* approaches zero for these complexes, resulting in zero overall overlap.

The consideration of conformational ensembles is a common feature of modern computational protein docking approaches to account for conformational changes due to binding (20, 164). Our results suggest that while native conformational ensembles are likely to yield good binding conformations on a global scale, small-scale structural adaptions at the binding interface seem to occur that are specifically caused by interactions with the binding partner.

5.2 Designing ubiquitin mutants to influence the binding behaviour

In chapter 3 I observed, that the pincer mode of ubiquitin is characteristically restricted when it is bound to different binding partners. In chapter 4, I intended to use this knowledge to selectively change the binding affinity of ubiquitin to certain binding partners by changing its population along this mode. Experimentally verified examples (152, 153) indicate that this can be achieved by mutations of residues that are not part of the binding interface of ubiquitin but of its hydrophobic core.

A protocol based on non-equilibrium free energy simulations was developed to estimate free energy differences between the two pincer mode states of different mutants in an effective way, allowing us to screen a total of 112 point mutations. For fifteen of these, the screening protocol returned promising values, eleven of which could be validated using a more extensive equilibrium free energy method. Considering the restrictions on computation time that are necessary to efficiently screen a large number of mutations, it can be expected that some results are not precise, as could also be seen in outliers in zero-cycles calculated with the method (section 4.2.3). Of the remaining eleven candidates that showed a clear preference for either the open or the closed state of ubiquitin, I introduced six into complexes that have been shown to to restrict ubiquitin dynamics to one of these states. It was expected, that an incompatible pairing of mutant and binding partner, that is a mutant shifted to the open state with a binding partner that restricts wild type ubiquitin to the closed state and vice versa, would destabilise the complex. Of twelve such pairings, this expectation was clearly fulfilled for eight, with a significant average decrease of binding affinity compared mutants that were introduced into compatible complexes. Of the remaining four pairings, two involve the binding partner *dsk2*. Wild type ubiquitin bound in this complex seems to be slightly less restricted to the open conformation than in complex with UCH-L3 (the binding partner in 1xd3).

I calculated changes in binding free energy up to $18.4\,\mathrm{kJ\,mol^{-1}}$. Published values for Binding affinities of ubiquitin complexes range from $1690\,\mu\mathrm{M}$ to $2.8\,\mu\mathrm{M}$, which (according to equation 1.2) corresponds to binding free energies from $\Delta G_D = -15.9\,\mathrm{kJ\,mol^{-1}}$ to $\Delta G_D = -31.9\,\mathrm{kJ\,mol^{-1}}$. The calculated changes in binding free energy for some of the mutants found in chapter 4 are of similar magnitude and hence should severely influence the binding behaviour of the mutants. In general, while the complexes used in 4 do not spontaneously switch between states in 10x100ns MD simulations, there is no information on how energetically unfavourable the other state really is, or if there are alternative binding modes that play a role for the mutant.

As seen in figure 4.11, the dynamics of the mutants seems to be determined by the binding partner in most cases, which requires a substantial shift in the population distribution along the pincer mode. In terms of the applicability of the different binding models, this would increase the role of induced fit effects in the case of incompatible mutant-complex pairs. However, as shown in figure 4.10, this change is in most cases connected to a significant decrease in binding affinity, indicating that the conformational flexibility of ubiquitin and its ability to bind by conformational selection are an important contribution to its function as they enable it to bind to a variety of binding partners.

5.3   Outlook

The results described in chapter 4 still have to be considered work in progress and further studies on the topic will be necessary to validate and better understand the results. We were able to identify several candidates of ubiquitin mutants selective for one of the *pincer mode* states, which should significantly influence binding behaviour. First attempts to validate this have had some promising but also some unexpected results - further investigations of the effects of binding should prove useful. in addition, experiments to verify the predicted changes in binding affinity for various mutants have been planned.

A topic of ongoing research are the mechanisms by which receptors select for different configurations of polyubiquitin chains (53). Recently published structures of such polybiquitin chains without and with binding partners have provided first insights into these mechanisms, but to our knowledge the influence of protein dynamics in this context has not been extensively investigated.

Over the course of this project, methods have been developed that are also applicable beyond its scope. The different applications of partial least squares discrimination analysis (PLS-DA) discussed in chapter 3 should be useful in many cases where structural differences between structural ensembles are of interest. The FGTI/CGI protocol described in chapter 4 could be applicable to a number of protein engineering problems where the goal is a shift in population along a generalised coordinate.

References

[1] C Sanchez, C Lachaize, F Janody, B Bellon, L Rder, J Euzenat, F Rechenmann, and B Jacq. **Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database**. *Nucleic Acids Research*, **27**(1):89–94, 1999. 1

[2] L Salwinski. **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Research*, **32**(90001):449D–451, 2004. 1

[3] I A Wilson and R L Stanfield. **Antibody-antigen interactions: new structures and new conformational changes**. *Current Opinion in Structural Biology*, **4**(6):857–867, 1994. 1

[4] T Pawson and P Nash. **Protein-protein interactions define specificity in signal transduction**. *Genes & Development*, **14**(9):1027–1047, 2000. 1

[5] R D Vale. **The Molecular Motor Toolbox for Intracellular Transport**. *Cell*, **112**(4):467–480, 2003. 1

