Modeling Flexibility in Protein-DNA and Protein-Ligand Complexes using Molecular Dynamics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Dipl.-Inform. Nina Monika Fischer
aus Coburg

Tübingen
2013

Tag der mündlichen Qualifikation	02.09.2013
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter	Prof. Dr. Klaus Harter

Abstract

Molecular dynamics simulations provide valuable insights into the inherent flexibility of biomolecules and complement the static structures obtained from X-ray crystallography. Structural dynamics of proteins and alterations at protein-DNA and protein-ligand binding interfaces can thus be investigated.

Regulation of gene expression plays a pivotal role in cellular processes and is modulated by special proteins binding to specific DNA sequences. WRKY proteins represent a large protein family in plants and are involved in the regulation of developmental processes, such as leaf senescence, as well as in the response to abiotic and biotic stress situations. The highly conserved WRKY DNA-binding domain recognizes predominantly the 'TTGACC' W-box consensus sequence. Since all WRKY proteins prefer this DNA sequence motif, it remains unclear so far how stimulus-specific responses are mediated. The W-box sequence might be more degenerate and yet undetected differences at the protein-DNA binding interface might exist. With molecular modeling and molecular dynamics simulations we constructed a first three-dimensional WRKY-DNA complex at atomic detail. Our modeling approach facilitates the investigation of structural differences and similarities of WRKY DNA-binding domains in complex with DNA while accounting for flexibility. Complexes of N- and C-terminal DNA-binding domains of *At*WRKY33 with DNA and *in silico* binding free energy predictions gave a first indication that the N-terminal domain interacts with DNA in a similar way as the C-terminal domain. Additionally, we found an amino acid variation between *At*WRKY11 and *At*WRKY50 relevant for DNA-binding specificity and thus could explain WRKY DNA-binding motif preferences at the structural level.

Despite significant advances in structure-based drug design techniques during the last years, one remaining challenge is still the accurate representation of flexible proteins in protein-ligand docking. We present a method that can generate and identify possible protein conformations when only a single protein structure is available. These different protein conformations represent the structural space of flexible proteins and can be used as input structures for molecular docking. Aldose reductase, dihydrofolate reductase, and HIV-1 protease are prominent drug targets known for their large backbone flexibility. We tested our method on these three proteins and could identify at least one protein conformation that is highly similar to one of the structures available in the PDB each. Another method we developed is an advanced molecular dynamics simulation and binding affinity prediction protocol for re-scoring docked ligand poses. Our approach can be employed after protein-ligand docking to discriminate between good and inferior binders while additionally accounting for flexibility. When applied to an urokinase ligand dataset, it improves the correlation of 0.57 between experimental binding affinities and docking scores significantly to 0.92. Thus, it can advance the rank-ordering of docking hit lists, especially when applied to structurally similar ligands.

Zusammenfassung

Molekulardynamische Simulationen bieten die Möglichkeit wertvolle Einblicke in die Biomolekülen inhärente Flexibilität zu erlangen und können die Strukturen ergänzen, die durch Röntgenkristallografie gewonnen worden und deshalb bewegungslos sind. Somit können die Flexibilität von Proteinen und die strukturellen Veränderungen an Interaktionsflächen von Protein-DNS und Protein-Ligand Komplexen untersucht werden.

Die Regulation von Genexpression spielt bei allen Entwicklungsprozessen in Zellen eine zentrale Rolle und wird von bestimmten Proteinen kontrolliert, die an spezifische DNS Sequenzen binden. In Pflanzen bilden WRKY Proteine eine große Familie solcher Proteine, die an der Regulation von Entwicklungsprozessen, wie Blattalterung als auch bei der Bewältigung von biotischen und abiotischen Stresssituationen beteiligt sind. Ihre hochkonservierte DNS-Bindedomäne erkennt hauptsächlich die Konsensussequenz ‚TTGACC'. Da alle WRKY Proteine das gleiche DNS-Bindemotiv bevorzugen, konnte bis jetzt nicht eindeutig erklärt werden auf welche Art und Weise stimulus-spezifische Antworten vermittelt werden. Möglicherweise ist die W-Box Sequenz degenerativer und es existieren bisher unbekannte Unterschiede an der DNS-Protein Interaktionsfläche. Durch molekulare Modellierung und molekulardynamische Simulationen konnten wir einen ersten drei-dimensionalen WRKY-DNS Komplex auf atomarer Ebene konstruieren. Unser Modellierungsansatz macht die Untersuchung von strukturellen Unterschieden und Gemeinsamkeiten unter Beachtung von Flexibilität von WRKY DNS-Bindedomänen, die sich in Komplex mit DNS befinden, möglich. Komplexe von N- und C-terminalen DNS-Bindedomänen von *At*WRKY33 mit DNS und computergestützte Bindungsenergieberechnungen geben erste Hinweise auf ähnliche DNS-Interaktionen der N-terminalen verglichen mit der C-terminalen Domäne. Außerdem haben wir eine Aminosäurevariation zwischen *At*WRKY11 und *At*WRKY50 entdeckt, die relevant für die DNS-Bindespezifität ist, und können somit auf struktureller Ebene die Präferenz von WRKY Proteinen zu bestimmten DNS-Bindemotiven erklären.

Trotz bedeutender Fortschritte, die in den letzten Jahren bei Methoden im Bereich des strukturbasierten Wirkstoffentwurfs erreicht wurden, bleibt eine der noch nicht vollständig gelösten Probleme die wirklichkeitsgetreue Repräsentation von flexiblen Proteinen in Protein-Ligand Docking. Wir stellen einen Ansatz vor, der aus einer einzigen verfügbaren Proteinkonformation verschiedene Proteinkonformationen generieren und identifizieren kann. Diese generierten Proteinkonformationen stellen den strukturellen Raum flexibler Proteine dar und können als Eingabestrukturen für Dockingprogramme verwendet werden. Aldosereduktase, Dihydrofolatreduktase und HIV-1 Protease, sind jeweils wichtige Wirkstofftargets und gleichzeitig bekannt für ihre großen Proteinrückgratsbewegungen. Unsere Meth-

ode ist anhand dieser drei Proteine getestet worden und es wurde jeweils eine Proteinkonformation gefunden, die sehr ähnlich zu einer der bekannten Proteinstrukturen ist, die in der PDB verfügbar sind. Eine weitere Methode, die wir entwickelt haben, ist ein detailliertes Protokoll für Molekulardynamische Simulationen und Bindungsaffinitätsberechnungen, das die Bindeaffinität gedockter Liganden erneut bewertet. Unsere Methode kann nach einem erfolgreichen Protein-Ligand Dockingdurchlauf angewendet werden, um zwischen gut und schlechter bindenden Inhibitoren zu unterscheiden während gleichzeitig die Flexibilität berücksichtigt wird. Ein Urkokinaseligandendatensatz ist getestet worden, wobei unsere Methode die Korrelation zwischen experimentellen Bindungsaffinitäten und Dockingwerten von 0,57 auf 0,92 verbessert. Unsere Methode kann demnach die Rangkorrelation von Dockingergebnissen erheblich optimieren, insbesondere, wenn sie auf strukturell ähnliche Liganden angewendet wird.

Acknowledgements

It would not have been possible to write this thesis without the help and support of many wonderful people, to only some of whom it is possible to give particular credit here. Over the past couple of years I learned a lot, scientifically and personally, and I am deeply grateful to everyone who accompanied me during that time.

First and foremost, I want to thank my advisor Prof. Oliver Kohlbacher for giving me the opportunity to explore a fascinating research area. I am forever indebted to Oliver for his caring supervision, encouraging words, and devotion to teaching I will always keep as standard. He did not only guide me through scientific problems, but also through life changing situations.

I am very much obliged to Prof. Klaus Harter for going to the time and effort of reviewing this thesis. Without the research at his department the chapter about protein-DNA complexes would not have been possible.

Furthermore, I am very grateful to all my collaborators, in particular to Luise Brand and Dierk Wanke. It is an incredible privilege working together with you.

I would like to express my gratitude to all current and former members of the Kohlbacher Lab for thinking company, care, and support. Special thanks go to: Annette Höglund, who introduced me to the beauty of protein structures; Charlotta Schärfe, Sebastian Briesemeister, and Sandra Gesing for sharing the office and many ups and downs with me; Lena Feldhahn and Sven Nahnsen for their special friendship; Charlotta Schärfe, Lena Feldhahn, Marc Röttig, and Wolfgang Schneider for their incomparable companionship and their shared interest in teaching. I am very grateful to all my former students for their hard work, inspiration, and motivation and would like to emphasize the work of Christopher Mohr, Mirco Michel, and Wolfgang Schneider. The past years at the Center for Bioinformatics Tübingen would have not been as enjoyable without many great people, in particular Simone Linz with whom I spent invaluable lunch hours. I would also like to thank all 'WSI-Forscherinnen' for laughter, openness, and encouragement, especially Kay Nieselt who is keeping us inspired. Furthermore, many thanks go to Claudia Walter for all administrative issues and to the systems administrators Muriel Quenzer, Jan Schulze, and Werner Dilling for maintaining the cluster systems.

I utterly appreciated all valuable comments on this manuscript by Oliver, Dierk, Philipp, Mirco, Chris, Charlotta, and Kathi. I am also very much obliged to David van der Spoel for his patience and providing working conditions so that I could finish this thesis while having started in his group.

Last but not least, I am grateful beyond what words can describe to my family and closest friends, who have supported me throughout my life and in particular during the last years, both by keeping me harmonious and reminding me of what really matters in life—I always feel loved—Thanks!

In accordance with the standard scientific protocol, I will use the personal pronoun *we* to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

1 Introduction **1**

2 Biological background **9**

2.1 Structural characteristics of proteins, DNA, and small molecules 10

2.1.1 Proteins 10

2.1.2 DNA 12

2.1.3 Small molecules 14

2.2 Interactions and other factors at binding interfaces 14

2.2.1 Electrostatic interactions 15

2.2.2 Hydrophobic interactions 16

2.2.3 The role of water 16

2.2.4 Binding free energy and entropic contribution 17

2.3 Protein-DNA complexes 18

2.3.1 Gene regulation 18

2.3.2 Transcription factors and their DNA binding motifs 19

2.3.3 Specific features of protein-DNA interactions 20

2.4 Protein-ligand complexes 21

2.4.1 Lock-and-key principle 21

2.4.2 Induced fit and flexibility within protein-ligand complexes 21

3 Theoretical background - computational chemistry **23**

3.1 Molecular mechanics 23

3.2 Molecular dynamics 25

3.3 Solvents in molecular modeling 25

3.3.1 Explicit solvent 25

3.3.2 Implicit solvent 27

3.4 Poisson Boltzmann calculations 27

3.5 In silico binding affinity predictions 29

4 Studying protein-DNA complexes at the atomic level **31**

4.1 Introduction 31

4.2 Modeling and refining protein-DNA complexes using molecular dynamics 35

4.2.1 Introduction 35

4.2.2 Materials and Methods . . . 36
4.2.3 Results . . . 40
4.2.3.1 Modeling arbitrary protein-DNA complexes . . . 40
4.2.3.2 Advanced molecular dynamics simulation protocol . . . 43
4.2.4 Discussion . . . 48
4.3 Studying specific features at binding interfaces of WRKY-DNA complexes . . . 51
4.3.1 Introduction . . . 51
4.3.2 Materials and Methods . . . 53
4.3.3 Results . . . 55
4.3.3.1 Structural analysis of *At*WRKY1 cDBD-DNA structures . . . 55
4.3.3.2 Binding specificities of *At*WRKY11 and *At*WRKY50 . . . 56
4.3.3.3 Structural details of *At*WRKY33 cDBD- and nDBD-DNA complexes 58
4.3.4 Discussion . . . 59

5 Studying flexibility of protein-ligand complexes 61
5.1 Introduction . . . 62
5.2 Representing major movements in protein-ligand docking . . . 63
5.2.1 Introduction . . . 63
5.2.2 Materials and Methods . . . 66
5.2.3 Results . . . 74
5.2.3.1 Conformational diversity . . . 74
5.2.3.2 Recurrent conformational changes . . . 79
5.2.3.3 Docking into structural representatives . . . 81
5.2.4 Discussion . . . 83
5.3 Rapid molecular dynamics simulation protocol for re-scoring docked ligand poses . . . 86
5.3.1 Introduction . . . 86
5.3.2 Materials and Methods . . . 87
5.3.3 Results . . . 90
5.3.3.1 Rapid molecular dynamics simulation protocol . . . 90
5.3.3.2 Optimized MM-PB/GBSA re-scoring parameter set . . . 91
5.3.3.3 Binding free energies for urokinase-ligand complexes . . . 91
5.3.4 Discussion . . . 93

6 Conclusion 97

A Abbreviations 102

B First Appendix 104

C Second Appendix 114

D Third Appendix **122**

E Fourth Appendix **128**

F Contributions and Publications **132**

1 Introduction

It is impressed on our minds in infancy that a certain arbitrary symbol indicates an existing fact; [...] until we throw all our preconceived impressions on one side, and seek the truth by independent observations from Nature herself. [210]

Eadweard Muybridge, 1957
explaining the discovery
detected in his study
The Horse in Motion

Motivation

In 1962, Max Perutz and Sir John Kendrew won the Nobel Prize in Chemistry for their research on protein structures. They determined the first well-defined, three-dimensional model of a protein at atomic detail using X-ray crystallography [142–145, 231, 232]. X-ray crystallography is still today a prominent method to obtain protein structures. With their findings they prepared the ground for studying molecular mechanisms of how biomolecules work in intricate detail. Crystal structures of proteins, protein-DNA complexes, or protein-ligand complexes can be analyzed to gain information about interactions between proteins and DNA or between proteins and ligands. Moreover, influences of mutations on the structure of proteins or other relevant structural characteristics can be studied at an atomic level. These structural insights reveal important details for understanding biochemical processes and functions of proteins.

One often neglected aspect of biochemical processes is the structural flexibility of the proteins involved. Imaging the structure of a specific protein during a dynamic process is comparable to the task Eadweard Muybridge tackled in 1878 to prove a certain position of a horse while galloping [209]. His study *The Horse in Motion*, see Fig. 1.1, was an intermediate step toward motion pictures. He used multiple cameras to capture motion in stop-action photographs. With that he could show an image of a horse at a certain point in time with all four hooves off the ground. This fact was not

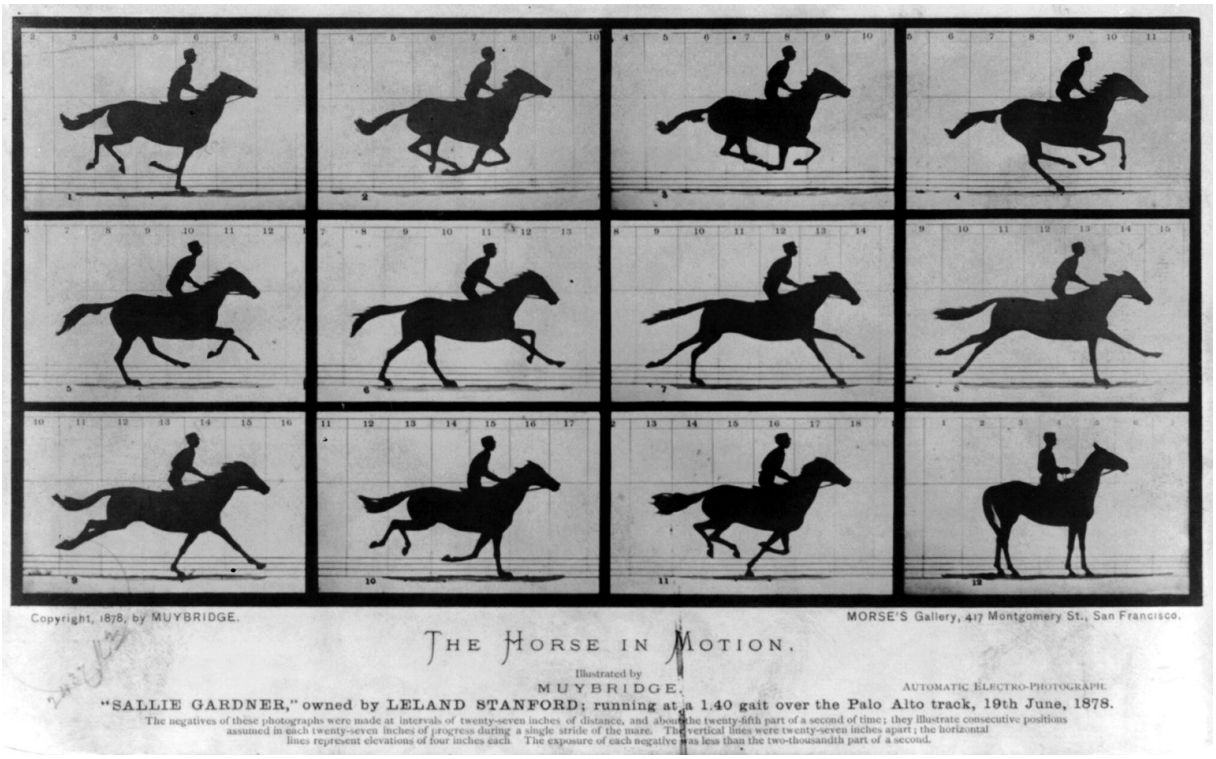


Figure 1.1: The first motion pictures. Eadweard Muybridge proved with his study *The Horse in Motion* that all four hooves of a horse are off the ground at the same time while galloping. These first motion pictures detect such an image while visualizing a dynamic event [208].

observed or portrayed by painters before, since it is almost impossible to detect a snapshot visible only for such a short period in time during a dynamic event.

As motion pictures illustrate dynamic events, molecular dynamics (MD) simulations can be employed to understand the inherent flexibility of biomolecules and thus complement the static structures obtained from X-ray crystallography. Crucial structural rearrangements of molecules can be explored computationally through MD simulations. Thereby, novel structural conformations of proteins, protein-DNA complexes, or protein-ligand complexes can be investigated. Analyzing MD simulation results yield a more complete understanding of the structure, flexibility, and conformational changes of proteins and specific interactions at protein-ligand or protein-DNA interfaces. Insights gained from these flexibility analyses complement rational drug design studies and structural investigations examining the regulation of gene expression.

Protein-DNA interactions

Regulating gene expression plays a central role in cellular events as it defines, amongst others, developmental processes. Proteins binding to target sequences on DNA are called transcription factors (TFs) and are essential for regulating gene expression. The interplay between TFs and their binding sites (TF binding sites, TFBSs) is usually complex and more often than not, multiple TFs act in concert to ensure correct spatial and temporal expression of multiple target genes [91]. Especially, understanding the function of proteins that participate in this process and the function of genes they regulate is of tremendous interest. Various approaches are used for characterizing gene function and identifying individual TFBSs at the sequence level. The most detailed picture is provided by the structural level, which gives insight into specific interactions between these DNA-binding proteins and the DNA double helix [137, 181, 287].