[6] P Uetz and R E Hughes. **Systematic and large-scale two-hybrid screens**. *Current Opinion in Microbiology*, **3**(3):303–308, 2000. 1

[7] C M Deane, Salwiski, I Xenarios, and D Eisenberg. **Protein Interactions Two Methods for Assessment of the Reliability of High Throughput Observations**. *Molecular & Cellular Proteomics*, **1**(5):349–356, 2002. 1

[8] E Sprinzak, S Sattath, and H Margalit. **How Reliable are Experimental ProteinProtein Interaction Data?** *Journal of Molecular Biology*, **327**(5):919–923, 2003. 1

[9] K D Wilkinson. **Quantitative analysis of protein-protein interactions**. *Methods in Molecular Biology*, **261**:1534, 2004. 1

[10] J Piehler. **New methodologies for measuring protein interactions in vivo and in vitro**. *Current Opinion in Structural Biology*, **15**(1):4–14, 2005. 1

[11] B A Shoemaker and A R Panchenko. **Deciphering Protein-Protein Interactions. Part I. Experimental Techniques and Databases**. *PLoS Comput Biol*, **3**(3):e42, 2007. 1

[12] O F Lange, N-A Lakomek, C Fares, G F Schroder, K F A Walter, S Becker, J Meiler, H Grubmuller, C Griesinger, and B L de Groot. **Recognition**

REFERENCES

Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution. *Science*, **320**(5882):1471–1475, 2008. 1, 6, 28, 32, 34, 36, 59

[13] P L Kastritis and A M J J Bonvin. **Are Scoring Functions in ProteinProtein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark**. *Journal of Proteome Research*, **9**(5):2216–2225, 2010. 2

[14] J Fernández-Recio and M J E Sternberg. **The 4th meeting on the Critical Assessment of Predicted Interaction (CAPRI) held at the Mare Nostrum, Barcelona**. *Proteins: Structure, Function, and Bioinformatics*, **78**(15):3065–3066, 2010. 2, 20

[15] M F Lensink and S J Wodak. **Docking and scoring protein interactions: CAPRI 2009**. *Proteins: Structure, Function, and Bioinformatics*, **78**(15):3073–3084, 2010. 2

[16] S Sacquin-Mora, A Carbone, and R Lavery. **Identification of Protein Interaction Partners and ProteinProtein Interaction Sites**. *Journal of Molecular Biology*, **382**(5):1276–1289, 2008. 2

[17] M N Wass, G Fuentes, C Pons, F Pazos, and A Valencia. **Towards the prediction of protein interaction partners using physical docking**. *Molecular Systems Biology*, **7**(1), 2011. 2

[18] M N Wass, A David, and M J E Sternberg. **Challenges for the prediction of macromolecular interactions**. *Current Opinion in Structural Biology*, **21**(3):382–390, 2011. 2

[19] J Martin and R Lavery. **Arbitrary proteinprotein docking targets biologically relevant interfaces**. *BMC Biophysics*, **5**(1):7, 2012. PMID: 22559010. 2

[20] J Janin. **Protein protein docking tested in blind predictions: the CAPRI experiment**. *Molecular BioSystems*, **6**(12):2351, 2010. 2, 61

[21] P L Kastritis and A M J J Bonvin. **On the binding affinity of macromolecular interactions: daring to ask why proteins interact**. *Journal of The Royal Society Interface*, **10**(79), 2013. 2

[22] E Fischer. **Einfluss der Configuration auf die Wirkung der Enzyme**. *Berichte der deutschen chemischen Gesellschaft*, **27**(3):2985–2993, 1894. 3

[23] D E Koshland. **Application of a Theory of Enzyme Specificity to Protein Synthesis**. *Proceedings of the National Academy of Sciences*, **44**(2):98–104, 1958. 3

[24] D E Koshland. **Enzyme flexibility and enzyme action**. *Journal of Cellular and Comparative Physiology*, **54**(S1):245258, 1959. 3

[25] B Ma, S Kumar, C-J Tsai, and R Nussinov. **Folding funnels and binding mechanisms**. *Protein Engineering*, **12**(9):713 –720, 1999. 3

[26] C-T Tsai, B Ma, and R Nussinov. **Folding and Binding Cascades: Shifts in Energy Landscapes**. *Proceedings of the National Academy of Sciences*, **96**(18):9970–9972, 1999. 3

[27] D D Boehr, R Nussinov, and P E Wright. **The role of dynamic conformational ensembles in biomolecular recognition**. *Nature Chemical Biology*, **5**(11):789–796, 2009. 3, 59

[28] F B Straub and G Szabolcsi. **O dinamicseszkij aszpektah sztuktur fermentov.(On the dynamic aspects of protein structure) In**. *Molecular Biology, Problems and Perspectives*, page 182187, 1964. 3

[29] R Grünberg, J Leckner, and M Nilges. **Complementarity of Structure Ensembles in Protein-Protein Binding**. *Structure*, **12**(12):2125–2136, 2004. 3, 59

[30] T Wlodarski and B Zagrovic. **Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin**. *Proceedings of the National Academy of Sciences*, **106**(46):19346–19351, 2009. 3, 8, 59

[31] P Csermely, R Palotai, and R Nussinov. **Induced fit, conformational selection and independent dynamic segments: an extended view of binding events**. *Trends in Biochemical Sciences*, **35**(10):539–546, 2010. 3, 59

[32] S Vijay-kumar, C E Bugg, and W J Cook. **Structure of ubiquitin refined at 1.8 resolution**. *Journal of Molecular Biology*, **194**(3):531–544, 1987. 4, 8, 32

[33] R Ramage, J Green, T W Muir, O M Ogunjobi, S Love, and K Shaw. **Synthetic, structural and biological studies of the ubiquitin system: the total chemical synthesis of ubiquitin**. *Biochemical Journal*, **299**:151–158, 1994. 4, 8, 32

[34] A Varshavsky. **The early history of the ubiquitin field**. *Protein Science : A Publication of the Protein Society*, **15**(3):647–654, 2006. PMID: 16501229 PMCID: PMC2249785. 4

[35] A Ciehanover, Y Hod, and A Hershko. **A heat-stable polypeptide component of an ATP-dependent proteolytic system from reticulocytes**. *Biochemical and Biophysical Research Communications*, **81**(4):1100–1105, 1978. 4

[36] A Hershko, A Ciechanover, H Heller, A L Haas, and I A Rose. **Proposed role of ATP in protein breakdown: conjugation of protein with multiple chains of the polypeptide of ATP-dependent proteolysis**. *Proceedings of the National Academy of Sciences*, **77**(4):1783–1786, 1980. 4

[37] K D Wilkinson, M K Urban, and A L Haas. **Ubiquitin is the ATP-dependent proteolysis factor I of rabbit reticulocytes**. *Journal of Biological Chemistry*, **255**(16):7529–7532, 1980. 4

[38] D Finley, B Bartel, and A Varshavsky. **The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis**. *Nature*, **338**(6214):394–401, 1989. 4

[39] L Hicke and H Riezman. **Ubiquitination of a Yeast Plasma Membrane Receptor Signals Its Ligand-Stimulated Endocytosis**. *Cell*, **84**(2):277–287, 1996. 4

[40] L Hicke and R Dunn. **Regulation of Membrane Protein Transport by Ubiquitin and Ubiquitin-Binding Proteins**. *Annual Review of Cell and Developmental Biology*, **19**(1):141–172, 2003. PMID: 14570567. 4

REFERENCES

[41] M G Goebl, J Yochem, S Jentsch, J P McGrath, A Varshavsky, and B Byers. **The yeast cell cycle gene CDC34 encodes a ubiquitin-conjugating enzyme**. *Science*, **241**(4871):1331–1335, 1988. 4

[42] S Jentsch, J P McGrath, and A Varshavsky. **The yeast DNA repair gene RAD6 encodes a ubiquitin-conjugating enzyme**. *Nature*, **329**(6135):131–134, 1987. 4

[43] M Hochstrasser and A Varshavsky. **In vivo degradation of a transcriptional regulator: The yeast 2 repressor**. *Cell*, **61**(4):697–708, 1990. 4

[44] E zkaynak, D Finley, and A Varshavsky. **The yeast ubiquitin gene: head-to-tail repeats encoding a polyubiquitin precursor protein**. *Nature*, **312**(5995):663–666, 1984. 4

[45] D Finley, E zkaynak, and A Varshavsky. **The yeast polyubiquitin gene is essential for resistance to high temperatures, starvation, and other stresses**. *Cell*, **48**(6):1035–1046, 1987. 4

[46] A Hershko and A Ciechanover. **The Ubiquitin System for Protein Degradation**. *Annual Review of Biochemistry*, **61**(1):761–807, 1992. PMID: 1323239. 4

[47] A Hershko and A Ciechanover. **The Ubiquitin System**. *Annual Review of Biochemistry*, **67**(1):425–479, 1998. PMID: 9759494. 4

[48] T Kirisako, K Kamei, S Murata, M Kato, H Fukumoto, M Kanie, S Sano, F Tokunaga, K Tanaka, and K Iwai. **A ubiquitin ligase complex assembles linear polyubiquitin chains**. *The EMBO Journal*, **25**(20):4877–4887, 2006. 4

[49] K Iwai and F Tokunaga. **Linear polyubiquitination: a new regulator of NF-B activation**. *EMBO reports*, **10**(7):706–713, 2009. 4

[50] D Komander, F Reyes-Turcu, J D F Licchesi, P Odenwaelder, K D Wilkinson, and D Barford. **Molecular discrimination of structurally equivalent Lys 63-linked and linear polyubiquitin chains**. *EMBO reports*, **10**:466 – 473, 2009. 4

[51] F Ikeda and I Dikic. **Atypical ubiquitin chains: new molecular signals. Protein Modifications: Beyond the Usual Suspects' Review Series**. *EMBO Reports*, **9**(6):536–542, 2008. PMID: 18516089 PMCID: PMC2427391. 4

[52] H T Kim, K P Kim, F Lledias, A F Kisselev, K M Scaglione, D Skowyra, S P Gygi, and A L Goldberg. **Certain Pairs of Ubiquitin-conjugating Enzymes (E2s) and Ubiquitin-Protein Ligases (E3s) Synthesize Nondegradable Forked Ubiquitin Chains Containing All Possible Isopeptide Linkages**. *Journal of Biological Chemistry*, **282**(24):17375–17386, 2007. 4