In order to understand gene regulation, a combination of experimental and theoretical approaches is frequently used that try to answer the following questions: Which is the TF's preferred binding site? Where do these TF binding motifs lie along the genome? Which genes are regulated by the TF? When investigating in more detail the specificity and especially the structural features of protein-DNA complexes we focus on the following issues: How are specific interactions formed between the TF and the DNA double helix? How do such interactions alter the structure of both TF and DNA? How do protein-DNA contacts account for specificity in detail? Answers to these questions about structural details can explain observations at the sequence level and even the functional level of gene expression.

In vitro methods have been revised in terms of efficiency and accuracy during the last years. State-of-the-art techniques range from large-scale experiments screening huge DNA sequence libraries for TFBSs to time-consuming small-scale experiments such as X-ray crystallography providing atomic details of TF-DNA interactions. Results of these experiments, as gene expression data, genomic sequences, and structural data are collected in databases [19, 20, 25, 234]. These data can then be used for the development of computational (*in silico*) methods. In theory, the more experimental binding site data for a certain TF is available, the more statistically significant is its proposed binding site model. However, in practice, artifacts such as insufficient sampling of DNA sequence space introduce often noise or biases into these TFBS models [168, 276]. Through tight feedback, verification, and close cooperation between different disciplines it is possible to facilitate more accurate approaches on the experimental as well as on the computational site.

Computational approaches can be assigned either to sequence-based or structure-based methods. Both kinds of methods try to represent binding preferences between proteins and DNA. The most basic TFBS model is based on sequence information alone and is termed consensus sequence [276]. It is usually obtained by aligning known TFBSs and presents the most likely nucleotide at each position of the binding site for a certain TF. Consensus sequences were the first sequence-based representations to be used for scanning the genome for TFBSs [88]. Position-specific scoring matrices (PSSMs) [44] are refinements of consensus sequences, since they capture the statistical occurrence of nucleotides at each position of the binding site. PSSMs possess intrinsic limitations, because independence between sequence positions is assumed by this method [16]. However, binding specificity is also achieved

through structural dynamics and chemical complementarity between TFs and their binding motifs. In order to gain a more complete picture of TF binding sites, it is therefore important to study the interactions underlying sequence- as well as structure-specific recognition between proteins and DNA.

Knowledge-based structural methods analyze protein-DNA co-crystal structures to study interactions between protein side chains and DNA base pairs at the binding interface [154, 186]. These methods yield insights into the complex network of interactions between proteins and DNA. However, they do not provide information on the dynamics of interactions or on flexibility of molecular binding partners. Molecular modeling thus holds great promise for addressing the issues where sequence- and non-dynamic structure-based methods fail. With MD simulations flexibility at the protein-DNA interface can be studied providing more information than rigid structures [76, 118]. These flexible complexes can be considered a more truthful representation of the biophysical reality of interactions [153]. In addition to structural re-arrangements, MD simulations study the effects of interdependence [312], hydration at the protein-DNA interface [66, 199], direct and indirect interactions, and mutations of amino acids and base pairs at the protein-DNA binding interface on the binding affinity. Existing methods are, however, often limited with respect to available protein-DNA complexes and implicit solvent MD simulations or constrained protein and DNA backbone atoms. If no experimental protein-DNA co-crystal structure is available, protein-DNA docking approaches can construct such a complex. Since it is a challenging problem to find the correct placement of the protein on the DNA strand, protein-DNA docking techniques have limitations. We describe an alternate technique in order to be able to study interactions at the protein-DNA binding site. In the context of elucidating protein-DNA complexes, we approach unanswered questions in this thesis. These range from modeling accurate protein-DNA complexes to performing non-constrained explicit solvent MD simulations on protein-DNA complexes. Thereby, we focus on interactions at the binding interface, their flexibility, and their specificity.

WRKY proteins comprise a large family of proteins in the plant kingdom and many questions remain open about their binding specificity. Only two protein structures [71, 308] and one protein-DNA complex [309] are stored in the protein databank (PDB). Therefore, structural information of different WRKY proteins and characteristics about specific interactions at the binding interface is not available. Hence, we introduce a molecular modeling protocol to model yet experimentally unresolved protein-DNA complexes. The resulting WRKY-DNA complex is in excellent agreement with the (only recently published) nuclear magnetic resonance (NMR) solution structure [309]. The protocol is thus able to build WRKY-DNA complexes with excellent accuracy. In the last section of Chapter 4, we focus on the analysis of specific interactions and dynamics at the binding interface of different WRKY-DNA complexes. *At*WRKY11 and *At*WRKY50, two representatives of the WRKY family are analyzed. Especially, their unique features and differences are studied at the molecular level using explicit solvent MD simulations which provide for the first time insights into which amino acids specifically interact with the DNA. We identified one amino acid important for specificity and can thereby offer a conclusive DNA-binding specificity model for the WRKY protein family.

Protein-ligand complexes

Specificity is not only an important aspect within protein-DNA binding mechanisms, but also crucial in structure-based drug design. Structure-based drug design involves several steps ranging from protein target identification, hit and lead discovery, lead optimization to clinical studies. We focus in this work on the step of identifying a small chemical molecule that binds and inhibits a target protein. Typically, in pharmaceutical technology large chemical databases are screened experimentally for promising hit structures. These *in vitro* screenings are expensive and their results yield little information on the atomic details of the interactions between the protein and the chemical compound. In order to reduce time and costs during this process and to be able to choose only the best compounds for clinical studies a lot of experimental and computational methods have been developed during the last years. One structure-based *in silico* method to identify small molecules which inhibit the target protein is protein-ligand docking. Docking methods try to find the best placement of small chemical molecules, also called ligands in this context, within the binding pocket of a protein and a respective binding affinity. Several placements of the same ligand are possible, whereby each resulting placement is termed ligand pose. Algorithms in docking methods are either using stochastic search strategies, an example is *AutoDock* [207], or systematic algorithms, e.g., incremental construction strategies as realized in *Glide* [89]. Both search strategies find usually good ligand poses in a short period of time. A scoring function calculates the binding affinity of each ligand pose with respect to its target protein. Various scoring functions exists [26, 74, 89, 146, 238, 240, 245] that try to represent fundamentals of biophysical interactions between ligand and protein atoms close to reality.

Since some proteins are highly flexible it is advantageous to incorporate knowledge about structural alterations while searching for new drugs. Side chain flexibility of residues within the binding pocket is represented by most docking programs. However, most state-of-the-art docking methods do not account for large internal protein motions. *FlexE* [51] is an example for a docking program which captures large backbone movements of proteins by using experimentally obtained protein structures. It docks into an ensemble of different protein conformations which represent the structural space. However, most protein conformations are yet experimentally unresolved for the majority of flexible proteins. Molecular modeling methods such as normal mode analysis or Monte Carlo sampling can also yield an ensemble of various protein structures. However, when applying MD simulations to protein crystal structures large scale structural rearrangements can be observed in a continuous manner. Thereby, transition states between specific protein conformations can additionally be obtained. In Chapter 5.1, we introduce MD simulation protocols for studying highly flexible proteins and a procedure, consisting of principal component analysis (PCA) and clustering methods, to identify distinct protein conformations, covering the most relevant regions of the protein's conformational space. In a later step these identified protein conformations can be used as input structures for docking.

As mentioned above molecular docking assesses the value of each ligand pose by determining a certain score. Scoring functions estimate binding free energies of protein-ligand complexes. They are fast, approximate, mathematical methods used to predict the strength of all interactions between the ligand and the protein. Finding the trade-off between speed and exact physics underlying these

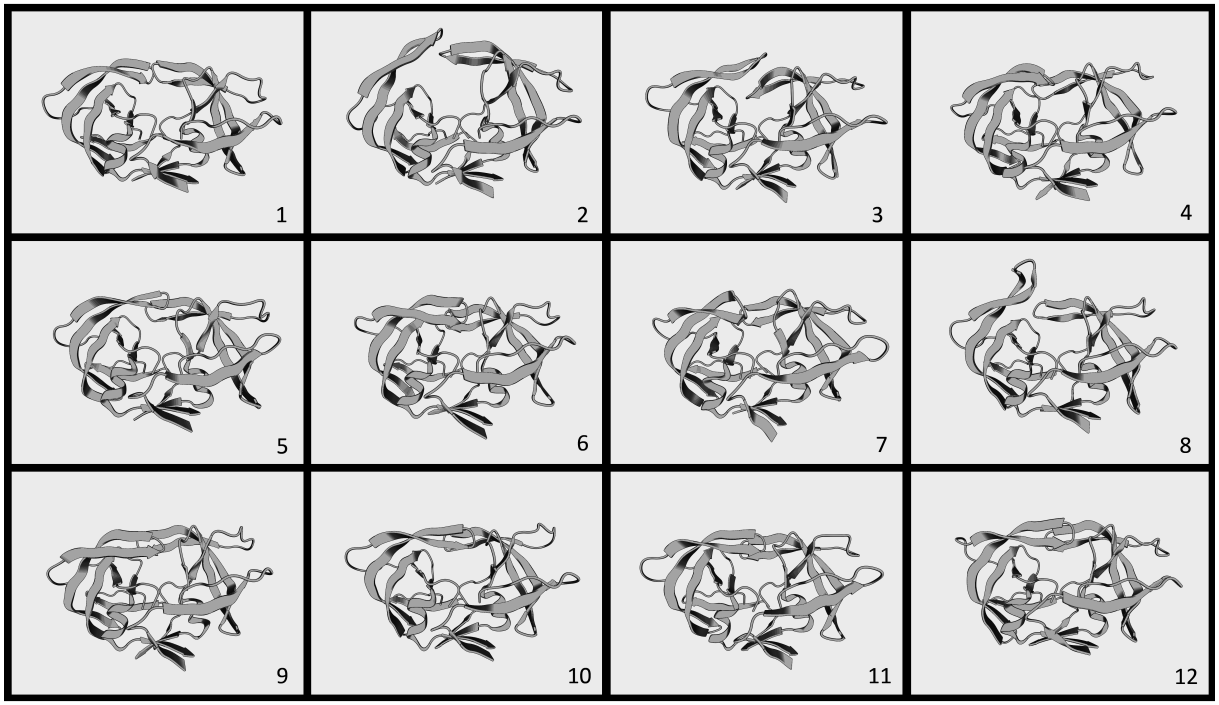


Figure 1.2: MD simulation of HIV-1 protease. HIV-1 protease (PDB id: 1hhp [270]) is a flexible protein, which undergoes conformational changes. During an MD simulation the flap regions of HIV-1 protease open up (2) and subsequently return to a "closed" conformation (6), before they open up again (8), and remain closed (9-12) in the last part of the simulation. Static crystal structures of HIV-1 protease are available in the PDB in the "closed" (6, 9-12) and "semi-open" (1) conformations, but not for intermediate states.

interactions is challenging. Additionally, it needs to be pointed out that physico-chemical interactions at binding interfaces are not yet completely understood and thus, cannot be described entirely by computational methods. Often, non-active compounds can be found at the top of docking hit-lists, whereas highly potent molecules might not be determined as good binders by docking methods. To improve such docking results more advanced scoring methods can be applied on a fraction of the best computationally screened compounds. Since the number of compounds was reduced, methods which model physico-chemical interactions at the binding interface with higher accuracy and have thus higher computational costs can be used. We developed such a re-scoring protocol based on a force field technique, which is presented in Chapter 5.3. Force fields represent intermolecular interactions superior than common scoring functions available in docking programs. In our re-scoring protocol we account also for flexibility using an implicit solvent MD simulation. Compared to the work of others [33, 34, 126, 159], our protocol performs with higher accuracy and is comparatively fast.

Structure of the thesis

This thesis is structured into six chapters. Following this introduction, the biological and computational background is introduced in Chapters 2 and 3, respectively. Chapter 4 describes and evaluates the methods and techniques we developed for modeling and invoking flexibility in protein-DNA complexes. Subsequently, our framework for protein flexibility studies, identifying the most prominent protein conformations, as first step toward developing a flexible docking method is presented in the first part of Chapter 5. In the second part of Chapter 5 our protocol that includes dynamics in rescoring docked ligand poses is described. Chapter 6 provides a general conclusion of the presented work. It highlights especially how MD simulations and thereby flexibility affects results studying protein-DNA and protein-ligand complexes.

2 Biological background

The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicated by any theory of protein structure. [144]

John C. Kendrew, 1958
reporting on myoglobin, the
first solved protein structure

The biological background of this thesis covers four parts. The first section highlights specific features of the molecular structure of proteins, DNA, and small molecules. Since an in-depth introduction about biochemical fundamentals is beyond the scope of this thesis, we refer to textbooks for a more complete coverage of this topic. However, knowledge about particular characteristics concerning these biomolecules is necessary to understand the issue of this work.

The second section describes the main interaction-types, the role of water, and the entropic effect at the interface of two biomolecules.

The third section contains a short introduction to gene regulation and gives insight into the biological importance of its two main players: DNA and proteins. The main function of DNA is to store, replicate, and propagate genetic information. Proteins, on the other hand, have a wider range of different structures and functions, whereby certain proteins are involved in regulating gene expression. Specific interactions between proteins and DNA are essential for gene regulation and their distinctive features will be described in detail.

Analogous to protein-DNA complexes, specific interactions at the binding interface of protein-ligand complexes contribute to the binding affinity. Small molecules that bind to proteins have specific features as well as the proteins they bind to. The fourth section mainly focuses on a comparison between out-dated and still valid assumptions to describe the binding process of protein-ligand complexes.

Parts of Section 2.1 and 2.3 have been submitted for publication [82].

2.1 Structural characteristics of proteins, DNA, and small molecules

2.1.1 Proteins

Amino acids are the fundamental building blocks of proteins. In general, 20 natural distinct amino acids are present in proteins. They differ in their respective side chain, which determines their physico-chemical properties. Based on their side chain, amino acids are often assigned to one of four categories: weak bases, weak acids, hydrophilic amino acids, and hydrophobic amino acids. The property of a side chain also determines the secondary structural preference of a protein's residue. All amino acids have two functional groups in common: an amino group (NH_2) and a carboxylic acid group (COOH). These two groups are linked together by one central carbon atom, termed C_α atom. To this carbon atom the eponymous side chain is also connected. Amino acids are encoded by two different naming schemes: the three-letter code and the one-letter code. The overall structure of proteins is organized at four distinct levels: the primary, secondary, tertiary, and quaternary structure.

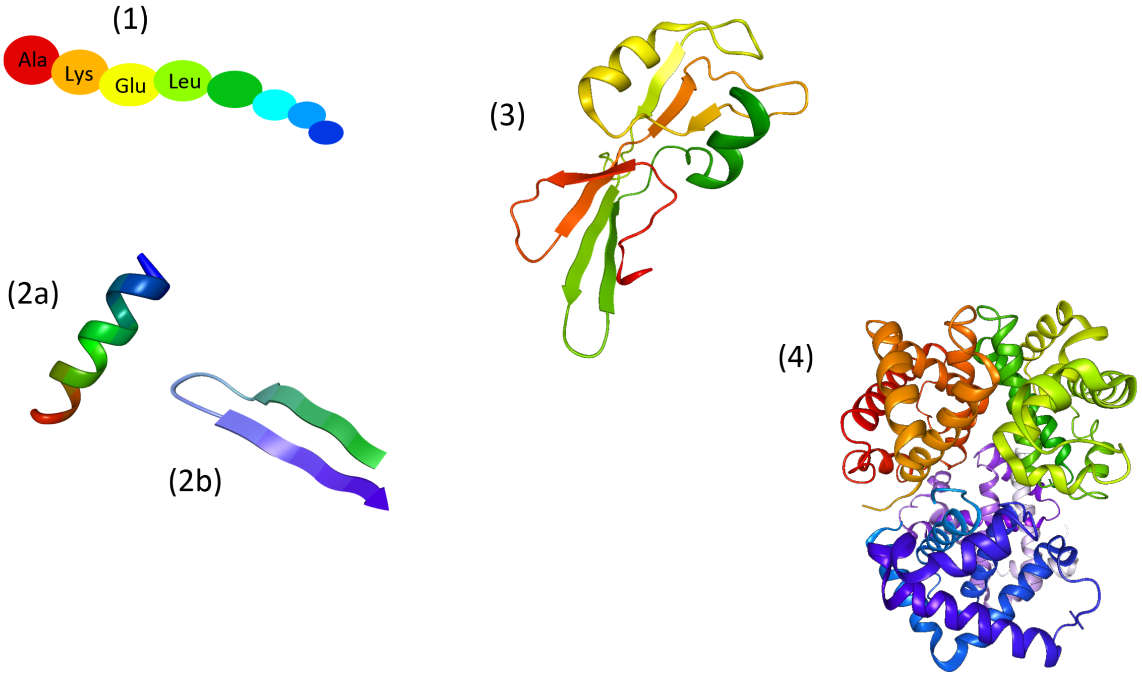


Figure 2.1: The protein structure is organized at four distinct levels. (1) The primary structure represents the amino acid sequence of the protein: Ala, Lys, Glu, Leu, and so on. Secondary structural elements of proteins are mainly α-helices (2a) and β-sheets (2b). The tertiary structure (3) is the overall fold of one protein chain (PDB id: 1odh) [53]. Quaternary protein structures (4) exist when multiple protein chains, in this case four protein chains (PDB id: 2hhb) [80], aggregate into larger complexes.

Primary structure

The protein's primary structure is the sequence of amino acids forming a polypeptide chain. Two amino acids are linked together by a peptide bond, which is formed between the nitrogen atom of the amino group (NH_2) of one amino acid and the carbon of the carboxylic acid group (COOH) of another amino acid. When two amino acids form a peptide, one water molecule is released, and what remains of each amino acid is called residue. The amino-terminal end of the protein's polypeptide chain is called N-terminus and the carboxy-terminal end C-terminus. Three atoms of the polypeptide chain form the protein's backbone: the N (the amide N), the C_α, and the C (the carbonyl C) atom. The order and composition of amino acids at the sequence level influences inherently a protein's three-dimensional structure.

Secondary structure

Repeating spatial segments in proteins are mainly either α-helices or β-sheets, which are connected by loops. When analyzing these secondary structural elements in more detail they can be differentiated to more specific subgroups. In general, all secondary structural elements are formed by hydrogen bonds (H-bonds). In α-helices, H-bonds are formed between backbone atoms of successive residues of the same chain. Each backbone amide N donates its hydrogen atom to the backbone carbonyl group of the amino acid four positions earlier. β-sheets consist of two or more β-strands connected laterally by backbone H-bonds. The majority of β-sheets form either parallel or antiparallel arrangements. When two β-strands are arranged in parallel the N-termini are oriented in the same direction, in antiparallel arrangements the successive β-strands show alternate directions (see Fig. 2.1 (2b)). The detailed H-bond pattern is different for parallel and antiparallel β-sheets. In anti-parallel β-sheets, two H-bonds are formed between two amino acids on opposite strands, when their backbone NH and CO groups are facing each other so that each backbone amide N can donate its hydrogen atom to the backbone carbonyl of the other strand. When in parallel β-sheets two amino acids i and j of different β-strands are adjacent, residue i may form H-bonds with residues $j-1$ and $j+1$. Residue j may form H-bonds with other residues or with none at all.

Tertiary structure

The overall three-dimensional fold of a single polypeptide chain is called tertiary structure. It includes all secondary structural elements and forms the specific three-dimensional structure of a protein. This structure also determines the protein's function. Internal amino acids form contacts and stabilize the overall shape of the protein. H-bonds between polar side chains, ionic bonds between charged side chains, hydrophobic interactions between non-polar side chains, and even covalent bonds between two cysteine side chains (disulfide bond) can be found.