[53] D Komander and M Rape. **The Ubiquitin Code**. *Annual Review of Biochemistry*, **81**(1):203–229, 2012. PMID: 22524316. 4, 62

[54] M Petroski. **The ubiquitin system, disease, and drug discovery**. *BMC Biochemistry*, **9**(Suppl 1):S7, 2008. 5

[55] A Mani and E P Gelmann. **The Ubiquitin-Proteasome Pathway and Its Role in Cancer**. *Journal of Clinical Oncology*, **23**(21):4776–4789, 2005. 5

[56] A M Sheehy, N C Gaddis, J D Choi, and M H Malim. **Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein**. *Nature*, **418**(6898):646–650, 2002. 5

[57] X Yu, Y Yu, B Liu, K Luo, W Kong, P Mao, and X-F Yu. **Induction of APOBEC3G Ubiquitination and Degradation by an HIV-1 Vif-Cul5-SCF Complex**. *Science*, **302**(5647):1056–1060, 2003. 5

[58] J E Davies, S Sarkar, and D C Rubinsztein. **The ubiquitin proteasome system in Huntington's disease and the spinocerebellar ataxias**. *BMC Biochemistry*, **8**(Suppl 1):S2, 2007. 5

[59] S C Upadhya and A N Hegde. **Role of the ubiquitin proteasome system in Alzheimer's disease**. *BMC Biochemistry*, **8**(Suppl 1):S12, 2007. 5

[60] I Dikic, S Wakatsuki, and K J Walters. **Ubiquitin-binding domains - from structures to functions**. *Nature Reviews Molecular Cell Biology*, **10**(10):659–671, 2009. 5, 6

[61] L Hicke, H L Schubert, and C P Hill. **Ubiquitin-binding domains**. *Nature Reviews Molecular Cell Biology*, **6**(8):610–621, 2005. 5

[62] J Hurley, S Lee, and G Prag. **Ubiquitin-binding domains**. *Biochemical Journal*, **399**:361–372, 2006. 5

[63] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. **The Protein Data Bank**. *Nucleic Acids Research*, **28**(1):235 –242, 2000. 6, 31

[64] M Hu, P Li, M Li, W Li, T Yao, J-W Wu, W Gu, R E Cohen, and Y Shi. **Crystal Structure of a UBP-Family Deubiquitinating Enzyme in Isolation and in Complex with Ubiquitin Aldehyde**. *Cell*, **111**(7):1041–1054, 2002. 8, 31

[65] G Prag, S Misra, E A Jones, R Ghirlando, B A Davies, B F Horazdovsky, and J H Hurley. **Mechanism of Ubiquitin Recognition by the CUE Domain of Vps9p**. *Cell*, **113**(5):609–620, 2003. 8, 31

[66] W I Sundquist, H L Schubert, B N Kelly, G C Hill, J M Holton, and C P Hill. **Ubiquitin Recognition by the Human TSG101 Protein**. *Molecular Cell*, **13**(6):783–789, 2004. 8, 31

[67] H Teo, D B Veprintsev, and R L Williams. **Structural Insights into Endosomal Sorting Complex Required for Transport (ESCRT-I) Recognition of Ubiquitinated Proteins**. *Journal of Biological Chemistry*, **279**(27):28689–28696, 2004. 8, 31

[68] S Misaghi, P J Galardy, W J N Meester, H O, H L Ploegh, and R Gaudet. **Structure of the Ubiquitin Hydrolase UCH-L3 Complexed with a Suicide Substrate**. *Journal of Biological Chemistry*, **280**(2):1512–1520, 2005. 8, 31

REFERENCES

[69] S Hirano, M Kawasaki, H Ura, R Kato, C Raiborg, H Stenmark, and S Wakatsuki. **Double-sided ubiquitin binding of Hrs-UIM in endosomal protein sorting**. *Nature Structural & Molecular Biology*, **13**(3):272–277, 2006. 8, 31

[70] S Lee, Y C Tsai, R Mattera, W J Smith, M S Kostelansky, A M Weissman, J S Bonifacino, and J H Hurley. **Structural basis for ubiquitin recognition and autoubiquitination by Rabex-5**. *Nature Structural & Molecular Biology*, **13**(3):264–271, 2006. 8, 31

[71] F E Reyes-Turcu, J R Horton, J E Mullally, A Heroux, X Cheng, and K D Wilkinson. **The Ubiquitin Binding Domain ZnF UBP Recognizes the C-Terminal Diglycine Motif of Unanchored Ubiquitin**. *Cell*, **124**(6):1197–1208, 2006. 8, 31

[72] S L Alam, C Langelier, F G Whitby, S Koirala, H Robinson, C P Hill, and W I Sundquist. **Structural basis for ubiquitin recognition by the human ESCRT-II EAP45 GLUE domain**. *Nature Structural & Molecular Biology*, **13**(11):1029–1030, 2006. 8, 31

[73] G V Avvakumov, J R Walker, G Bernstein, S Xue, P J Finerty Jr., F MacKenzie, J Weigelt, M Sundstrom, C H Arrowsmith, A M Edwards, A Bockjarev, and S Dhe-Paganon. **Covalent Ubiquitin-USP2 Complex**. *Protein Data Bank*, 2006. 8, 31