Quaternary structure

Multiple polypeptide chains can aggregate to larger complexes. These quaternary structures are built

out of several three-dimensional tertiary protein structures. The interactions that hold different tertiary protein structures together are the same interactions that are present internally in tertiary structures. Hemoglobin (see Fig. 2.1 (4)), the oxygen-transport protein present in red blood cells consists of four polypeptide chains. The aggregation of these four tertiary protein structure enables its function.

2.1.2 DNA

Living organisms possess two different kinds of nucleic acids: ribonucleic acids (RNA) and deoxyribonucleic acids (DNA). Both RNA and DNA contain nucleotides as building blocks. Each nucleotide is built out of three different structural components. Two of these structural units are a sugar- and a phosphate-moiety, whereby DNA contains a deoxyribose and RNA contains a ribose as sugar. The third component, the base, is linked to the sugar ring and exhibits four distinct types in DNA: adenine (A), guanine (G), cytosine (C), and thymine (T). These bases belong either to the purine (A and G) or the pyrimidine group (T and C). One purine A forms two H-binds with one pyrimidine T and one purine G forms three H-bonds with one pyrimidine C, see Fig. 2.2. These interactions build base pairs (bps) between A and T and G and C. The edges of the bases reaching into the minor and major groove of the DNA form a specific pattern of H-bond acceptors and donors that can be recognized

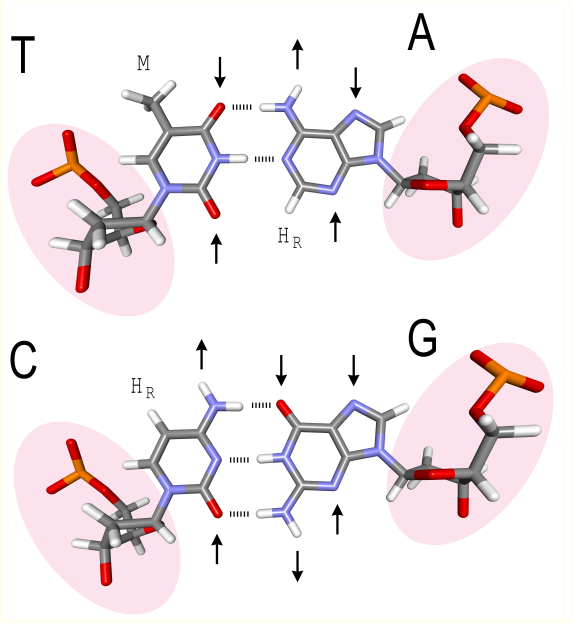


Figure 2.2: Characteristics of DNA base pairs. Intermolecular H-bonds (dotted lines) in cytosine-guanine (C-G) and thymine-adenine (T-A) bps, stabilize the DNA double helix. The bp edges form a pattern of H-bond acceptors and donors (indicated by arrows) which can be recognized by amino acid side chains of proteins. The letter M marks the methyl group of thymine (T) and H_R indicates ring hydrogen donors. The major groove of the double helix is oriented upwards, whereas the minor groove is oriented downwards. The chemical composition of the DNA sugar-phosphate backbone is constant and independent of the bp sequence and is placed at the outer left and the outer right side. Figure based on Fig. 1 in [82].

by amino acid side chains of proteins [261], see Fig. 2.2 for an illustration. The polar characteristic of the DNA double helix results primarily from the negatively charged sugar-phosphate backbone, which is independent of the bp sequence.

The shape of the famous Watson-Crick double helix [298] arises from two single DNA strands where pairs of complementary bases form intermolecular H-bonds. As the nucleotide strands are not directly placed opposite of each other, two differently sized grooves emerge. In the ordinary B-DNA the major groove is 22 Å wide, whereas the minor groove is 12 Å wide, see Fig. 2.3. Double-stranded DNA can

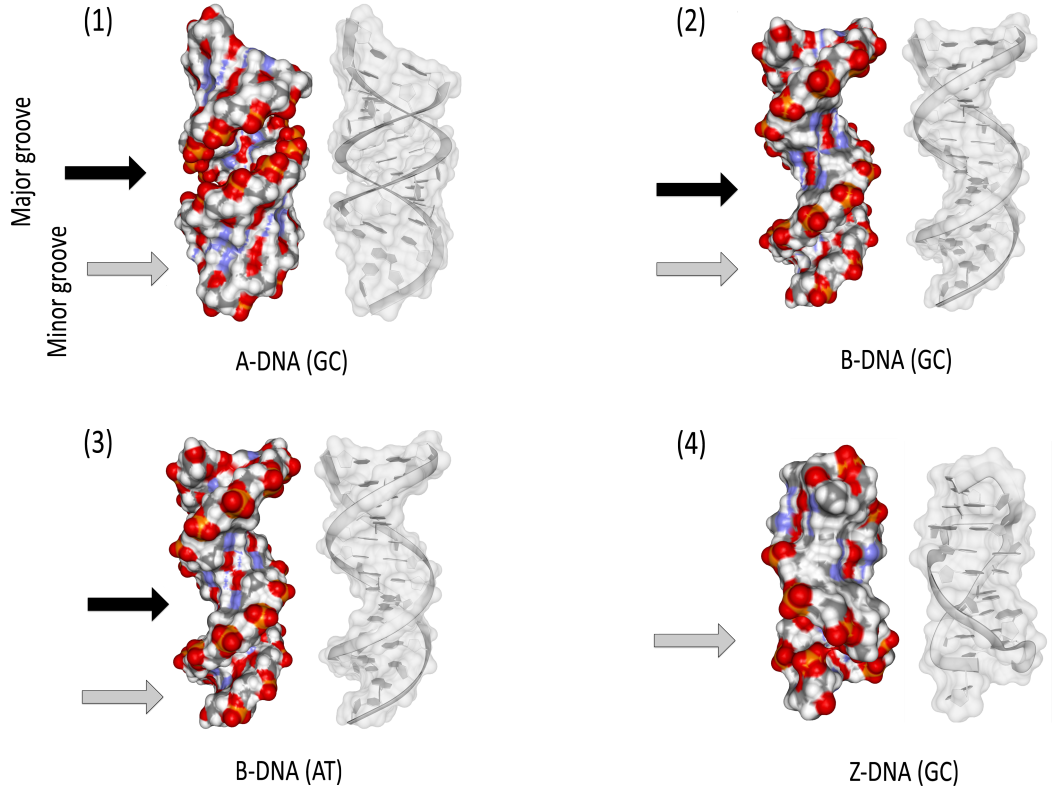


Figure 2.3: Three-dimensional conformations of A-DNA, B-DNA, and Z-DNA helices. Each DNA helix comprises 14 bps and was constructed with the Nucleic Acid Builder (*NAB*) integrated in the *AMBER* software package [38], except the Z-DNA helix. The coordinates of the Z-DNA were obtained from the PDB (PDB id: 1r4d [47]). Each molecular shape is colored with respect to its atoms (oxygen atoms are colored red, nitrogen atoms are colored blue, and carbon atoms are colored gray) to approximate the electrostatic potential. Additionally, the DNA helix is displayed as cartoon representation. (1) A-DNA with a narrow, deep major groove, and a wide, shallow minor groove. The model consists of the alternating sequence of G-C bps. (2) B-DNA with a wide, shallow major groove and a narrow, deep minor groove, composed of alternating G-C bps. (3) B-DNA with an alternating sequence of A-T bps. Since the models were built using *NAB*, they do not reflect sequence-dependent minor or major groove widths. However, the differences of the electrostatic potential between GC and AT B-DNA are indicated. The AT B-DNA exhibits predominantly neutral thymine methyl groups in the major groove, colored gray, as opposed to less neutral atoms in the major groove of the GC B-DNA. (4) Z-DNA lacks a major groove and its minor groove is negatively charged, narrow, and deep. Figure based on Fig. 1 in [247] and Fig. 2 in [82].

undergo conformational changes that distort the classical Watson-Crick double helix. These three-dimensional geometric variations in the double helix are mainly dependent on the primary nucleotide sequence. Stretches with C-G bps are more stable and rigid since they are stabilized by three H-bonds, in contrast to two H-bonds between A-T pairs. The structure of the helix can deviate from the ideal B-DNA in local rearrangements (e.g., minor groove width and DNA kinks) or in an overall global shape (e.g., DNA bending, A-DNA, and Z-DNA). Different structural characteristics of A-, B-, and Z-DNA are illustrated in Fig. 2.3. DNA regions of at least three consecutive ApA (an A-T bp is directly followed by another A-T bp, connected by the phosphate group), TpT, or ApT bp steps narrow the minor groove of the DNA, thus enhancing local negative electrostatic potentials [116, 155, 211, 246, 247]. In DNA sequences where CpA and TpA dinucleotides are predominantly present, so-called Hoogsteen bps exist in thermal equilibrium with standard Watson-Crick bps [120, 214]. Hoogsteen bps are normal Watson-Crick bps, predominantly A and T bps, which are flipped over like an upside down step of a 'normal' ladder. The sugar-phosphate backbone of the helix remains stable, whereas the two bases are turned upside down. Thereby, the double helix structure adopts a Z-DNA conformation.

2.1.3 Small molecules

A small molecule is a low molecular weight organic compound, also called ligand in this context. Ligands can bind to a target or receptor, e.g., a protein. In this work we focus on interactions between ligands and proteins. Small molecules are not only able to form interactions with a protein, but are also able to modulate the protein's activity or function. Ligands which serve as drugs against a certain disease have additional characteristics; some of these features are described in the following. The upper molecular weight limit is approximately 500 Da, which allows drug molecules to rapidly diffuse across cell membranes. Therefore, they are usually able to reach intracellular binding sites and their place of pharmacological action. These compounds are often either weak acids or bases and contain multiple titratable functional groups. Their charged state is dependent on the pH of the surrounding solvent or on the properties of protein residues in their vicinity. Small molecules used as drugs are mostly rigid. They can bind with higher affinity in the protein's binding pocket than flexible ligands since they are not able to loose as many degrees of freedom upon binding. Hence, binding of rigid ligands is entropically slightly more favorable than binding flexible ones. Other important features which characterize a well-effective drug, e.g., bioavailability and non-toxicity, are not covered in this work.

2.2 Interactions and other factors at binding interfaces

Proteins form complexes with DNA in order to regulate gene expression. This binding process between transcription factors (TFs) and the DNA double helix is always reversible. Whereas, organic compounds can bind to proteins both by forming a chemical bond and by non-covalent interactions. For example, acetylsalicylic acid (the active agent in aspirin) or omeprazole (the active agent in omeprazole) react chemically with their target proteins, thereby forming a covalent bond. However,

we focus on ligands which form non-covalent interactions with their target proteins in the following. Non-covalent interactions are reversible and differ in their nature, strength, optimal orientation, and distance. There are basically two types of interactions: electrostatic and hydrophobic interactions. In this section we highlight basics about these interaction types as well as the role of water and the role of entropy in binding.

2.2.1 Electrostatic interactions

Electrostatic interactions can be described by Coulomb's law, since they result from the existence of two or more charged particles in a system. Coulomb's law states that equally charged particles (with respect to their sign) repel, whereas oppositely charged particles attract each other. More precisely, Coulomb's law contains the following: the magnitude of the electrostatic force is proportional to the product of electrical point charges (q_i and q_j) and inversely proportional to the distance between these charges (r_{ij}).

Ionic bonds are among the strongest non-covalent interactions. An ionic bond, also called salt bridge, is formed by favorable electrostatic interactions between two oppositely charged ions and can account for up to 5 kcal/mol per bond of the binding energy. The distance between these two ions is typically about 2.7-3.0 Å. Salt bridges often arise in proteins between the anionic carboxylate of either aspartic acid (Asp) or glutamic acid (Glu) and the cationic ammonium of lysine (Lys) or the guanidinium of arginine (Arg). Other amino acids such as histidine (His), tyrosine (Tyr), and serine (Ser) can also form salt bridges depending on outside factors affecting their pK_a's. At protein-DNA binding interfaces positively charged residues (Asp and Glu) are able to form ionic bonds with the negatively charged DNA backbone.

H-bonds are almost as strong as ionic bonds and can also account for up to 5 kcal/mol per bond. An H-bond is formed between a proton donor and acceptor group. The strength of an H-bond is highly dependent on the distance and angles between the involved atoms. The donor group contains a strong electronegative charged heteroatom, such as O, N, or F, and a polar hydrogen atom. This positive charge can interact with a lone pair of another heteroatom, which becomes the proton acceptor group. The optimal distance between the proton donor and acceptor atom in water is 2.8 to 3.2 Å. Optimal angles are 150 to 180° between the proton donor, the hydrogen, and the protein acceptor atom, and 100 to 180° between the hydrogen atom, the proton acceptor, and the atom attached to the proton acceptor atom. H-bonds can also be formed with water molecules. If H-bond donor and acceptor groups lie too far apart to interact directly, a water molecule can bridge the distance between them. These water-mediated H-bonds can cause some water molecules to become trapped at the interaction interface of protein-DNA or protein-ligand complexes.

Van der Waals interactions are typically weaker than both ionic and H-bonds, while contributing with about 1 kcal/mol to the binding energy. These interactions can only be established between two molecules upon binding, as they rely on dipoles which need to be induced temporarily at a particular

moment. Polar molecules carry for instance positive partial charges and cause a negative polarization of a non-polar molecule.

Cation-π or π-π interactions occur between groups with delocalized π-electrons. Cations offer a positive charge and thus can interact with delocalized π-electrons of aromatic rings or other π-conjugated systems. Aromatic rings are able to interact with each other by π-π-stacking, thus forming face-to-face or edge-to-face orientations. These interactions are caused by intermolecular overlapping of p-orbitals in π-conjugated systems and become stronger as more π-electrons are delocalized.

Metal interactions, which are of course not purely electrostatic, can be observed in protein-ligand complexes, when the protein contains a metal ion. Metal ions are able to form bonds with oppositely charged ligand groups or form coordinated bonds. Functional groups of ligands which are suited well for metal interaction are: thiol groups, hydroxamic acids, acidic groups, and heterocyclic nitrogen groups.

2.2.2 Hydrophobic interactions

At large hydrophobic areas on the surface of biomolecules, polar water molecules form an ordered pattern which leads to a cage-like water shell around these areas. These water molecules have restricted mobilities in contrast to water molecules in bulk solvent. When two large hydrophobic areas are located in close proximity, water molecules dissociate from their surfaces. The number of released water molecules is directly proportional to the magnitude of binding energy. Thus, the hydrophobic effect basically is based on the entropic effect.

2.2.3 The role of water

The binding process of two biomolecules in general is influenced by the surrounding solvent, which primarily is water in cellular environments. Consequently the influence of water molecules upon binding has to be taken into account. The amazingly unique properties of water molecules provide a highly mobile environment. Their dipole nature results in four possible H-bond interaction sites on the surface of water molecules. Since H-bonds between water molecules can easily be broken and established, water molecules form H-bonds among each other in a very flexible and variable way. This phenomenon reflects the highly dynamic property of water as solvent at room temperature.

Water molecules which form more than one H-bond with the protein or the DNA molecule can be detected in X-ray diffraction (XRD) structures. These spatially stable water molecules sometimes remain at the DNA surface or in protein binding pockets. At these positions they are able to form water-mediated H-bonds between the protein and DNA or the protein and a ligand. Around certain DNA sequences, they usually form specific water molecule patterns [253–255], described in more detail in Section 2.3.3. Buried water molecules deep inside of protein's binding pockets usually stay inside the binding pocket since it is energetically more favorable than migrating back to the solvent.

2.2.4 Binding free energy and entropic contribution

The energetic magnitude of a complex formation between two biomolecules can be described by the Gibbs free energy. The change in Gibbs free energy ΔG is equal to the change in enthalpy ΔH minus the entropic term:

$$\Delta G = \Delta H - T \cdot \Delta S \tag{2.1}$$

The entropic term is described by the absolute temperature T and ΔS, the change in entropy. Under standard conditions Eq. 2.1 becomes:

$$\Delta G^{\circ} = \Delta H^{\circ} - T\Delta S^{\circ} = -RTln(K_{eq}) \tag{2.2}$$

In Eq. 2.2, R is the gas constant and K_{eq} an equilibrium constant. The binding free energy contains both enthalpic and entropic contributions that are frequently thought to compensate each other [73].

In order to accurately estimate the binding free energy of a system, the entropic component needs to be taken into account. The entropy is a measure for the disorder of a system, whereupon nature tends towards maximum disorder for an isolated system. A more precise way to define entropy is to measure the multiplicity of a system associated with the state of its objects. When examining the number of possible orientations of protein side chains or the internal flexibility of ligands one can observe that there are more options to arrange protein side chains and internal bonds and angles of ligands when both molecules are free as opposed to be bound. Therefore, the multiplicity of this system is reduced upon binding, which results in loss of entropy.

The entropic contribution can be estimated as described in the following: due to complex formation of two particles, three transational and three rotational degrees of freedom are lost. This corresponds energetically to approximately $6 \cdot \frac{1}{2}RT \approx 7.5$ kJ/mol. This value is far from being exact and more advanced models exist. However, it is difficult to calculate entropic energy contributions accurately. Understanding the interplay of different factors, their interrelated processes, and their impact is the goal of various studies. Chervenak and Toone [48] examined the enthalpic influence of solvent rearrangements on the binding enthalpy of protein-ligand complexes. They found out that rearrangements of solvent molecules account for about 10% of the binding enthalpy of the investigated complexes. Singh and Warshel [267] present a restraintrelease approach to evaluate the polar and hydrophobic contributions and provide the first microscopic estimate of the magnitude of all contributions to binding entropy. However, it is questionable to simply rely on the additive formulation of all binding contributions to represent the whole picture of binding affinity. Especially, when examining enthalpic and entropic contributions separately sometimes the following can be observed: the loss of an H-bond goes hand in hand with a loss in entropy [151].

2.3 Protein-DNA complexes

In 1967, Ptashne discovered that proteins which bind to specific target sequences on DNA are essential for regulating gene expression [235]. These proteins are called transcription factors (TFs). Specific interaction between TFs and *cis*-regulatory elements, located upstream, downstream or within introns of genes along the DNA, is a key mechanism in gene regulation. The focus of the following subsections is to introduce the basic mechanism underlying transcriptional control of gene expression, the proteins responsible for this control, and specific protein-DNA interactions.

2.3.1 Gene regulation

Gregor Mendel studied inheritance in pea plants (*Pisum sativum*) in the 1860s and was the first to describe a discrete biological entity responsible for visible traits in plants. A few years later these units were called genes. Genes are DNA sequences that encode for mRNA which are amongst others directly involved in building enzymes and structural proteins of cells. The process of gene expression, by which information of a gene is decoded into its gene product is called transcription. Gene regulation is very complex, therefore, we only give a brief introduction and overview.

The complexity of gene expression in an organism correlates well with the complexity of the corresponding organism. In eukaryotes and prokaryotes not only the genome size and the number of genes vary, but also the complexity of the regulatory network controlling the transcription process. However, the underlying biochemical processes are mainly consistent for both. In prokaryotes, gene regulation is first of all necessary to adapt the organism to varying environmental conditions. On the other side, in multi-cellular organisms different cells have specific functions. In these fully differentiated cells the once successfully established complex transcription program has then a lower demand for adaptations.

Eukaryotic genes are divided into several regions. Segments which contain information for proteins (coding regions) are called exons. In between these regions lie non-coding DNA units (introns). Regulation of eukaryotic gene expression occurs at different cellular levels and locations. In eukaryotes DNA is mostly wrapped around histone proteins and thereby tightly packed to chromatin. The events leading to functional proteins comprise different steps, starting with chromatin remodeling during gene expression followed by several other steps: transcription, mRNA splicing, mRNA transport out of the nucleus, translation of mRNA, degradation of mRNA, and post-translational modification of proteins. Chromatin remodeling and modification steps are highly important in gene regulation, nonetheless they are still not well understood. We focus on transcriptional regulation, which is carried out by specialized DNA-binding proteins. This system is highly complex and forms a pivotal network of important concurrent processes, e.g., combinatorial control, synergy, activation, repression, feedback, and feed-forward loops.

Protein-DNA interactions are highly specific and enable TFs to recognize short transcription factor binding sites (TFBSs), typically located near genes in non-coding genomic DNA. These intergenic regions were first dismissed as non-functional. However, it soon became clear that these non-coding DNA segments are involved in regulatory processes and that they are responsible for activating or

silencing genes during transcription.

Usually a set of basal TFs is responsible for recruiting the transcription initiation complex. This enables the RNA polymerase II to bind to the proximal promoter, which results in the binding of the transcription initiation complex at a location called TATA box. *Cis*-regulatory regions that contain a more diverse set of binding site patterns than TATA boxes are located several kilobases away from the transcription start site. However, the flexibility of the DNA double helix enables protein-protein interactions between TF-*cis*-regulatory complexes and TF-proximal promoter complexes, thereby directing transcription.

2.3.2 Transcription factors and their DNA binding motifs

TFs typically contain a DNA-binding domain (DBD) and one or multiple interaction domains that can bind to other proteins. TFs can be classified according to the fold of their DBD [180], of which the helix-turn-helix (HTH), zinc finger (ZF), basic leucine zipper (bZip), and basic helix-loop-helix (bHLH) motifs are the most common ones. One representative of each TF classification group is shown in Fig. 2.4. As there is only a limited number of DNA-binding folds for all TFs that target additionally

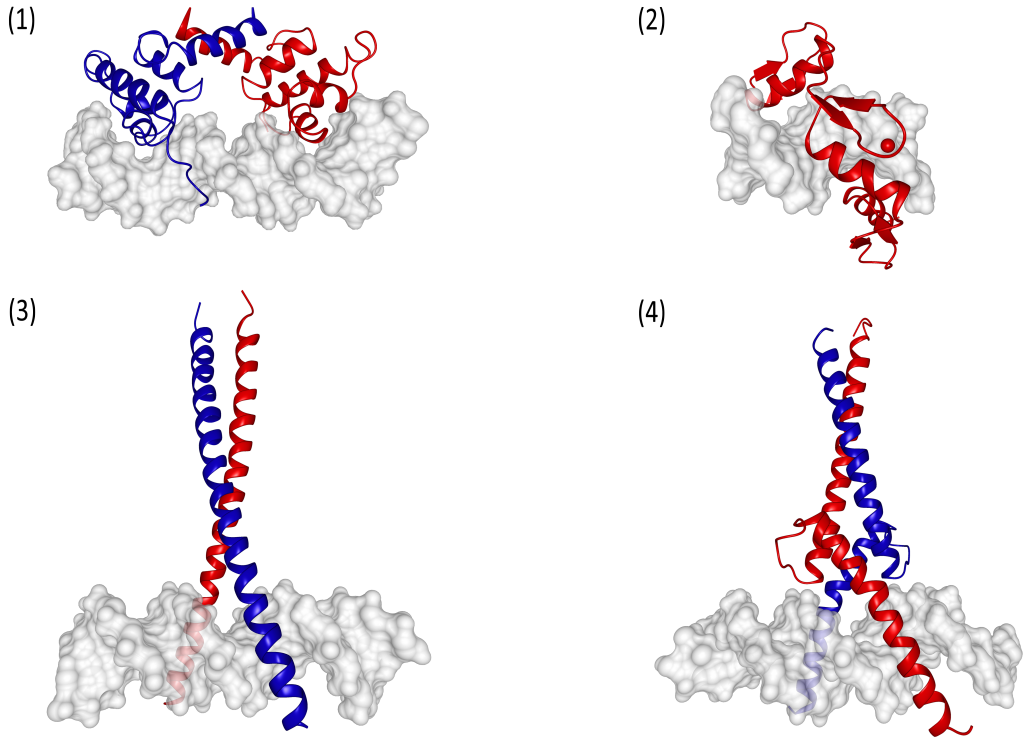


Figure 2.4: Binding domain motifs of TFs. (1) λ-repressor comprising a HTH motif, binds the DNA helix as a dimer, one monomer is colored blue and the other red (PDB id: 1lmb [14]). (2) Zif268 is a ZF protein consisting of three ZFs, the zinc ions are each coordinated by two histidine and two cysteine residues (PDB id: 1aay [75]). (3) bZip protein consisting of two α-helices, shown in blue and red (PDB id: 2h7h [218]). (4) TF MyoD consists of two bHLH motifs, colored in blue and red, respectively (PDB id: 1mdy [183]). Figure based on Fig. 3 in [82].

very specific, short, and degenerate DNA sequences, TFs have evolved a variety of mechanisms to overcome this limitation. These arrangements include DNA binding with tandem repeats of similar DNA sequence motifs or formation of homo- and heterodimers together with other proteins [265].

2.3.3 Specific features of protein-DNA interactions

Roughly two thirds of all interactions between TFs and DNA are non-specific and are formed with the sugar-phosphate backbone. They are highly important for the overall stability of TF-DNA complexes. The negatively charged phosphodiester backbone of nucleotide acids typically interacts with positively charged lysine and arginine residues along the protein interface [181].

The remaining third of all interactions is responsible for specificity. Two-thirds of these specific interactions involve complex H-bond patterns. An obvious amino acid and base preference is revealed by the distribution of H-bonds [185, 225, 226], but no generalizable recognition code can easily be established [189, 224]. H-bonds can be further classified into subgroups based on the type of interaction between amino acids and bps. There are either single interactions where only one H-bond exists between an amino acid and its corresponding base or interactions where amino acid side chains interact with bases forming two or more H-bonds. A third H-bond type exists when amino acids interact with more than one base at the same time [265].

VdW interactions are usually non-specific and formed between amino acids and the DNA backbone. Specific vdW interactions can exist and—if present—they involve methyl groups of thymine [181]. Thymine methyl groups can also form CH-π interactions with ring systems of amino acids.

Besides vdW contacts and H-bonds, water-mediated interactions can frequently be found in protein-DNA complexes. Water molecules act as contact mediators and space-fillers at the protein-DNA interface and play a key role in complex formation [132, 133, 259, 305]. Water molecules cluster at distinct hydrogen-bonding sites around the bases and are not evenly spread over the molecular surface of the DNA [253–255]. The bound solvent clusters can serve as recognition motifs for TFs. This means, an atomic description of water molecules at the interface is required for a complete molecular formulation of protein-DNA interactions. However, the main role of interfacial water molecules is to stabilize either the protein or DNA separately by solvating protein and DNA atoms and not to mediate protein-DNA interactions [94].

TFs recognize their distinct DNA binding site not only directly through specific contacts with amino acids, but also indirectly. There are numerous reports [45, 111, 157, 158, 162–164, 257] where a mutation of a base that is not in contact with any amino acid side chain was found to affect the binding affinity. The effect of varying DNA groove shapes on the electrostatic potential offers an additional mode of protein-DNA recognition. Marginal local alterations in groove shape can modulate the electrostatic potential and thus change the binding affinity. Even closely related members of the same protein family can distinguish small differences in nucleotide sequences which induce alterations of minor groove shapes and electrostatic potentials. This phenomenon was recently identified by studying Hox proteins [139] which indicates that indirect readout of DNA is important for gene regulation.

2.4 Protein-ligand complexes

In analogy to the binding process between proteins and DNA during gene expression, small molecules can bind to proteins, thereby forming protein-ligand complexes. This process is important in drug design where ligands inhibit their target protein upon binding. In the following subsections we focus on principles how this complex forming procedure can be described.

2.4.1 Lock-and-key principle

In 1894 Emil Fischer proposed the so-called lock-and-key principle [81], which describes the formation of a protein-ligand complex figuratively.

> *„Enzym und Glucosid [müssen] wie Schloss und Schlüsel zu einander passen [...], um eine chemische Wirkung aufeinander ausüben zu können.“* [81]
>
> *"Enzyme and glycoside must fit together like a key and a lock in order to initiate a chemical action upon each other."* [99]

This model describes the binding process, but implies that enzyme and substrate (or in our case: protein and ligand) are both rigid bodies that feature complementary areas on their surfaces. Additionally, it is limited to the assumption that the protein is an enzyme, where the natural substrate binds to its enzyme binding pocket. Ligands often bind to enzyme binding sites. In these cases the ligand needs to have a stronger binding affinity than the natural substrate. However, ligands can also bind to other druggable binding sites, so-called allosteric sites of proteins. In those cases the protein's structure is usually altered to change its regular function. This effect cannot be described by the simple lock-and-key model anymore. In both cases ligand binding sites are usually easily accessible lying at the protein surface. However, they bury the small ligand to a certain degree and specific interactions are formed between the ligand and protein side chains.

2.4.2 Induced fit and flexibility within protein-ligand complexes

Flexibility of both molecules is neglected as well as the ability to form water-mediated interactions when relying on the lock-and-key principle exclusively. However, it is still of fundamental importance and remains a valid assumption for many protein-protein interactions [23]. Nonetheless, the lock-and-key principle has been revised by other models. David Koshland postulated the induced fit hypothesis [156]. Here, both binding partners, in our case protein and ligand, are no longer considered to be rigid bodies. When both binding partners are in close proximity they might interact with each other in a way that induces conformational changes at their interaction sites. This induced fit phenomenon is the basis for the entire complex formation process. However, the induced fit model does not hold for all protein-ligand complexes. Additionally, structural alterations of both, protein and ligand, are

limited in a natural way since large induced fit effects have negative effects on the overall binding free energy, especially on entropy and protein-ligand specificity.

Another theory assumes that different structural conformations of one three-dimensional protein structure exist more or less side by side. One ligand can then stabilize one protein conformation upon binding. And different ligands could stabilize distinct binding site conformations of one protein upon binding. This phenomenon can be observed by the protein aldose reductase (AR), where different ligands bind to distinct protein conformations [273]. We studied the flexibility of AR in detail and results are presented in Chapter 5.2.

3 Theoretical background - computational chemistry

An atom is a body which cannot be cut in two. A molecule is the smallest possible portion of a particular substance. No one has ever seen or handled a single molecule. Molecular science, therefore, is one of those branches of study which deal with things invisible and imperceptible by our senses, and which cannot be subjected to direct experiment. [190]

James C. Maxwell, 1873
describing the new concept and
his understanding of molecules

Structure-based computational methods can be used for modeling protein-DNA and protein-ligand complexes *in silico*, thereby providing valuable insights into physical interactions at an atomic level. Not only the pure structure of biological macromolecules can be modeled computationally, but also their conformational changes can be examined by various computational approaches. In the following sections the underlying theory on molecular mechanics (MM) and molecular dynamics (MD) is described.

3.1 Molecular mechanics

Force fields (FFs) are used to model molecules by calculating the energy of the system as a function of their atomic positions [166]. MM represents the interactions occurring within a system based on a rather simple theory. The basic idea of MM is that bonds and angles have standard values, which are consistent for all bonded interactions. Each bonded and non-bonded interaction energy is modeled by an analytical term consisting of empirical parameters. These parameters are developed by calibrating

the FF with experimental data mostly on small molecules and results from quantum mechanical (QM) calculations. Even if the parameters are tested on a relatively small number of test cases, these FFs can still be applied to larger molecules and a wider range of problems. The popular *AMBER* FF [57, 290] consists of five contributions:

$$\begin{aligned}\mathcal{V}(\mathbf{r}^N) = &\sum_{\text{bonds}} \frac{K_r}{2}(r - r_{eq})^2 + \sum_{\text{angles}} \frac{K_\theta}{2}(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\ &+ \sum_{i<j} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\end{aligned} \tag{3.1}$$

In Eq. (3.1) the potential energy is described by $\mathcal{V}(\mathbf{r}^N)$, which is a function of the positions $\mathbf{r}$ of N particles (usually atoms) in a system. The first contribution in this equation is a harmonic term and specifies the resulting energy change of stretching or compressing a chemical bond of length r in reference to its average length r_{eq} determined by experimental data or QM results. The more the bond length r deviates from the reference value r_{eq}, the more increases the energy within the system. The second energy contribution is also a harmonic term and expresses the opening or closing of angle θ, as compared to the reference angle θ_{eq}. Each angle between three specific atoms A, B, and C can be described by a specific value θ_{eq} representing the most common angle between these atoms determined during FF calibrations. The third term is a torsion potential, which is represented as a cosine series expansion and describes the change of energy by rotating around a single bond. The torsion angle is ϕ multiplied by n, which represents the number of minimum points in the function when the bond is rotated through $360°$. The phase factor γ determines the minimum state of the torsion angle and K_r, K_θ, and V_n are empirically determined constants.

The last two terms describe non-bonded interactions. The first non-bonded term represents vdW interactions between temporarily induced dipoles. The repulsion and dispersion forces are represented by a Lennard-Jones potential. The collision parameter σ between two atoms i and j is set to a value for which the energy is zero. The Lennard-Jones potential is constructed from a repulsive part r_{ij}^{-12} and an attractive part r_{ij}^{-6}. Combining these two components the Lennard-Jones potential represents the non-bonded vdWs interaction energy between two atoms.

The second non-bonded contribution represents the energy of electrostatic interactions between two molecules or between two parts of a molecule. It is approximated by the sum of interactions between pairs of point charges q_i and q_j of atoms i and j and their distance r_{ij} using Coulomb's law.

In more advanced FFs additional terms, for example special hydrogen bond terms or cross-terms, are included [127]. Cross-terms describe the coupling of internal coordinates. When an angle between atoms A, B, and C is decreased, simultaneously the bond lengths of atom pairs A and B and B and C are also increased in general.

3.2 Molecular dynamics

MD simulations allow studying dynamical properties of biological macromolecules *in silico*, for example, proteins, nucleic acids, and membranes. MD is a theoretical approach trying to describe structural alterations of biomolecules. These dynamical events play a key role in all living organisms especially in binding of biomolecules. In MD simulations varying conformations of a biomolecule over a certain time results in a trajectory. A trajectory specifies how positions and velocities of atoms in a system change throughout a simulation, from the starting coordinates to the end of the MD run. Trajectories are obtained by solving the differential equation gained from Newton's Law of Motion ($\mathbf{F} = m \cdot \mathbf{a}$) shown in Eq. (3.2). The underlying core of an MD simulation is MM, whereby the FFs used in MD need to be differentiable. Such a differentiable FF is for example the *AMBER* FF, as described in Eq. (3.1).

$$\frac{d^2\mathbf{r}_i(t)}{dt^2} = \frac{-\nabla\mathcal{V}(\mathbf{r}_i(t))}{m_i} = \frac{\mathbf{F}_i(t)}{m_i} = \mathbf{a} \tag{3.2}$$

In Eq. (3.2) m_i is the mass of atom i, which is moving along vector $\mathbf{r}_i$ with $\mathbf{F}_i$, the directional force on the atom. The force $\mathbf{F}_i$ exerted on atom i by all other atoms in the molecular system is given by the negative gradient of the potential energy function $\mathcal{V}$ (see Eq. (3.1)) which in turn depends on the coordinates of all N atoms in the system.

3.3 Solvents in molecular modeling

Water is known to be the universal solvent on earth, which is essential to all known forms of life. In order to simulate biomolecules *in silico* in a realistic way, especially their behavior, action, and dynamics inside biological cells, the environment needs to be modeled as accurately as possible. In molecular modeling there are mainly two approaches to represent a solvent environment: explicit and implicit solvent models.

3.3.1 Explicit solvent

In explicit solvent models, discrete water molecules are placed around the molecule approximating the properties of solvent-solvent and solute-solvent interactions. The transferable intermolecular potential three point (TIP3P) water molecule [138] is one representative for an *in silico* water molecule. It places positive charges on the two hydrogen atoms (q_1 = 0.417) and a negative charge on the oxygen atom (q_2 = -0.834). The third interaction site is known as the Lennard-Jones interaction point and is located at the center of the oxygen atom. With this interaction point the vdW interaction between two water molecules is estimated. Each TIP3P water molecule is assumed to have a rigid geometry. Other water models, for example TIP4P [138] and ST2 [275], have also a rigid geometry, but have four and five interaction sites, respectively. The use of a rigid water molecule model is obviously an approximation, but on the other hand results in less computational effort. Some properties cannot be

determined like the internal flexibility of water molecules, which would be indispensable for calculating vibrational properties.

For explicit solvent simulations, in which biomolecules are surrounded by discrete water molecules, periodic boundary conditions are generally applied. Approximately three quarters of water molecules present in a water box are at the surface of the solvent rather than being placed in the bulk region. The behavior of water molecules at the surface is different and would be influenced by the walls of the water box. In order to avoid the possibility that this affects the behavior of the biomolecule during MD simulations, the water box would need to be sufficiently large. However, a large number of water molecules results in high computational time. The number of water molecules as well as the computational time is decreased by placing the molecule in a periodic boundary box, shown in Fig. 3.1 [166]. For example, a cubic box of particles is replicated in all directions (26 times in 3D, eight times

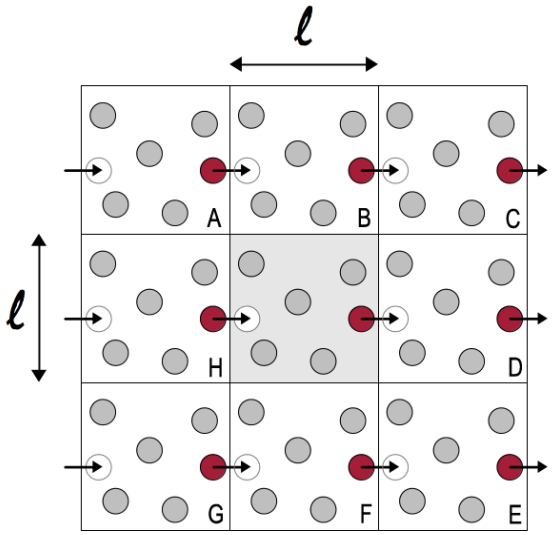


Figure 3.1: Periodic boundary box. Explicit solvent simulations are usually performed under periodic boundary conditions. In 2D the original system (gray box) with box length l is replicated eight times (A, B, C, ... , H) resulting in a periodic boundary box. Water molecules (gray circles) experience forces as if they were in bulk fluid.

in 2D (gray cell in Fig. 3.1)) to give a periodic array, in such a way that the water molecules experience forces as if they were in bulk fluid. If a particle, for example, the red particle in Fig. 3.1, leaves the gray cell, an imaginary particle (the red particle of cell H and indicated as transparent particle in the gray cell in Fig. 3.1) will enter the cell on the opposite site, while the number of particles in the gray box remain constant during MD simulations [61]. Even if periodic boundary conditions are widely used in MD simulations they produce artifacts. One drawback is that it is not possible to achieve fluctuations with wave lengths greater than the length of one cell.