[74] P Peschard, G Kozlov, T Lin, I A Mirza, A M Berghuis, S Lipkowitz, M Park, and K Gehring. **Structural Basis for Ubiquitin-Mediated Dimerization and Activation of the Ubiquitin Protein Ligase Cbl-b**. *Molecular Cell*, **27**(3):474–485, 2007. 8, 31

[75] N-A Lakomek, O F Lange, K F A Walter, C Fars, D Egger, P Lunkenheimer, J Meiler, H Grubmller, S Becker, B L deGroot, and C Griesinger. **Residual dipolar couplings as a tool to study molecular recognition of ubiquitin**. *Biochemical Society Transactions*, **36**(6):1433, 2008. 6

[76] K Lindorff-Larsen, R B Best, M A DePristo, C M Dobson, and M Vendruscolo. **Simultaneous determination of protein structure and dynamics**. *Nature*, **433**(7022):128–132, 2005. 6

[77] B Richter, J Gsponer, P Vrnai, X Salvatella, and M Vendruscolo. **The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins**. *Journal of Biomolecular NMR*, **37**(2):117–135, 2007. 6

[78] D Long and R Brüschweiler. **In Silico Elucidation of the Recognition Dynamics of Ubiquitin**. *PLoS Computational Biologyl*, **7**(4):e1002035, 2011. 8, 36, 59

[79] M Born and R Oppenheimer. **Zur Quantentheorie der Molekeln**. *Annalen der Physik*, **389**(20):457484, 1927. 12

[80] D van der Spoel, E Lindahl, B Hess, A van Buuren, E Apol, P J Meulenhoff, D P Tieleman, A L T M Sijbers, K A Feenstra, R van Drunen, and H J C Berendsen. *Gromacs User Manual 4.5.4*. 2010. 12, 13, 14

[81] W F VAN GUNSTEREN AND H J C BERENDSEN. **Groningen molecular simulation (GROMOS) library manual**. *Biomos, Groningen*, **24**:682704, 1987. 13

[82] WILFRED VAN GUNSTEREN, SR BILLETER, AA EISING, PHILIPPE HNENBERGER, P KRGER, ALAN MARK, WRP SCOTT, AND ILARIO TIRONI. *Biomolecular Simulation: The GROMOS96 manual and user guide.* 1996. 13

[83] A D MACKERELL, D BASHFORD, M BELLOTT, R L DUNBRACK JR., J D EVANSECK, M J FIELD, S FISCHER, J GAO, H GUO, S HA, D JOSEPH-MCCARTHY, L KUCHNIR, K KUCZERA, F T K LAU, C MATTOS, S MICHNICK, T NGO, D T NGUYEN, B PRODHOM, W E REIHER, B ROUX, M SCHLENKRICH, J C SMITH, R STOTE, J STRAUB, M WATANABE, J WIRKIEWICZ-KUCZERA, D YIN, AND M KARPLUS. **All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins**. *The Journal of Physical Chemistry B*, **102**(18):3586–3616, 1998. 13

[84] A D MACKERELL, M FEIG, AND C L BROOKS. **Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations**. *Journal of Computational Chemistry*, **25**(11):14001415, 2004. 13

[85] W L JORGENSEN, D S MAXWELL, AND J TIRADO-RIVES. **Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids**. *Journal of the American Chemical Society*, **118**(45):11225–11236, 1996. 13

[86] W D CORNELL, P CIEPLAK, C I BAYLY, I R GOULD, K M MERZ, D M FERGUSON, D C SPELLMEYER, T FOX, J W CALDWELL, AND P A KOLLMAN. **A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules**. *Journal of the American Chemical Society*, **117**(19):5179–5197, 1995. 13

[87] J WANG, P CIEPLAK, AND P A KOLLMAN. **How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?** *Journal of Computational Chemistry*, **21**(12):10491074, 2000. 13

[88] 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: Structure, Function, and Bioinformatics*, **65**(3):712–725, 2006. 13, 32

[89] D-W LI AND R BRSCHWEILER. **NMRBased Protein Potentials**. *Angewandte Chemie*, **122**(38):6930–6932, 2010. 13

[90] 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: Structure, Function, and Bioinformatics*, pages NA–NA, 2010. 13

[91] J E JONES. **On the Determination of Molecular Fields. II. From the Equation of State of a Gas**. *Proceedings of the Royal Society of London. Series A*, **106**(738):463–477, 1924. 13

REFERENCES

[92] H Fan, A E Mark, J Zhu, and B Honig. **Comparative study of generalized Born models: Protein dynamics**. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(19):6760 –6764, 2005. 14

[93] Y M Rhee, E J Sorin, G Jayachandran, E Lindahl, and V S Pande. **Simulations of the role of water in the protein-folding mechanism**. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(17):6456 –6461, 2004. 14

[94] R Zhou and b J Berne. **Can a continuum solvent model reproduce the free energy landscape of a -hairpin folding in water?** *Proceedings of the National Academy of Sciences*, **99**(20):12777 –12782, 2002. 14

[95] H J C Berendsen, J P M Postma, W F Van Gunsteren, and J Hermans. **Interaction models for water in relation to protein hydration**. *Intermolecular forces*, **11**(1):331342, 1981. 14

[96] W L Jorgensen, J Chandrasekhar, J D Madura, R W Impey, and M L Klein. **Comparison of simple potential functions for simulating liquid water**. *The Journal of Chemical Physics*, **79**(2):926–935, 1983. 14