3.3.2 Implicit solvent

Implicit solvent models are used for analyzing solvation effects without explicit treatment of water molecules. In an implicit solvent model the medium, in this case water, is represented by a continuum with electrostatic, entropic, and viscous properties reflecting the properties of water. A continuum electrostatic model of a biomolecule in water is commonly described by a low-dielectric medium surrounded by a high-dielectric ($\epsilon_s = 78.5$) continuous medium without any explicit charges representing water. The dielectric constant of the protein is usually set equal to the dielectric constant of small, non-polar organic compounds (ϵ_m = 2.0 - 4.0) [99, 166]. It should be kept in mind that proteins do not have a uniform consistent core, but consist of polar and non-polar regions causing diverse interior dielectric constants. These values range from 2.0 to 20.0 [9, 10, 258]. However, it is difficult to account for these diverse values when modeling the interior dielectric constant of proteins. In describing TF-DNA complexes, the DNA molecule also needs to be modeled. Modeling DNA in implicit solvent is a challenge since nucleic acids interact strongly with their surrounding solvent. However, continuum models cannot describe the polar interactions of nucleic acids with their environment sufficiently. Although implicit solvent models lack some important features they are successfully applied to calculating binding free energies of protein-protein [98, 100, 216], protein-RNA [103, 241], and protein-ligand [29, 242, 316] complexes.

3.4 Poisson Boltzmann calculations

The electrostatic contribution of the solvation free energy can be calculated with the Poisson Boltzmann (PB) method. This approach is commonly applied when implicit solvent models are used. The PB equation can be solved by the finite-difference method, which is more efficient for nonlinear PB or the boundary element method, which should be used for linear PB.

The Poisson equation, Eq. (3.3), relates the variation in the electrostatic potential within a dielectric continuum environment to the charge density.

$$\nabla \cdot \epsilon_s(\mathbf{r}) \nabla \Phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{3.3}$$

In the Poisson equation, which is one of the fundamental equations of classical electrostatics, Φ is the electrostatic potential, ϵ_s the dielectric constant of the solvent, ρ is the charge density, and $\mathbf{r}$ is the position vector. The electrostatic potential, the dielectric constant, and the charge density are all position-dependent. The electrostatic potential is a scalar quantity, which when multiplied by the charge of a particle, is equivalent to the energy required to transfer this particle from a field-free location into the electrostatic field. The Poisson equation tells us how the electrostatic potential varies throughout space, due to the charge distribution.

In a biological environment, the biomolecules are not surrounded by pure water alone. The solvent also consists of ions determining the salt concentration of the solution, e.g., sodium and chloride ions.

The salt concentration can be represented by the Boltzmann distribution, described in Eq (3.4).

$$n_i = n_i^0 e^{-q_i \Phi / k_B T} \tag{3.4}$$

In the Boltzmann distribution n_i^0 represents the density of ions in bulk solution, q_i is the charge on the ion, Φ is the electrostatic potential in that area of space, k_B is the Boltzmann constant, and T the temperature in Kelvin. The Boltzmann distribution, which is employed here to provide the density of mobile counterions, implies an accumulation of anions where the potential is positive and of cations where the potential is negative. When the effects of Eq. (3.4) are incorporated into the Poisson equation, Eq. (3.3), it results in the PB equation, see Eq. (3.5).

$$\nabla \cdot \epsilon_s(\mathbf{r}) \nabla \Phi(\mathbf{r}) - \kappa' \sinh[\Phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) \tag{3.5}$$

κ' is related to the Debye-Hückel inverse length, κ, see Eq. (3.6) and incorporates salt effects into the PB equation.

$$\kappa^2 = \frac{\kappa'^2}{\epsilon_s} = \frac{8\pi N_A e^2 I}{1000 \epsilon_s k_B T} \tag{3.6}$$

In Eq. (3.6), e is the elementary charge, I is the ionic strength of the solution and N_A is Avogadro's number. The nonlinear PB equation can be written by expanding the hyperbolic sine function as a Taylor series.

$$\nabla \cdot \epsilon_s(\mathbf{r}) \nabla \Phi(\mathbf{r}) - \kappa' \Phi(\mathbf{r}) \left[1 + \frac{\Phi(\mathbf{r})^2}{6} + \frac{\Phi(\mathbf{r})^4}{120} + \cdots \right] = -4\pi\rho(\mathbf{r}) \tag{3.7}$$

Using only the first term in the expansion of Eq. (3.7), the linear PB equation is obtained, see Eq. (3.8).

$$\nabla \cdot \epsilon_s(\mathbf{r}) \nabla \Phi(\mathbf{r}) - \kappa' \Phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{3.8}$$

The linear PB equation is computationally easier to solve, but it is more reasonable to choose the nonlinear equation for DNA or RNA calculations, where ionic strength effects are known to be important. The PB equation can be solved by finite difference and boundary element approaches. The finite difference method has first been introduced only for proteins by Warwicker *et al.* [297] and then implemented by several other groups. For example, Honig *et al.* developed the *DelPhi* program which is widely in use for PB calculations [150, 213, 243, 244]. This approach is stated to be more efficient for solving the nonlinear PB equation in contrast to the boundary element method [239], which seems to be more efficient for the linear equation. Another software package for solving the PB equation is *APBS* [12]. It uses the *Finite Element ToolKit* [13, 198] to solve the PB equation numerically. *APBS* was designed to efficiently evaluate electrostatic properties for many biomolecular simulation studies and provides implicit solvent models which accurately account for both repulsive and attractive interactions between solute and solvent.

3.5 In silico binding affinity predictions

PB theory can be used to calculate the electrostatic contribution of the solvation free energy term for binding affinities between protein and DNA or protein and ligand molecules. Other contributions to the binding free energy are obtained using MM or MD simulations. One approach to predict binding affinities *in silico* will be described in the following.

The overall objective of the Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA) method is to calculate the free energy difference between two states which most often represent the bound and unbound state of two solvated molecules A and B. Since both binding partners A and B and the positioning of water molecules change upon binding we indicate the change with an asterik (*) in the following.

$$[A]_{aq} + [B]_{aq} \Leftrightarrow [A^*B^*]_{aq^*} \tag{3.9}$$

Ideally, we would like to calculate the free energy of binding between two biomolecules A and B directly. However, in such a simulation the majority of energy contributions would come from solvent-solvent interactions. Thus, the fluctuations in total energy would be an order of magnitude larger than the actual binding free energy and the calculation would take an exorbitant amount of time to converge. A more effective method to determine binding free energies of two biomolecules is to decompose the calculation according to the thermodynamic cycle (Fig. 3.2).

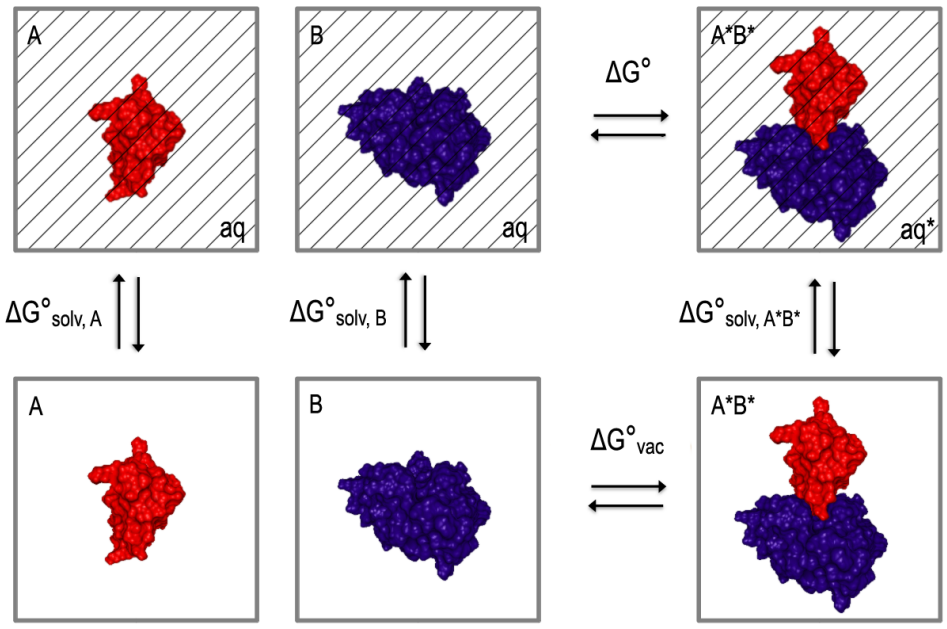


Figure 3.2: Thermodynamic cycle. The thermodynamic cycle describes the binding process of two biomolecules A (red) and B (blue). The upper half describes the complex formation in aqueous solvent and the lower half in vacuum that both results in the complex A*B*. The thermodynamic cycle describes an approach for determining the binding free energy ΔG° computationally.

The binding free energy ΔG° can then be calculated as follows:

$$\Delta G^\circ = \Delta G^\circ_{vac} + \Delta G^\circ_{solv,A^*B^*} - (\Delta G^\circ_{solv,A} + \Delta G^\circ_{solv,B}) \tag{3.10}$$

The different contributions to the binding free energy ΔG° are calculated separately in the MM-PBSA approach. Solvation free energies ΔG°_{solv} are calculated by either solving the linear or nonlinear PB equation for each of the three states, which represents the electrostatic contribution to the solvation free energy and accounting for hydrophobic contributions with an empirical term.

$$\Delta G^\circ_{solv} = G^\circ_{elec,\epsilon_s} - G^\circ_{elec,\epsilon_m} + \Delta G^\circ_{hydrophobic} \tag{3.11}$$

ΔG°_{vac} is determined by the average interaction energy between molecule A and B. Additionally, the entropy change upon binding can be taken into account if desired.

$$\Delta G^\circ_{vac} = \Delta E^\circ_{MM} - T \cdot \Delta S^\circ \tag{3.12}$$

The entropic contribution can be calculated by performing normal mode analysis (NMA) on the three structures. In practice, entropic contributions are often neglected if only states of similar entropy are compared such as two similar molecules A binding to the same molecule B. The reason for this is that NMA calculations are computationally expensive and introduce uncertainty in the results, since they tend to have large margin of errors.

4 Studying protein-DNA complexes at the atomic level

The lactose repressor selects one out of six million nucleotide sequences in the Escherichia coli genome and binds to it to prevent the expression of the genes for lactose metabolism. How does this protein, a 150,000-dalton tetramer of identical subunits, recognize its target? [96]

Walter Gilbert, 1973
introducing his study on the
lac operator

In this chapter we will now turn toward studying protein-DNA complexes at the structural level. We will highlight specific interactions between TFs and the DNA double helix while additionally considering flexibility. For studying these interactions on a molecular level, protein-DNA complex structures are required, however, they are often not available in structural databases [21, 22]. Therefore, we present in the first part of this chapter an universally applicable approach that models arbitrary protein-DNA complexes. To refine such modeled, or also experimentally obtained, protein-DNA complexes we introduce an advanced MD relaxation and simulation protocol which yields even more accurate protein-DNA complex models. In the second part of this chapter, we study selected representatives of the WRKY protein family. We elucidate structural characteristics and unique features of different WRKY proteins at the protein-DNA binding interface at atomic detail using molecular modeling and dynamics simulation approaches.

4.1 Introduction

Regulatory DNA sequences harbor essential information for gene regulation and comprise information from upstream signaling cascades. This is mediated by several colluding factors, amongst others by

direct binding of proteins to short DNA motifs in regulatory regions. In these areas several signaling pathways are presumably able to merge in order to fine-tune gene expression [92, 294]. Experimental approaches shed light on TFBSs by performing mutation studies along DNA sequences to investigate the binding affinity and specificity of one TF to different DNA motifs. Not only DNA sequences can be mutated *in vitro*, but also protein residues at the binding interface. Thereby, amino acids which are directly involved in the binding process can be identified and their contribution to the overall binding affinity can be estimated. Traditional experimental methods for determining the specificity of DNA-binding proteins are laborious, however, tremendous advances have been achieved during the last years which led to the development of high-throughput methods [18, 35, 148, 184, 227, 277, 314]. Information gained from these mutation studies in combination with predictions of TF binding locations provides more detailed views of the regulatory circuit of cells and the effects of variation on gene expression.

By studying in detail the structure of DNA, Nikolova *et al.* [214, 215] discovered in 2011 that DNA double helices exhibit not always a perfect Watson-Crick double helix structure in the unbound state. Their studies highlight the significance of structural information even for nucleic acids. They found that so-called Hoogsten base pairs (bps) do not only occur in DNA when it is damaged or is bound to a drug or protein, which was readily accepted until then, but are also present in normal DNA. This implies, that more layers of information are stored in the genetic code in contrast to sequence information alone. Therefore, proteins might not only induce structural changes to DNA upon binding, but also recognize specific double helix structures. These findings stress the importance of incorporating structural features in studying binding interfaces of protein-DNA complexes.

Protein-DNA complexes were among the first macromolecular structures characterized by X-ray crystallography. Since Gilbert and Müller-Hill [1, 97] isolated the lac repressor with its operon, studying specific protein-DNA interactions at the binding interface has fascinated many researchers. Especially the following questions are of great interest while analyzing protein-DNA complexes: How are specific interactions formed between TFs and the DNA double helix? How do such interactions alter the structure of involved macromolecules? How do certain protein-DNA interactions account for specificity in detail? How do TFs differentiate their specific and short DNA binding motif from the many similar DNA sequences in a given genome? Answers to these questions can then explain observations at the sequence and even functional level of protein families and confirm experimental results from specificity and binding affinity studies.

Studying the structure of protein-DNA complexes, especially at the binding interface leads to deeper insights into the recognition process and specific interactions. Computational structure-based studies at atomic resolution facilitate to gain knowledge about the structural and chemical complementarity between TFs and their binding sites. The recognition of certain DNA sequence motifs by proteins is based on the overall structure of both, proteins and DNA and on specific contacts at the binding interface. For correct and orientation-dependent binding, TFs take several biophysical properties into account: the sequence consensus of bps, local charge and hydrophobicity, size of the major or minor groove, DNA-shape, and bending properties of DNA [246]. This information is stored in the

consecutive sequence of different bps. However, it is difficult to provide protein-DNA complexes by experimental methods for all existing TFs and especially for all possible mutations at the binding interface. Molecular modeling techniques do not only study existing protein-DNA complexes, but can also be applied to insert single mutations at the binding interface of existing protein-DNA complexes and estimate the influence of these alterations to the binding affinity. In a review [174], Liu and Bradley describe state-of-the-art methods [2–4, 8, 15, 68, 108, 173, 205, 206, 237, 260, 266, 278, 313] for modeling protein-DNA interaction specificity at atomic detail. They conclude that the success of existing structure-based molecular modeling methods depends on the quality of energy functions and conformational sampling algorithms. These algorithms are important for modeling structural dynamics and accounting for optimal refinement after mutating bps or protein side chains. Conservative approaches estimating binding affinities using exclusively experimental, high-resolution protein-DNA complexes seem to give the best results even though they account not at all or only a little for flexibility [3, 4, 68, 206]. However, when introducing mutations at the binding interface, modeling flexibility improves the prediction accuracy for the binding affinity [8, 266, 311].

When no experimental protein-DNA complex is available, a common computational approach to obtain such a complex for studying the binding interface is docking. There exist some docking approaches which even account for flexibility at the protein-DNA interface [65, 67]. *HADDOCK* (High Ambiguity Driven protein-protein DOCKing), developed in the group of Bonvin [285, 286], is an information-driven flexible docking approach for the modeling of biomolecular complexes. *HADDOCK* distinguishes itself from other docking methods in the fact that it encodes information from identified or predicted protein interfaces in ambiguous interaction restraints to drive the docking process. These restraints are defined as ambiguous distances between all residues shown to be involved in the interaction. It is even able to account for DNA flexibility and captures the bend and twist geometries of DNA, which are induced upon protein binding. For these reasons it is known as the best protein-DNA docking method. In common docking approaches full structural flexibility of both biomolecules is restricted and mostly predefined protein side-chain orientations are used. This is also a routine procedure in molecular modeling approaches, which incorporate flexibility in native or mutated protein-DNA complexes. Existing methods are mostly limited to implicit solvent simulations or constrained[1] protein and DNA backbone atoms [8, 266, 311]. This implies that there exist restrictions in typically applied methods for studying protein-DNA specificity. In this chapter, we present a molecular modeling approach for constructing arbitrary protein-DNA complexes. Additionally, we introduce an advanced explicit solvent MD simulation protocol to overcome limitations of existing methods. MD simulation protocols often imply high computational requirements, nevertheless they can yield detailed knowledge when applied on individual case studies for analyzing specific protein-DNA interactions and dynamics.

In this chapter we study protein-DNA complexes and their interactions at the binding interface

[1] Although it is common in the MD simulation community to use the term *restraint* when, for example, an energy penalty is applied to prevent a system or part of a system to deviate from a defined value (atomic position, intramolecular distance, or dihedral angle) and *constraint* when the length of bonds or angles are fixed, we are going to use *constraint* for both in the following.

of mainly one protein family: the WRKY proteins. WRKY proteins are a plant-specific subfamily of the large Glia cell missing 1-FLYWCH (GCM1-FLYWCH) superfamily of zinc-finger TFs that are found in eukaryote organisms and it has been speculated that this superclass evolved from prokaryote transposases [11, 77, 78, 187]. The WRKY protein family comprises 72 members in *Arabidopsis thaliana* (*At*), but can be found amongst others in sweet potato, rice, parsley, and grape [299]. They help plants to overcome different stress situations [140, 169, 264, 284, 293, 307, 315] and are involved in the regulation of developmental processes [115, 135, 178]. This makes them an interesting target in basic plant research as well as in crop and agricultural industry. The hallmark of all WRKY proteins is their highly conserved DNA-binding domain (DBD). This domain can be divided in an almost invariant

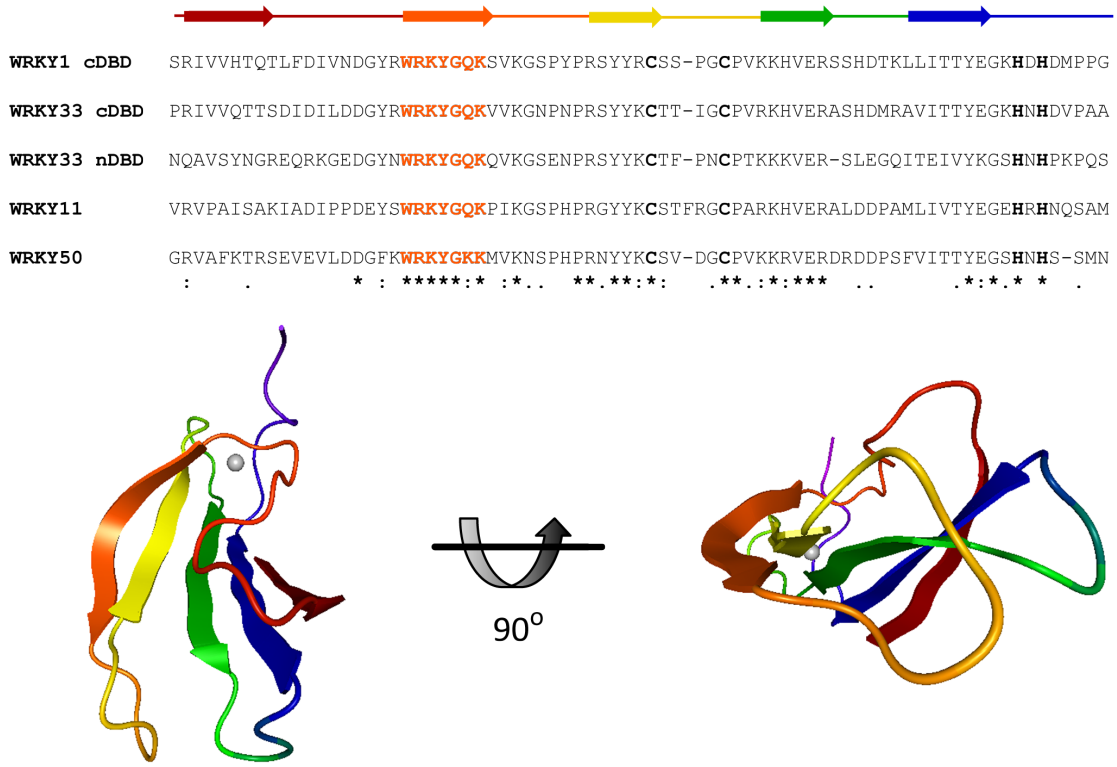


Figure 4.1: *At*WRKY DNA-binding domains. A multiple sequence alignment of different *At*WRKY DBDs is illustrated, where (*) indicates the same amino acid, (:) amino acid with similar chemical properties, and (.) majority of amino acids with similar chemical properties. Above the alignment, the general secondary structure based on the crystal structure of *At*WRKY1 cDBD (PDB id: 2ayd) is given. Below the alignment the structure of *At*WRKY1 cDBD is illustrated, which consists of five β-strands. The second β-strand that is formed by the following amino acids: WRKYGQK interacts with the DNA double helix. Figure based on Fig. 2 in [30].

WRKYGQK amino acid sequence and either a C2-C2 or a C2-H2 zinc complexing finger structure [78]. Depending on the type of zinc finger (ZF) and the number of WRKY-domains per protein, the WRKY family can be divided into several groups (I-III) and corresponding subgroups. Two DBDs are present in group I WRKY proteins, one C-terminal WRKY DNA-binding domain (cDBD) and one

N-terminal WRKY DNA-binding domain (nDBD), while group II and III WRKY proteins have only one single WRKY DBD. It is assumed that the grouping reflects a certain functional diversity of the WRKY proteins [249], which is not yet fully understood. The overall structure of WRKY DBDs is about 60 amino acids long and consists of five β-strands. As mentioned before, the DBD of WRKY TFs contains the almost invariant amino acid residues tryptophane (W), arginine (R), lysine (K), and tyrosine (Y), which are responsible for the protein name. The WRKYGQK residues lie consecutively along a β-strand, kinked in the middle of the sequence by the glycine residue. This β-strand forms specific contacts with the major groove of the DNA, thereby revealing a specific DNA-binding motif. Almost all WRKY proteins favor a specific DNA sequence for binding, termed the W-box (5'-TTGAC-C/T-3') [43, 56, 64, 248, 293].

In this thesis we focus on the analysis of the structure, specific interactions with DNA, and dynamics at the binding interface for *At*WRKY1 cDBD, 11 DBD, 50 DBD, 33 cDBD, and 33 nDBD. The N-terminal DBD of group I WRKY proteins has never been shown to bind to DNA. The structures of the two domains of *At*WRKY33 (cDBD and nDBD), are compared to each other and binding affinities are determined for both domains bound to DNA in the last part of this chapter. *At*WRKY11 DBD and 50 DBD exhibit an amino acid difference within the highly conserved second β-strand. Unique features of *At*WRKY11 and 50 are studied at the molecular level which provide for the first time hints to amino acids that specifically interfere with the DNA-binding process and thereby offer a conclusive molecular mechanistic model for the WRKYGQK consensus in controlling DNA-binding specificity.

4.2 Modeling and refining protein-DNA complexes using molecular dynamics

In this section we describe an approach for modeling an arbitrary protein-DNA complex, which can be used when no protein-DNA complex structure is available. Nonetheless, two structures must be given: the protein structure of interest and a protein-DNA complex of a related protein. Additionally, we introduce an advanced MD relaxation and simulation protocol for modeled or experimentally obtained protein-DNA complexes.

4.2.1 Introduction

For studying three-dimensional biomolecular structures the PDB [21] is an inevitable source for data. At the moment (June 2012) the PDB holds about 82,000 three-dimensional structures obtained mainly by XRD or NMR experiments. Different types of biomolecular structures are stored in the PDB, which can be classified into four main groups: proteins, DNA, RNA, and protein-nucleic acid complexes. These different biomolecular structure types occur in varying numbers and account for approximately 76,000 proteins, 1,350 DNA, 950 RNA, and 3,700 protein-nucleic acid complexes. These numbers state clearly that more protein structures are available than, for example, protein-DNA complexes. Methods which only rely on existing protein-DNA complexes can study binding interfaces of these

complexes and can include single bp or side chain mutations, but cannot expand their studies to cover the whole structural space of protein-DNA complexes.

As mentioned previously, docking is a common approach in molecular modeling to obtain a protein-DNA complex, when both biomolecules are only separately available. Nevertheless, typically the input structures for docking methods need to be obtained experimentally. However, especially the surface of XRD or NMR structures looks differently when no binding partner has been taken into account during creation. For instance, in XRD structures crystal cell packing is sometimes dense, thereby side chains of proteins in neighboring crystal cells form contacts with each other. That implies that these protein structures are refined with respect to artificial conditions. DNA double helices can also be shaped differently in XRD or NMR experiments when no protein binding partner is available. However, when the DNA double helix is modeled *in silico* it comprises ideal angles and atom positions and would feature an idealistic structure. In either case, biomolecules need to adapt their structures during or after docking to represent a protein-DNA complex more truthfully. However, flexibility is a compute-intensive task that protein-DNA docking programs do not account for or include only partially.

We describe a modeling procedure for arbitrary protein-DNA complexes which can be applied when the protein of interest is not bound to its DNA binding partner. Both biomolecules are also primarily considered static as in most docking approaches. To overcome these rigid conditions, flexibility is induced by molecular dynamics to recalibrate the structure of the modeled protein-DNA complex. We introduce an advanced explicit solvent MD simulation protocol that permits these structural alterations. Especially, amino acids at the binding interface can then shift in order to adopt optimal orientations toward the DNA double helix and the double helix can also relax its structure. These structural alterations lead to an overall structural change of the protein-DNA complex while specific interactions between protein side chains and DNA bps are being formed. This MD relaxation and simulation protocol can also be used to equilibrate experimentally obtained protein-DNA structures.

4.2.2 Materials and Methods

Protein-DNA complexes

In 2010, there was only one WRKY protein crystal structure available in the PDB, which is the protein structure of *At*WRKY1 cDBD (PDB id: 2ayd) [71]. There was also an NMR structure of *At*WRKY4 cDBD (PDB id: 1wj2) [308] available that is as *At*WRKY1 cDBD not bound to DNA. The general structure of both proteins is highly resembling and features all characteristics as illustrated in Fig. 4.1.

After searching for similar protein structures of *At*WRKY1 cDBD using the *Dali* server [119], a protein-DNA complex structure, a Glia cell missing *Mus musculus* (*Mm*GCM1a) TF bound to DNA (PDB id: 1odh [53], Z-score > 6.2, in October 2010) was returned. A similar search has previously been described by Duan *et al.* [71]. The *Mm*GCM1a domain consists of two domains, which contain two ZF motifs, whereby the large and the small domain displays five- and three-stranded β-sheet sections, respectively. The *Mm*GCM1a domain comprises a novel mode of sequence-specific DNA

recognition, where the five-stranded β-sheets fit into the major groove of the DNA [53], see Fig. 4.2 (1). Residues lying along the β-strand which is closest to the DNA form contacts with the DNA.

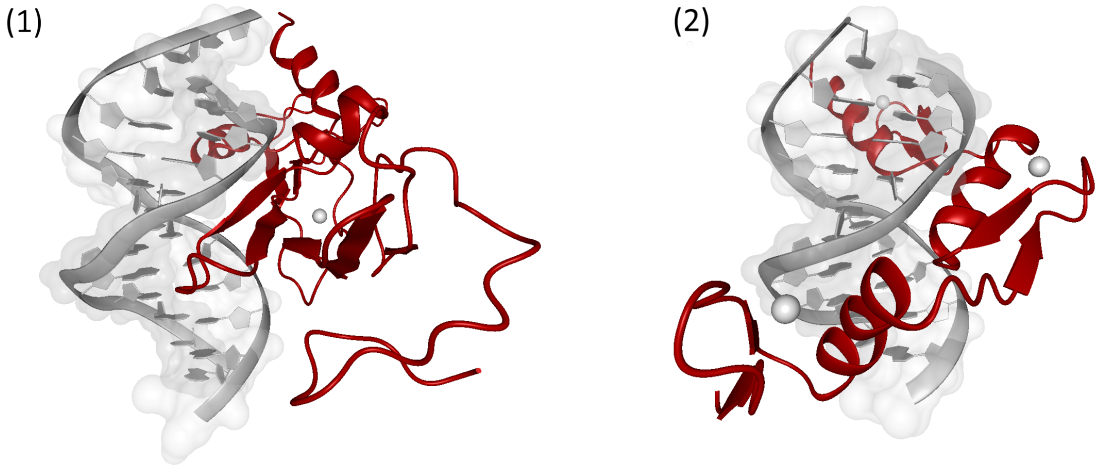


Figure 4.2: *Mm*GCM1a and Zif268 bound to DNA. Proteins (shown in red) bind along the DNA (shown in gray) major groove by forming contacts with the DNA backbone and bps. In (1) the *Mm*GCM1a protein is illustrated which interacts with the DNA with a β-sheet, formed by five β-strands. The smaller domain consisting of three β-strands is partially placed behind the double helix. The DNA-binding domain (DBD) is built up of five β-strands which is structurally similar to the DBD of WRKY proteins. The Zif268 (2) binds with three α-helices to the DNA major groove.

In 2012, the first WRKY-DNA complex was experimentally obtained and published in the PDB (PDB id: 2lex) [309]. This protein-DNA complex consists of the C-terminal DBD of *At*WRKY4 and a DNA double helix with the following sequence: 5'-CGCCTTTGACCAGCGC-3'. 20 NMR solution structures were published in the mentioned PDB file. We used the first NMR model of this WRKY-DNA complex for validating our modeled WRKY-DNA structures.

For demonstrating the applicability of our advanced molecular dynamics relaxation and simulation protocol for protein-DNA complexes we performed simulations with another TF-DNA complex: the well studied ZF Zif268 protein bound to DNA (PDB id: 1aay) [75] illustrated in Fig. 4.2 (2). This TF comprises three ZF motifs, whereas each ZF binds to a three bp long DNA sequence.

Mapping

We mapped the *At*WRKY1 cDBD protein [71] onto the *Mm*GCM1a protein [53] using RMSD-minimizing superposition (as implemented by the atom bijection method in BALL [114]). We stored 33 C_α atoms of both proteins (listed in Table B.1 in the appendix) in a paired list as input for the atom bijection method. This mapping yields a root-mean-square deviation (RMSD) of 3.18 Å for all 68 WRKY protein C_α atoms. Taking only the aligned β-sheet C_α atoms (26 C_α atoms) into account (listed in Table B.2 in the appendix) the RMSD drops to 0.88 Å. Mapping the *At*WRKY1 cDBD structure onto the *Mm*GCM1a protein results in a protein structure that comprises all atoms of *At*WRKY1 cDBD [71] now oriented in a way that all C_α atoms are optimal superimposed with all

C_α atoms of the *Mm*GCM1a protein [53]. The DNA is still missing in the resulting structure, but can be easily inserted by taking all DNA atoms of the GCM1-DNA complex structure (chain B and C of PDB structure 1odh) and adding them to the beginning of the rotated and translated 2ayd structure.

Base pair mutations

The original 1odh (chain B) DNA sequence (5'-CGATGCGGGTGCA-3'), in the created *At*WRKY1 cDBD-DNA complex model is mutated to a DNA sequence containing the WRKY's W-box sequence (5'-TTGACC-3'). We used the DNAMutator method in *BALL* (version 1.4.2) [114] (DNAMutator is a simple, geometry based method with local reoptimization for DNA bp mutations) to mutate bps by keeping DNA backbone atoms, deleting 1odh nucleobase atoms, and inserting new nucleobase atoms in the same plane as the old ones. We used a parsley promoter sequence (5'-TCAAAGTTGACCAATAAT-3') [248] as a reference DNA sequence and mutated the bases according to the binding profile Duan *et al.* [71] developed for the WRKY-DNA complex. This results in the following DNA sequence for the WRKY-DNA complex: 5'-AAAGTTGACCAAT-3' (chain B). The antiparallel antisense DNA strand (chain C) was mutated at the same time correspondingly, since our program considers bps.

Docking

To compare our protocol for modeling protein-DNA complexes with other established approaches we docked the *At*WRKY1 cDBD protein (PDB id: 2ayd) [71] to a DNA sequence using *HADDOCK* [65, 67]. Default values were used for docking using the interface of the *HADDOCK* webserver, which requires two input structures. The first input structure is the DNA double helix constructed using the NAB module in *AMBER 11* [39]. Its sequence comprises the parsley promoter sequence (5'-TCAAAGTTGACCAATAAT-3') which contains the W-box segment at positions seven through twelve. The second input structure is the *At*WRKY1 cDBD protein (PDB id: 2ayd) without its zinc ion. The active residues that are thought to be directly involved in the interaction are specified [71, 309]. For the DNA molecule the following bases were specified: 7, 8, 9, 10, 11, 12, which defines the W-box: 5'-TTGACC-3'. For the *At*WRKY1 cDBD we choose the following residues: 312, 314, 315, 316, 317, 318, 319, these numbers correspond to the residue IDs of the WRKYGQK amino acids of the *At*WRKY1 cDBD protein (PDB id: 2ayd).

Preprocessing steps for molecular dynamics simulations

The input for the actual MD simulation protocol is a preprocessed protein-DNA complex. We describe the preprocessing steps for the *At*WRKY1 cDBD-DNA complex in the following. For all other complexes [53, 75, 309] the preprocessing is performed accordingly. It is inevitable for MD simulations that all residues and all atoms are present in the PDB file. Therefore, we checked our structure (PDB id: 2ayd [53]) with *what-check* [121] to identify missing atoms and residues. Additionally, orientations of side chains are checked as well as the protonation states of certain amino acids (His, Cys, Arg, Glu, and Lys). The *At*WRKY1 cDBD protein passes all tests and is then used as input for the H^{++}

web server [6, 101]. H^{++} determines protonation states of all amino acid residues. Since *At*WRKY1 cDBD contains a zinc ion which is coordinated by two histidine (resdidue ids: 361 and 363) and two cysteine residues (residue ids: 332 and 337) these four residues are protonated correspondingly (see Fig. B.1 in the appendix). Specific residues need to be renamed according to the naming conventions of *AMBER*, which results in the following residue name changes: HIS 298, 342, 348, and 361 to HIE, HIS 363 to HID and CYS 332 and 337 to CYM (the numbers correspond to residue ids in the original PDB file 2ayd). As described previously, we added 13 DNA bps (chain B and C of PDB id: 1odh) to the beginning of the PDB file. Therefore, we changed the residue id of the first amino acid of PDB structure 2ayd to residue id 27 and all other residue ids accordingly. At the end of the modified PDB file CONECT rows describing bonds between S atoms of cysteins and N atoms of histidines to the zinc ion are also added. Additionally, the residue name of the zinc ion is changed to Z4.

This modified PDB file containing all protein-DNA complex atoms and corresponding positions serves as input for *LEaP*. *LEaP* is a preparatory program that can create topology and coordinate files from PDB files and is part of *AMBER*. We used the ff99sb [123] for simulating protein-DNA complexes, which is a non-polarizable force field based on the Cornell *et al.* [57] force field. It is suitable for proteins and nucleic acids and stated to perform best for biomolecules, when we started the project in 2010. For this force field the parm99.dat parameters and specific parameters for nucleic acids stored in the DNA_CI.lib and frcmod.parmbsc0 files [229] are loaded. Since zinc ions should remain at their original positions, coordinated by certain amino acid side chains, parameters for these zinc ions are also loaded. The parameters for bonds, angles, dihedrals, and improper torsions are specified in Z4 library and frcmod files (listed in Table B.4 and B.5 in the appendix). The biomolecule is placed into an octahedral waterbox of approximately 9,000 TIP3P water molecules. The distance

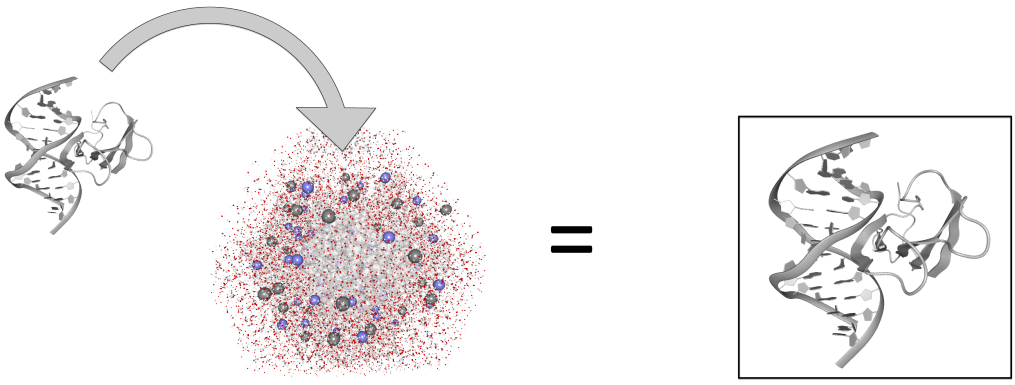


Figure 4.3: Preprocessing The protein-DNA complex is placed into an octahedral waterbox and counterions (sodium ions shown in blue and chloride ions shown in gray) are added. This system is represented as squared box in the following.

between the atoms at the surface of the protein-DNA complex and the edge of the waterbox is at least 15 Å. In order to neutralize the system, counterions are added to the solvent. In this case 17 sodium ions are placed around the protein-DNA complex. To mimic *in vitro* and more importantly

in vivo conditions a salt concentration of 0.2 mol/l is added to the solvent. This results in a total amount of 50 sodium and 33 chloride ions which are added to the waterbox by replacing a water molecule each.