[97] H J C Berendsen, J R Grigera, and T P Straatsma. **The missing term in effective pair potentials**. *The Journal of Physical Chemistry*, **91**(24):6269–6271, 1987. 14, 32

[98] R W Hockney, S P Goel, and J W Eastwood. **Quiet high-resolution computer models of a plasma**. *Journal of Computational Physics*, **14**(2):148–158, 1974. 16

[99] A K Mazur. **Common Molecular Dynamics Algorithms Revisited: Accuracy and Optimal Time Steps of StrmerLeapfrog Integrators**. *Journal of Computational Physics*, **136**(2):354–365, 1997. 16, 18

[100] I Ohmine, H Tanaka, and P G Wolynes. **Large local energy fluctuations in water. II. Cooperative motions and fluctuations**. *The Journal of Chemical Physics*, **89**(9):5852–5860, 1988. 17

[101] D van der Spoel and P J van Maaren. **The Origin of Layer Structure Artifacts in Simulations of Liquid Water**. *Journal of Chemical Theory and Computation*, **2**(1):1–11, 2006. 17

[102] H Bekker. **Unification of box shapes in molecular simulations**. *Journal of Computational Chemistry*, **18**(15):19301942, 1997. 17

[103] T Darden, D York, and L Pedersen. **Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems**. *The Journal of Chemical Physics*, **98**(12):10089, 1993. 17, 32

[104] U Essmann, L Perera, M L Berkowitz, T Darden, H Lee, and L G Pedersen. **A smooth particle mesh Ewald method**. *The Journal of Chemical Physics*, **103**(19):8577, 1995. 17, 32

[105] 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**. *The Journal of Chemical Physics*, **81**(8):3684, 1984. 18, 32

[106] S Nos. **A molecular dynamics method for simulations in the canonical ensemble**. *Molecular Physics*, **52**(2):255–268, 1984. 18

[107] W G Hoover. **Canonical dynamics: Equilibrium phase-space distributions**. *Physical Review A*, **31**(3):1695–1697, 1985. 18

[108] G Bussi, D Donadio, and M Parrinello. **Canonical sampling through velocity rescaling**. *The Journal of Chemical Physics*, **126**(1):014101, 2007. 18, 32

[109] G J Martyna, M E Tuckerman, D J Tobias, and M L Klein. **Explicit reversible integrators for extended systems dynamics**. *Molecular Physics*, **87**(5):1117–1157, 1996. 18

[110] S Nos and M L Klein. **Constant pressure molecular dynamics for molecular systems**. *Molecular Physics*, **50**(5):1055–1076, 1983. 18

[111] M Parrinello and A Rahman. **Polymorphic transitions in single crystals: A new molecular dynamics method**. *Journal of Applied Physics*, **52**(12):7182–7190, 1981. 18

[112] K A Feenstra, B Hess, and H J C Berendsen. **Improving efficiency of large timescale molecular dynamics simulations of hydrogenrich systems**. *Journal of Computational Chemistry*, **20**(8):786–798, 1999. 18, 19, 32

[113] S Miyamoto and P A Kollman. **Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models**. *Journal of Computational Chemistry*, **13**(8):952–962, 1992. 18, 32

[114] B Hess, H Bekker, H J C Berendsen, and J G E M Fraaije. **LINCS: A linear constraint solver for molecular simulations**. *Journal of Computational Chemistry*, **18**(12):1463–1472, 1997. 18, 32

[115] R W Zwanzig. **HighTemperature Equation of State by a Perturbation Method. I. Nonpolar Gases**. *The Journal of Chemical Physics*, **22**(8):1420–1426, 1954. 20

[116] N Lu and D A Kofke. **Accuracy of free-energy perturbation calculations in molecular simulation. I. Modeling**. *The Journal of Chemical Physics*, **114**(17):7303–7311, 2001. 21

[117] N Lu and D A Kofke. **Accuracy of free-energy perturbation calculations in molecular simulation. II. Heuristics**. *The Journal of Chemical Physics*, **115**(15):6866–6875, 2001. 21

[118] N Lu, J K Singh, and D A Kofke. **Appropriate methods to combine forward and reverse free-energy perturbation averages**. *The Journal of Chemical Physics*, **118**(7):2977–2984, 2003. 21

[119] J G Kirkwood. **Statistical Mechanics of Fluid Mixtures**. *The Journal of Chemical Physics*, **3**(5):300–313, 1935. 21, 23

[120] C Jarzynski. **Nonequilibrium Equality for Free Energy Differences**. *Physical Review Letters*, **78**(14):2690–2693, 1997. 22

REFERENCES

[121] C Jarzynski. **Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach**. *Physical Review E*, **56**(5):5018–5035, 1997. 22

[122] R C Lua and A Y Grosberg. **Practical Applicability of the Jarzynski Relation in Statistical Mechanics: A Pedagogical Example**. *The Journal of Physical Chemistry B*, **109**(14):6805–6811, 2005. 22

[123] C Jarzynski. **Rare events and the convergence of exponentially averaged work values**. *Physical Review E*, **73**(4):046105, 2006. 22

[124] G Crooks. **Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems**. *Journal of Statistical Physics*, **90**(5):1481–1487, 1998. 22

[125] G E Crooks and C Jarzynski. **Work distribution for the adiabatic compression of a dilute and interacting classical gas**. *Physical Review E*, **75**(2):021116, 2007. 22

[126] M Goette and H Grubmller. **Accuracy and convergence of free energy differences calculated from nonequilibrium switching processes**. *Journal of Computational Chemistry*, **30**(3):447456, 2009. 22

[127] H J C Berendsen. *Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics*. Cambridge University Press, 2007. 23

[128] G M Torrie and J P Valleau. **Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling**. *Journal of Computational Physics*, **23**(2):187–199, 1977. 23

[129] S Kumar, J M Rosenberg, D Bouzida, R H Swendsen, and P A Kollman. **The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method**. *Journal of Computational Chemistry*, **13**(8):1011–1021, 1992. 23, 52

[130] J S Hub, B L de Groot, and D van der Spoel. **g_whamA Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates**. *Journal of Chemical Theory and Computation*, **6**(12):3713–3720, 2010. 23, 52

[131] A Kitao, F Hirata, and N G. **The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum**. *Chemical Physics*, **158**(2-3):447–472, 1991. 24

[132] A E García. **Large-amplitude nonlinear motions in proteins**. *Physical Review Letters*, **68**(17):2696–2699, 1992. 24

[133] A Amadei, A B M Linssen, and H J C Berendsen. **Essential dynamics of proteins**. *Proteins: Structure, Function, and Bioinformatics*, **17**(4):412–425, 1993. 24, 25

[134] J S Hub and B L de Groot. **Detection of Functional Modes in Protein Dynamics**. *PLoS Computational Biology*, **5**(8):e1000480, 2009. 25

[135] T Krivobokova, R Briones, J S Hub, A Munk, and B L deGroot. **Partial Least-Squares Functional Mode Analysis: Application to the Membrane Proteins AQP1, Aqy1, and CLC-ec1**. *Biophysical Journal*, **103**(4):786–796, 2012. 25

[136] I S Helland. **On the structure of partial least squares regression**. *Communications in Statistics – Simulation and Computation*, **17**(2):581–607, 1988. 26

[137] M C Denham. **Implementing partial least squares**. *Statistics and Computing*, **5**(3):191–202, 1995. 26

[138] J MacQueen. **Some methods for classification and analysis of multivariate observations**. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, **1**, page 14, 1967. 28

[139] S Lloyd. **Least squares quantization in PCM**. *IEEE Transactions on Information Theory*, **28**(2):129 – 137, 1982. 28

[140] A P Dempster, N M Laird, and D B Rubin. **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1):1–38, 1977. ArticleType: research-article / Full publication date: 1977 / Copyright 1977 Royal Statistical Society. 28

[141] B Efron. **Bootstrap Methods: Another Look at the Jackknife**. *The Annals of Statistics*, **7**(1):1–26, 1979. 30

[142] D N Politis and J P Romano. **The Stationary Bootstrap**. *Journal of the American Statistical Association*, **89**(428):1303–1313, 1994. 30, 52

[143] J H Peters and B L de Groot. **Ubiquitin Dynamics in Complexes Reveal Molecular Recognition Mechanisms Beyond Induced Fit and Conformational Selection**. *PLoS Computational Biology*, **8**(10):e1002704, 2012. 31

[144] O F Lange, D van der Spoel, and B L de Groot. **Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data**. *Biophysical Journal*, **99**(2):647–655, 2010. 31

[145] B Hess, C Kutzner, D van der Spoel, and E Lindahl. **GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation**. *Journal of Chemical Theory and Computation*, **4**(3):435–447, 2008. 31, 55

[146] B Hess and N F A van der Vegt. **Hydration Thermodynamic Properties of Amino Acid Analogues: A Systematic Comparison of Biomolecular Force Fields and Water Models**. *The Journal of Physical Chemistry B*, **110**(35):17616–17626, 2006. 31

[147] E J Sorin and V S Pande. **Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations**. *Biophysical Journal*, **88**(4):2472–2493, 2005. 31

[148] J Schlitter. **Estimation of absolute and relative entropies of macromolecules using the covariance matrix**. *Chemical Physics Letters*, **215**(6):617–621, 1993. 32, 36

[149] M Barker and W Rayens. **Partial least squares for discrimination**. *Journal of Chemometrics*, **17**(3):166–173, 2003. 38

REFERENCES

[150] K E Sloper-Mould, J C Jemc, C M Pickart, and L Hicke. **Distinct Functional Surface Regions on Ubiquitin**. *Journal of Biological Chemistry*, **276**(32):30483–30489, 2001. 43

[151] K D Wilkinson, E Laleli-Sahin, J Urbauer, C N Larsen, G H Shih, A L Haas, S T R Walsh, and A J Wand. **The binding site for UCH-L3 on ubiquitin: mutagenesis and NMR studies on the complex between ubiquitin and UCH-L3**. *Journal of Molecular Biology*, **291**(5):1067–1077, 1999. 43, 54

[152] A Haririnia, R Verma, N Purohit, M Z Twarog, R J Deshaies, D Bolon, and D Fushman. **Mutations in the Hydrophobic Core of Ubiquitin Differentially Affect Its Recognition by Receptor Proteins**. *Journal of Molecular Biology*, **375**(4):979–996, 2008. 43, 49, 61