4.2.3 Results

In the following we first illustrate a modeling approach for arbitrary protein-DNA complexes. We compare our WRKY-DNA complex with results obtained by the state-of-the-art docking program *HADDOCK* [65, 67] and a recently published WRKY-DNA complex NMR structure (PDB id: 2lex) [309]. In the second part we describe an advanced explicit solvent MD relaxation and simulation protocol and its applicability to other protein-DNA complexes

4.2.3.1 Modeling arbitrary protein-DNA complexes

The modeling procedure described in the following is applicable to protein-DNA complexes which exhibit a similar initial situation we faced for the WRKY proteins. We successfully performed this approach for other TFs to model a protein-DNA structure, however, the results are not described in this work. The following two prerequisites must be fulfilled. Firstly, the three dimensional structure of the protein of interest is stored in the PDB or a good homology model can be constructed. Secondly, a closely related protein that is additionally bound to DNA is also available. Then the steps to model an arbitrary protein-DNA complex can be performed, as described in the following for the *At*WRKY1 cDBD-DNA complex.

1. *Given: One protein crystal structure*
   *At*WRKY1 cDBD structure (PDB id: 2ayd)

2. *Search the* Dali *webserver for similar protein structures, which are bound to DNA*
   The *Mm*GCM1-DNA complex (PDB id: 1odh [53]) was returned.

3. *Mapping the protein structure onto the protein-DNA complex*
   *At*WRKY1 cDBD is superimposed with the *Mm*GCM1 protein, illustrated in Fig. 4.4 (1).

4. *Constructing a protein-DNA complex*
   Combining the mapped *At*WRKY1 cDBD structure and the DNA double helix of the *Mm*GCM1-DNA complex in one PDB file.

5. *Base pair mutations*
   The DNA of the *Mm*GCM1-DNA complex is mutated in order that it features the WRKY DNA binding motif using a proposed experimental binding model [71]: 5'-AAAGTTGACCAAT-3'.

In 2012, an *At*WRKY4 cDBD-DNA complex structure obtained by NMR experiments was published (PDB id: 2lex). Thus, we can compare our *At*WRKY1 cDBD-DNA model, which was constructed without the knowledge of this protein-DNA complex to this NMR structure and verify our results.

We mapped our *At*WRKY1 cDBD-DNA complex model onto the *At*WRKY4 cDBD-DNA structure (PDB id: 2lex) using all phosphate atom positions of the W-box motif as input for the atom bijection method in *BALL* [114]. In Fig. 4.4 (2) the superposition of both complexes is shown and the six bp long W-box motif is illustrated in surface representation. Since the W-box motif was used as reference

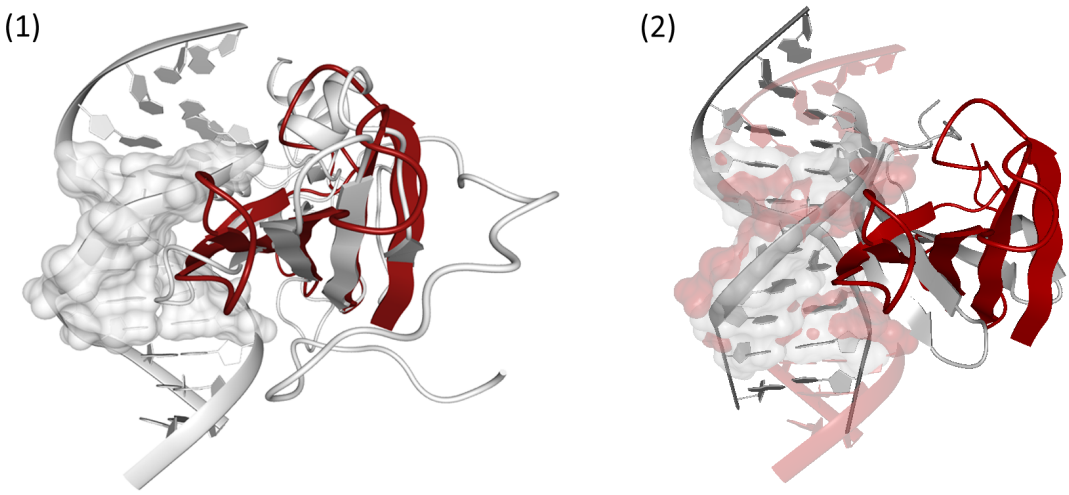


Figure 4.4: WRKY-DNA complex structures. (1) The *At*WRKY1 cDBD (in red) (PDB id: 2ayd) is superimposed with the *Mm*GCM1a protein bound to DNA (in gray) (PDB id: 1odh). The W-box sequence is indicated in surface representation. (2) In gray the NMR *At*WRKY4 cDBD-DNA complex structure (PDB id: 2lex) is shown superimposed to our computationally modeled *At*WRKY1 cDBD-DNA complexes in red.

in the mapping the surfaces of both DNA sequences overlay perfectly. The shape of the DNA double helix is different for the 2lex and the 1odh structures, which results in suboptimal structural matching of the terminal bps. The overall position of the *At*WRKY1 cDBD structure in our modeled complex is the same as the *At*WRKY4 cDBD structure position in the NMR structure with respect to the DNA binding motif. The essential difference lies in the orientation of the β-sheet plane of *At*WRKY1 cDBD relative to the DNA double helix. In our model the orientation between the DNA double helix and the β-sheet plane is more flat as compared to the β-sheet plane of *At*WRKY4 cDBD which is steeper (Fig. 4.4). Our modeled β-sheet plane resembles the β-sheet plane of *Mm*GCM1a in 1odh (Fig. 4.4 (1)).

In order to compare our modeling procedure to state-of-the-art methods we performed a docking run with *At*WRKY1 cDBD and the parsley promoter DNA sequence containing the W-box motif. The two biomolecules were docked using the webserver *HADDOCK*. In total 16 protein-DNA complexes were retrieved as output, whereas four protein-DNA complexes belong to the same cluster of structures. The best representatives of each cluster with respect to their energy values, resulting in four protein-DNA complexes, are selected and illustrated in Fig. 4.5. We mapped the four *HADDOCK* output complexes onto the *At*WRKY4 cDBD-DNA structure (PDB id: 2lex) as described for the mapping

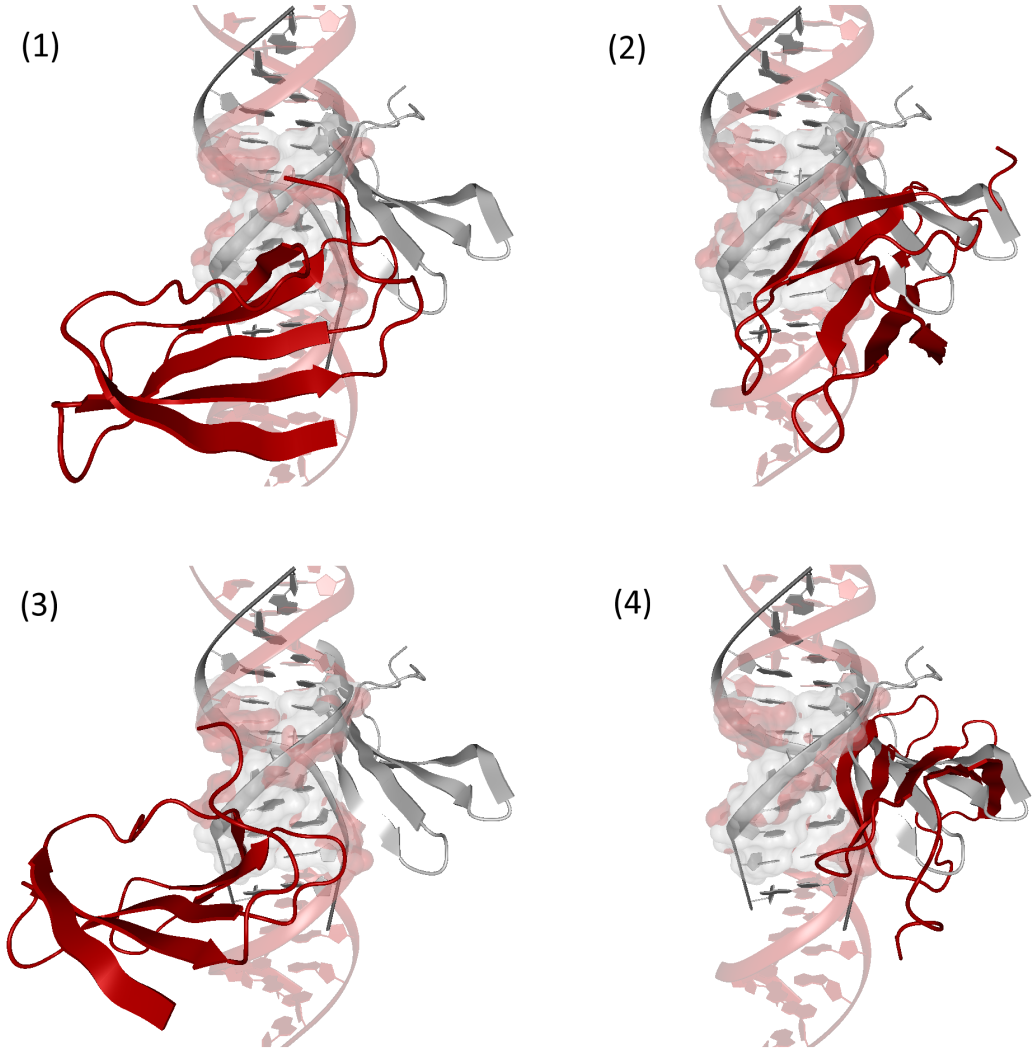


Figure 4.5: WRKY-DNA docking structures. In gray the NMR *At*WRKY4 cDBD-DNA complex structure is shown superimposed to docked protein-DNA complexes in red.

of our modeled protein-DNA complex. Only one docked complex structure features the protein at the correct position, however all β-sheets are facing the complete opposite direction (Fig. 4.5 (4)). In all other docking results the protein was not even placed at the correct position, which can be seen in Fig. 4.5 for (1), (2), and (3). RMSD values calculated between the modeled and the NMR protein-DNA complex confirm the visual inspection. We include 32 C_α atoms of selected β-sheet residues in the RMSD calculation (listed in Table B.3, which yield the following values: (1) 28.59 Å, (2) 17.03 Å, (3) 32.52 Å, and (4) 13.20 Å. For our protein-DNA complex we yield an RMSD of 13.02 Å for the 32 C_α atoms. Since our protein-DNA complex yields the lowest RMSD value and exhibits the WRKY protein at the correct position and orientation we can conclude that our modeling approach yields a valid protein-DNA complex.

4.2.3.2 Advanced molecular dynamics simulation protocol

Flexibility needs to be taken into account while studying protein-DNA complexes to gain a more accurate picture. In the following sections we describe an MD simulation protocol which induces flexibility to protein-DNA complexes. The MD relaxation and simulation protocol for protein-DNA complexes consists of three basic steps: minimization, relaxation, and production run. The parameter and topology files generated by *LEaP* are used as input files for the *AMBER* simulation program *sander* [39]. The structures are first minimized to remove initial steric clashes. After energy minimization, MD simulations are performed to equilibrate the solvent density. The structure is iteratively relaxed moving the atoms of the system along the negative energy gradient until a sufficiently low energy is obtained. MD samples more configurational space than minimization and allows structures to cross over small potential energy barriers. Conformations are sampled at regular time points during a simulation and stored (as snapshots). These snapshots can then be studied to obtain detailed information, especially about interactions at the binding interface. The single steps of this explicit solvent MD simulation protocol using the *AMBER* simulation program *sander* are described in the following (and details are listed in Table B.6 and Table B.7 in the appendix).

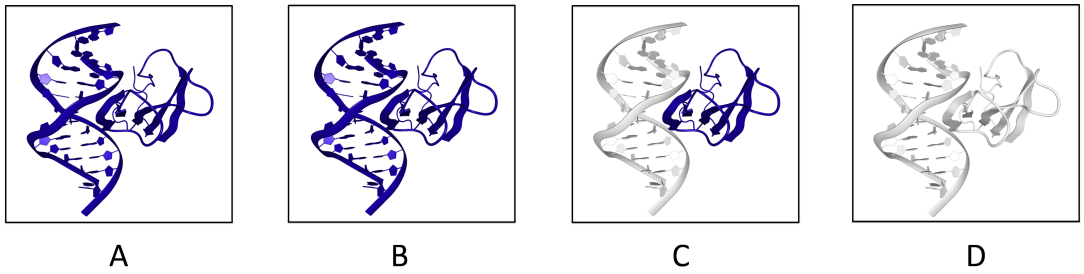
Minimization

Each minimization (A-D in Fig. 4.6) is performed using 4,000 steps of steepest descent (SD) algorithm followed by 1,000 steps of conjugate gradient (CG). The basic parameters are equal for all minimization steps (A-D), while the number of constrained atoms differ. During the first minimization step (A), all atoms of the protein-DNA complex are constrained with positional constraints. In the second minimization step (B), all solute heavy atoms are constrained. During the third minimization step (C), only main chain atoms of the protein are constrained and in the last step (D) all atoms of the system are free to move. The minimization steps try to remove initial steric clashes and provide a structure which is suitable as input for an MD simulation.

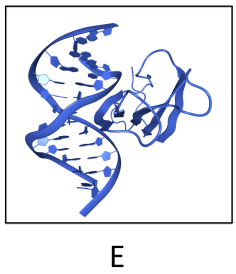
Equilibration (NVT-MD)

After minimization, an equilibration step (step E in Fig. 4.6) with constant volume NVT-MD is carried out for 100 ps, during which the system is heated up from 100 K to 300 K. Every 1,000 steps the translational center-of-mass motion is removed [42] to avoid energy drains [49, 107]. When 300 K are reached, the system is kept at this temperature using a canonical (constant T) ensemble. For temperature regulation, Langevin thermostat with a collision frequency 2.0 ps^{-1} is applied. The particle mesh Ewald (PME) method [61] is used to treat long range electrostatic interactions. SHAKE [250] is applied to constrain bond length involving bonds to hydrogen atoms, therefore a time step of 2 fs is satisfying. Since this is an explicit water MD simulation and SHAKE is applied, a special three-point algorithm [200] is used. All DNA double helix atoms and protein atoms are constrained with an harmonic force constant of 5 kcal/(mol·$Å^2$) during the NVT-MD.

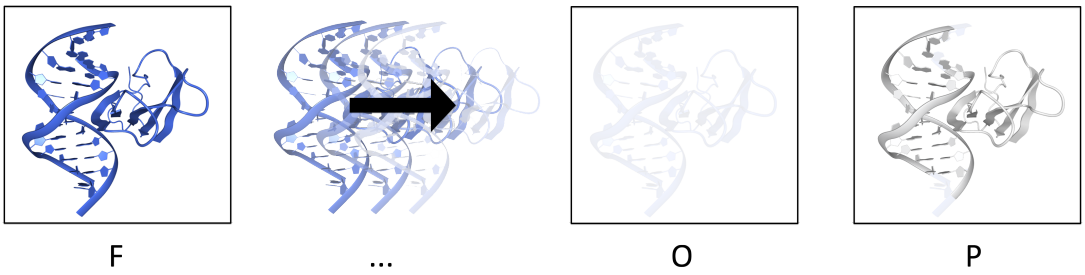
1. Minimization (A-D)



2. Equilibration, NVT-MD (E)



3. Equilibration, NPT-MD (F-P)



4. Production run (Q)

Figure 4.6: MD simulation procotol. Initially, the system is minimized in four steps (A-D). Then it is equilibrated at constant volume (NVT-MD), shown in step E. Steps F to P indicate the equilibration at constant pressure (NPT-MD). Finally, the production run is performed in step Q. The constrained atoms are indicated in blue, whereby the lighter the color blue the lesser constraint are the atoms. However, since no H-atoms are represented in detail it is only an approximate illustration.

Equilibration (NPT-MD)

The setup for the next ten equilibration steps (steps F-O in Fig. 4.6) are basically the same, only the weight for the positional constraints differ, which is gradually reduced. In the last step (step P) only the flanking bps of the DNA double helix are constraint. In total eleven steps (steps F-P) with constant pressure using isotropic position scaling and constant temperature NPT-MD at 300 K are performed to adjust the solvent density. The first step (step F) is 100 ps long and all following steps (steps G-P) 40 ps. Again, every 1000 steps the translational center-of-mass motion is removed. The velocity information is read in from the previous equilibration step in order to start the second equilibration at the point the constant volume equilibration ended. For a periodic boundary system a constant pressure simulation [17] is the only way to equilibrate the density. For example, when placing the biomolecule in a water box using *LEaP* it sometimes happens that gaps between water molecules arise, which can result in holes during constant volume equilibration. Those holes need to be filled again with water molecules.

Production run

Finally, an MD simulation (step Q in 4.6) at 300 K is carried out for 20 ns. Only the two flanking bps of the DNA double helix, four nucleotides in total, are constrained with a harmonic force constant of 1 kcal/(mol·Å^2). This simulation is called production run, since the system should now be in equilibrium and stable over time.

The quality of the protocol is shown in Fig. 4.7. The RMSD values over the time span of the whole production run show overall stability for the *At*WRKY1 cDBD-DNA complex. The protein C_α atoms are stable as well as the DNA double helix. To verify that our relaxation and simulation

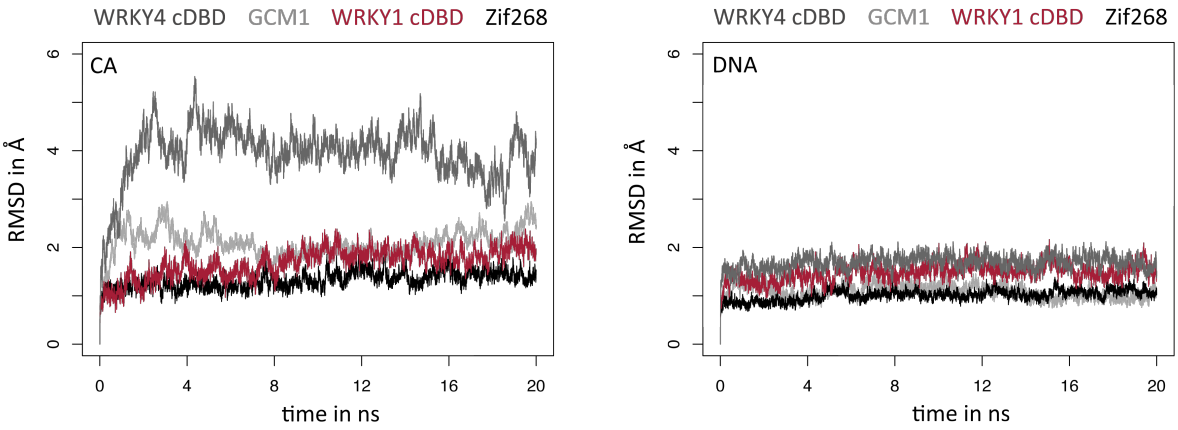


Figure 4.7: RMSD values during MD simulations. The RMSD values for all C_α atoms and all DNA atoms in reference to the initial structure of the production run are calculated for four protein-DNA complexes: WRKY1 cDBD (our model), Zif268 (XRD), GCM (XRD), and WRKY4 cDBD (NMR). The modeled protein-DNA complex acts as a high-quality, experimental XRD protein-DNA complex in contrast to the NMR structure which undergoes large structural changes.

protocol is not only suitable for the modeled *At*WRKY1 cDBD-DNA we applied it to the following protein-DNA complexes: the GCM-DNA complex (PDB id: 1odh), the Zif268-DNA complex (PDB id: 1aay), and the *At*WRKY4 cDBD-DNA complex (PDB id: 2lex). For the two XRD protein-DNA complexes we could also show that they are stable and keep their initial structures apart from some structural alterations especially in the loop regions and at the C- and N-terminal domains (cDBDs and nDBDs) of the proteins. Even, the DNA double helix is stable throughout the simulation and all zinc ions remain at their initial positions. However, the NMR protein-DNA complex (PDB id: 2lex) undergoes more structural alterations than the two protein-DNA complexes obtained by XRD or the modeled protein-DNA complex. Unfortunately, we observe that not only the protein C_α atoms are highly flexible, but also the DNA does not retain its perfect double helix structure. This could not be observed by solely analyzing RMSD values. We examined snapshots of the production run and for the *At*WRKY4 cDBD-DNA complex and observed that H-bonds between bases were broken that leads to a distorted double helix structure, see Fig. 4.8.