[153] Y Zhang, L Zhou, L Rouge, A H Phillips, C Lam, P Liu, W Sandoval, E Helgason, J M Murray, I E Wertz, and J E Corn. **Conformational stabilization of ubiquitin yields potent and selective inhibitors of USP7**. *Nature Chemical Biology*, **9**(1):51–58, 2013. 43, 44, 61

[154] A Amadei, A B M Linssen, B L de Groot, D M F van Aalten, and H J C Berendsen. **An Efficient Method for Sampling the Essential Subspace of Proteins**. *Journal of Biomolecular Structure and Dynamics*, **13**(4):615–625, 1996. 47

[155] B L de Groot, A Amadei, R M Scheek, N a J van Nuland, and H J C Berendsen. **An extended sampling of the configurational space of HPr from E. coli**. *Proteins: Structure, Function, and Bioinformatics*, **26**(3):314322, 1996. 47

[156] B L de Groot, A Amadei, D M F van Aalten, and H J C Berendsen. **Towards an Exhaustive Sampling of the Configurational Spaces of the Two Forms of the Peptide Hormone Guanylin**. *Journal of Biomolecular Structure and Dynamics*, **13**(5):741–751, 1996. 47

[157] G Vriend. **WHAT IF: A molecular modeling and drug design program**. *Journal of Molecular Graphics*, **8**(1):52–56, 1990. 49, 55

[158] D Seeliger and B L de Groot. **Protein Thermostability Calculations Using Alchemical Free Energy Simulations**. *Biophysical Journal*, **98**(10):2309–2316, 2010. PMID: 20483340 PMCID: 2872215. 49, 55

[159] R B Fenwick, S Esteban-Martn, and X Salvatella. **Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles**. *European Biophysics Journal*, **40**:1339–1355, 2011. 59

[160] H-X Zhou. **From Induced Fit to Conformational Selection: A Continuum of Binding Mechanism Controlled by the Timescale of Conformational Transitions**. *Biophysical Journal*, **98**(6):L15–L17, 2010. 59

[161] T R Weikl and C von Deuster. **Selected-fit versus induced-fit protein binding: Kinetic differences and mutational analysis**. *Proteins: Structure, Function, and Bioinformatics*, **75**(1):104110, 2009. 59

74

[162] I H MOAL AND P A BATES. **Kinetic Rate Constant Prediction Supports the Conformational Selection Mechanism of Protein Binding**. *PLoS Computational Biology*, **8**(1):e1002351, 2012. 59

[163] R GRÜNBERG, M NILGES, AND J LECKNER. **Flexibility and Conformational Entropy in Protein-Protein Binding**. *Structure*, **14**(4):683–693, 2006. 60

[164] G KUZU, O KESKIN, A GURSOY, AND R NUSSINO. **Expanding the Conformational Selection Paradigm in Protein-Ligand Docking**. In RICCARDO BARON AND JOHN M. WALKER, editors, *Computational Drug Discovery and Design*, **819** of *Methods in Molecular Biology*, pages 59–74. Springer New York, 2012. 61

Declaration

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other German or foreign examination board.

The thesis work was conducted from 2009 to 2013 under the supervision of Prof. Dr. Bert de Groot at the Max Planck Institute for Biophysical Chemistry.

Göttingen, 27.02.2013

Jan Henning Peters

Curriculum Vitae: Jan Henning Peters

Personal Information

Name: Jan Henning Peters
Adress: Stumpfe Eiche 13, 37077 Göttingen
Date and city of birth:: 09. September, 1978, Lippstadt
Nationality: German

Education

03.2009–   Max Planck Institute for Biophysical Chemistry, Göttingen
Thesis: On the Effect of Binding on Ubiquitin Dynamics Supervisor: Prof. Dr. Bert de Groot

10.2002–03.2009   University of Leipzig
International Physics Studies Progam (IPSP)
Thesis: *Effects of Laser-induced Heating in Nanopore- and Microcapillary Systems*
Supervisor: Dr. Ulrich Keyser, University of Leipuzig

12.2008   Graduated with degree *Master of Science*

10.1999–09.2002   University of Osnabrück
Cognive Science
Thesis: *Dynamische Verarbeitung melodischer Sequenzen mit einem Neuro-Fuzzy-System*
Supervisor: Dr. Tillman Weyde, University of Osnabrück

11.2003   Graduated with degree *Bacherlor of Science*, final grade: "very good" (B)

09.1991–06.1998 Ostendorf-Gynmasium Lippstadt

06.1998 *Abitur* (final grade: 1.5)

Articles in peer-reviewd journals

- J H Peters, M Tartz and H Neumann, Using design of experiments method to model beamlet properties. Review of Scientific Instruments, vol. 77, no. 3, p. 03B905, 2006.

- J H Peters and B L de Groot. Ubiquitin Dynamics in Complexes Reveal Molecular Recognition Mechanisms Beyond Induced Fit and Conformational Selection. PLoS Computational Biology, 8(10):e1002704, 2012. 31

Oral conference contributions

- *Dynamic Properties of Ubiquitin in Complexes* Workshop on Computer Simulation and Theory of Macromolecules, Hünfeld, 15.04.2011

- *Ubiquitin Dynamics in Complexes* Annual Meeting of the German Biophysical Society, Göttingen, 25.09.2012