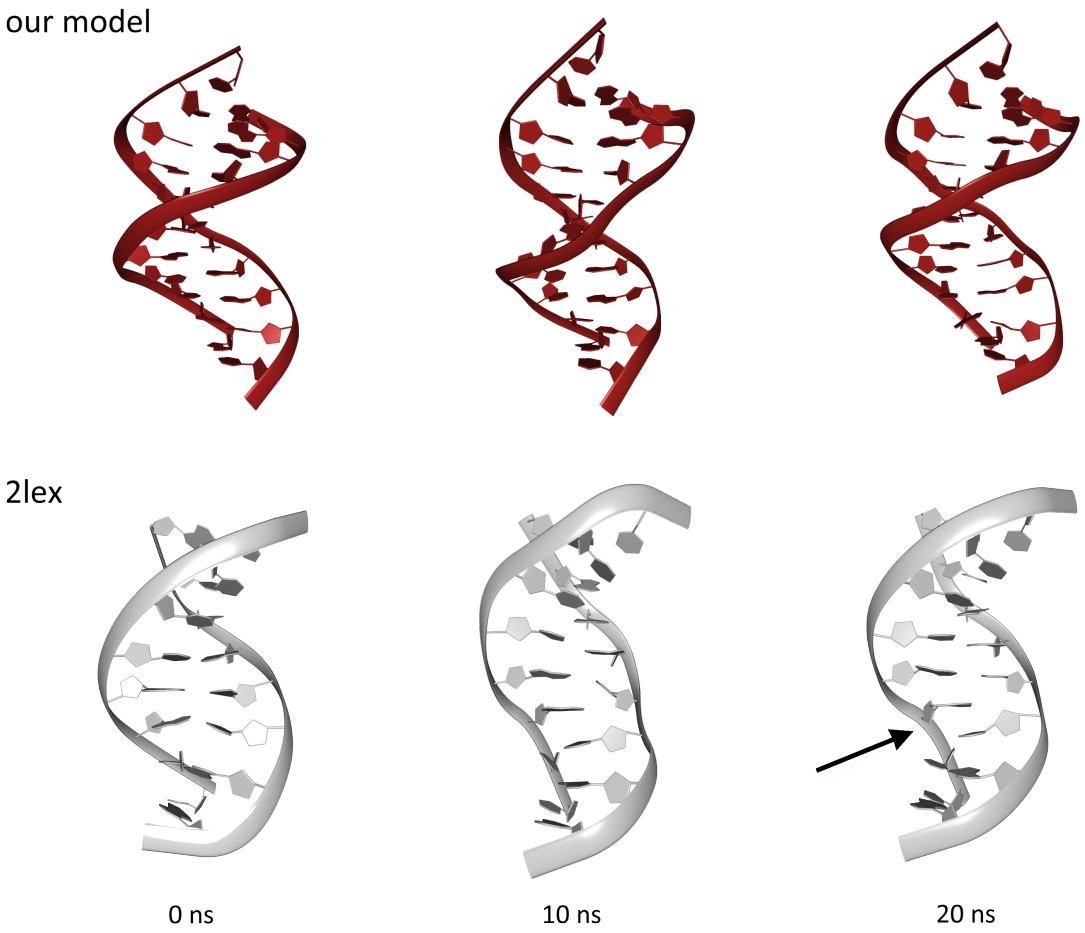


Figure 4.8: DNA structure during MD simulation. DNA double helices of our *At*WRKY1 cDBD-DNA complex (in red) and of the *At*WRKY4 cDBD-DNA (PDB id: 2lex, in gray) are illustrated. The DNA structures were extracted from the MD simulation production run at 0, 10, and 20 ns. The arrow points to a broken bp in 2lex.

The RMSD values for the modeled *At*WRKY1 cDBD-DNA complex indicate a stable MD simulation. To illustrate this fact, we extracted snapshots at certain time points of the production run of the modeled WRKY-DNA complex. This detailed inspection should give structural insights if the MD simulation protocol creates valid protein-DNA complexes. Comparisons between the *At*WRKY4

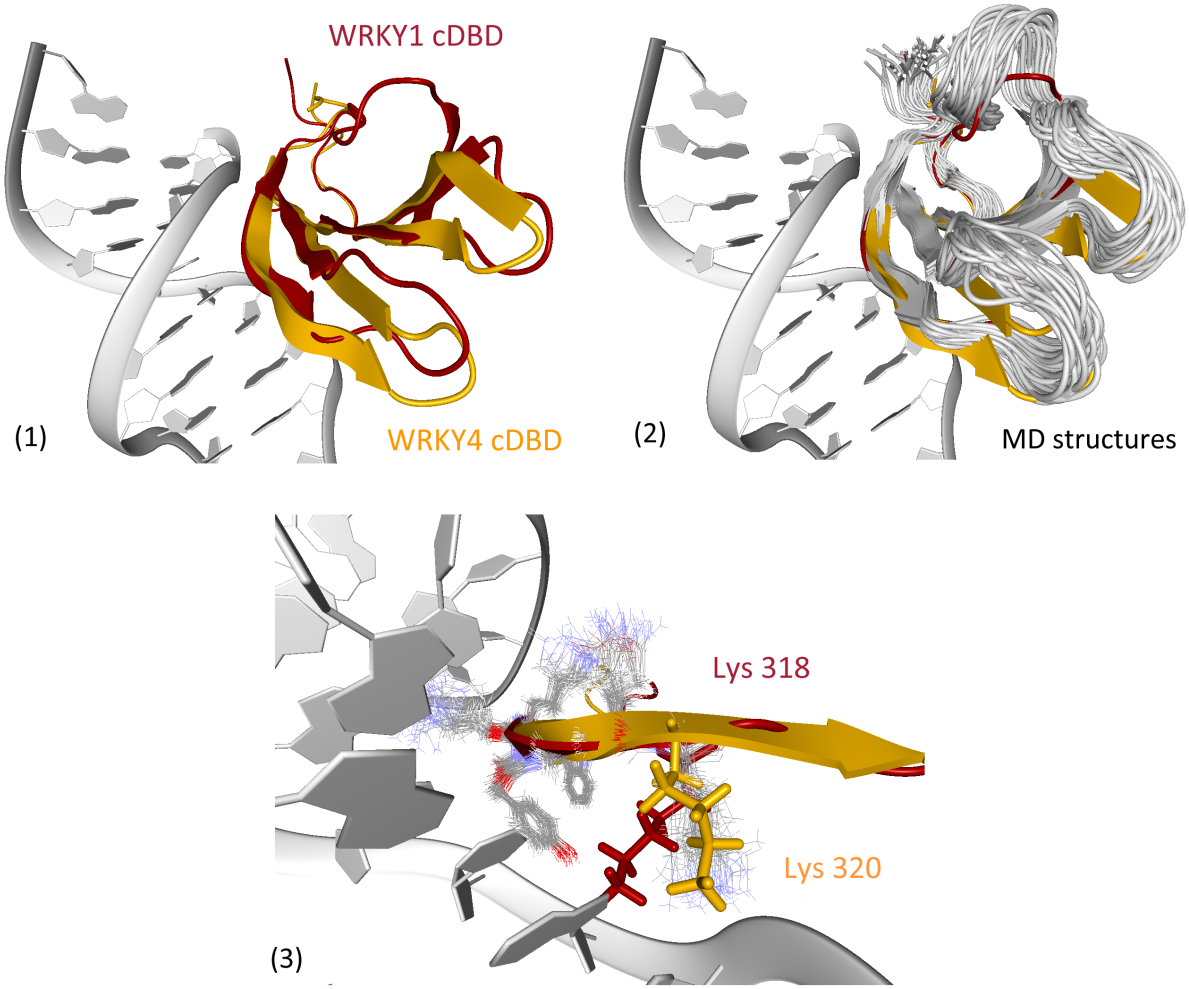


Figure 4.9: Refined *At*WRKY1 cDBD structure The NMR protein-DNA complex is shown in (1) (*At*WRKY4 cDBD in yellow and the DNA in gray) superimposed with the starting structure of *At*WRKY1 cDBD in red. In (2) *At*WRKY1 cDBD protein structures extracted from the MD simulation are shown in gray which move toward the yellow *At*WRKY4 cDBD structure. (3) Not only the backbone atoms are refined, but also side chains move toward preferred orientations. The initial orientation of lysine (red) moves during the MD simulation (snapshots of lysine shown in line mode) toward the lysine orientation of *At*WRKY4 cDBD (yellow).

cDBD-DNA NMR structure and the snapshots extracted between 15 and 20 ns for the *At*WRKY1 cDBD-DNA complex show that the *At*WRKY1 cDBD protein obtains a similar structure, see Fig. 4.9 (1) and (2). In Fig. 4.9 (1) the input structure of *At*WRKY1 cDBD superimposed with the *At*WRKY4 cDBD structure is shown. The four β-strands are superimposed perfectly, but the loop regions differ. In Fig. 4.9 (2) the dynamics of the loop regions are visible, while the atoms of the β-strands show not

as much flexibility during the last 5 ns of the simulation. The loop region atoms of *At*WRKY1 cDBD shift to positions comparable of *At*WRKY4 cDBD. Residues which were supposed to interact with the DNA shift toward the double helix and form contacts, see Fig. 4.9 (3). This can clearly be seen for lysine 318 of *At*WRKY1 cDBD that shifts towards the position of lysine 320 of *At*WRKY4 cDBD. This indicates that residues at the surface of *At*WRKY1 cDBD (PDB id: 2ayd) are now refined with respect to its DNA binding partner. Thus, the flexibility induced by our advanced MD relaxation and simulation protocol yields an accurate WRKY-DNA complex. Especially, the binding interface can now be studied in detail.

4.2.4 Discussion

In the first part of this section we introduced a molecular modeling approach for modeling arbitrary protein-DNA complexes. During this first modeling approach no structural flexibility was included. The approach performs well and even outperforms a state-of-the-art docking program. We include as much structural information as possible in our modeling procedure and could show that it results in a good model for the WRKY-DNA complex.

When comparing our complex model to the NMR structure, published only recently in 2012, a few differences appear which are discussed in the following. We tried to superimpose the WRKY structure perfectly with the *Mm*GCM1a protein. Therefore, the orientation of the β-sheet plane characterizes in our model the orientation of the *Mm*GCM1a protein. The angle between the β-sheet plane and the DNA is much smaller in the *At*WRKY4 cDBD-DNA complex, as compared to the GCM1-DNA complex. However, we would like to point out that *At*WRKY1 and *At*WRKY4 belong to the group I subfamily class of WRKY proteins, which feature two WRKY-DNA binding domains (an N-terminal and C-terminal DNA binding domain). The structure of the N-terminal domain as well as the whole protein structure of both proteins is not available in databases. The complete protein structure might interact with the DNA even differently as proposed in the NMR *At*WRKY4 cDBD-DNA structure. Since studies conclude that the N-terminal domain interacts also with the DNA the "steepness" or "flatness" of the C-terminal DBD β-strands could be affected. Both domains are connected by a linker which determines the length and probably the orientation of both DBDs, but we can only speculate about the true binding of *At*WRKY1 or *At*WRKY4 proteins to the DNA. Since, the *Mm*GCM1a protein exhibits two domains one of which is a closely related five β-strands in WRKY proteins it serves as perfect initial template structure for modeling a WRKY-DNA complex.

Compared to the docking program *HADDOCK* we yield a protein-DNA complex, which represents a three dimensional structure closer to experimental complexes. However, we did not perform the advanced docking procedure, whereby flexibility is included. When comparing our modeling results to the docked protein-DNA complexes we restricted the evaluation of both methods to the correct overall placement of the protein onto the DNA double helix regarding its position and orientation. Since, our own simple modeling approach does not account for structural alterations we wanted to use a comparable docking setup for creating protein-DNA complexes. One task which seems almost impossible to achieve in protein-DNA docking is the placement of the TF at its DNA sequence motif,

since there exist too many possibilities and the major groove of the DNA has no clear distinguishable features. Our modeling approach uses a related protein, already bound to the DNA strand, which makes the placement process tremendously easier. An additional advantage using or modeling procedure is that it always returns the same protein position and orientation relative to the DNA double helix. A docking program such as *HADDOCK* would return different possible orientations. When one would like to compare similar protein structures and their specificities at the binding interface our method is more suitable than docking. However, it should be noted that our approach requires a lot of prior information. When no closely related protein-DNA complex is available, our method is not applicable.

Mapping a protein structure onto a protein which is bound to DNA gives a first idea how the protein is placed on the DNA, but gives no or only little information about specific interactions at the binding site. Most protein side chains of the mapped protein are not oriented toward the double helix or in a worst case they overlap with DNA atoms. Therefore, we introduce an explicit solvent MD relaxation and simulation protocol. Molecular modeling methods which account for flexibility at protein-DNA interfaces typically use implicit solvent simulations and constrain protein and DNA backbone atoms [8, 266, 311]. Comparable to these approaches we are able to mutate amino acids and DNA bps at the binding interface. The DNAMutator in *BALL* places the bps in exact the same plane as the original bp and preserves the backbone atoms. When mutating protein side chains the rotamer closest to the original side chain position is selected by default. It is possible to choose another rotamer in order to avoid steric clashes with DNA atoms. With or without mutations DNA needs a careful relaxation protocol during MD simulation, in which Cartesian constraints are reduced step-by-step, otherwise the two DNA strands drift apart. The negatively charged DNA backbone, only two or three H-bonds between bases, and π-π-stacking of bps which is not properly modeled in force fields make DNA molecules a difficult simulation object. However, we are able to simulate protein-DNA complexes without constraints, apart from a marginal constraint at the two flanking bps of the DNA double helix and are able to model truthful representations of biomolecular structures surrounded by water molecules and counter ions. Our modeled WRKY-DNA complex is stable and the protein does not drift away from the DNA or alters its position drastically. The "flatness" of the β-sheet plane is also maintained. The reason for this might be that we used the DNA helix structure which was optimized with respect to the *Mm*GCM1a protein. Comparing the DNA structure of the *Mm*GCM1a protein-DNA complex to the *At*WRKY4 cDBD-DNA complex the major groove looks slightly different. However, the extent of this discrepancy is comparable to structural differences between the different NMR models. Surprisingly, the NMR model of the *At*WRKY4 cDBD-DNA complex was not stable throughout the production run. One reason might be that the initial structure of the NMR model is not perfectly suitable for MD simulations. However, our modeled *At*WRKY1 cDBD-DNA complex structure performs as good as the XRD protein-DNA complexes. In contrast to other molecular modeling approaches we cannot only insert single mutations at the binding interface, but can also apply our approach to study different proteins bound to the same DNA sequence motif or mutate all bps within the DNA sequence. Since we account for overall flexibility these mutated

structures are able to relax during our advanced explicit solvent MD simulation protocol.

During the production run of the MD simulation we could observe that side chains at the binding interface shift toward the positions of *At*WRKY4 cDBD side chains. Using explicit solvent simulations we could additionally study the distribution of water molecules and ions along the binding interface. Using the MD simulation protocol we are one step closer toward representing a truthful structural image of a protein-DNA complex. Common molecular modeling approaches, which incorporate flexibility in native or mutated protein-DNA complexes, are mostly limited to implicit solvent simulations or frozen protein and DNA backbones which might lead to bias in binding site prediction. We were able to establish a protocol without these limitations, which can be used for studying interactions at the binding interface and predicting specificity of different WRKY proteins.

4.3 Studying specific features at binding interfaces of WRKY-DNA complexes

In the first part of this chapter we described a modeling approach and a stable MD relaxation and simulation protocol for protein-DNA complexes. Both procedures are universally valid and now applied to a specific case study. In the following, we identify the optimal bp assembly of the DNA double helix for WRKY proteins in three dimensions and detect unique features of WRKY proteins at the binding interface using molecular dynamics. In the first part of this section, binding affinities and specific interactions between different double helix structures and *At*WRKY1 cDBD are determined. Then, in order to be able to highlight specific features of different WRKY proteins at the protein-DNA binding interface we have been creating homology models for *At*WRKY11, *At*WRKY50, *At*WRKY33 cDBD, and *At*WRKY33 nDBD. In the last part of this section, we identify differences between *At*WRKY11- and *At*WRKY50-DNA complexes at the binding interface and study the structures of *At*WRKY33 cDBD and *At*WRKY33 nDBD proteins.

4.3.1 Introduction

WRKY proteins comprise a large family of TFs. Although it is known that they are, for instance, important for pathogen defense in plants, the specific function for most individual proteins of this family is unclear. Since their sequence is highly conserved, it is assumed that their structures are also very similar. Additionally, it is common knowledge that all WRKY proteins favor a certain sequence motif, termed the W-box: 5'-TTGACC-3'. Does this imply that possibly all 72 WRKY

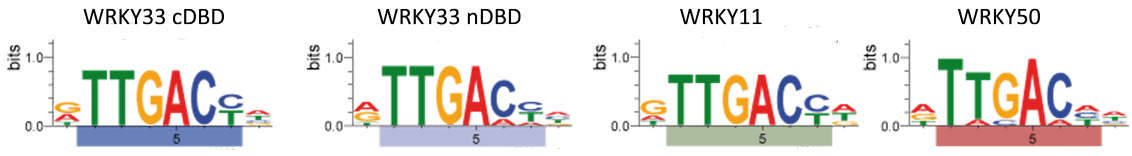


Figure 4.10: DNA sequence logos. DNA-binding studies reveal that WRKY proteins bind to a consensus sequence 5'-TTGACY-3', whereby WRKY50 proteins are also able to bind to 5'-TWGACY-3'. DNA sequence logos of different WRKY proteins show varying conserved sequence positions.

TFs in *Arabidopsis thaliana* bind to more or less identical DNA sequences? One might ask if WRKY proteins are able to regulate interchangeably the same genes and have the same function. But, why do 72 WRKY proteins exist in *Arabidopsis thaliana* or sometimes over 100 different WRKY proteins in other species? During the last years functional studies have been performed in combination with sequence-based computational approaches to answer these questions. The W-box sequence motif and its flanking bps were thoroughly analyzed by mutation studies [50]. The results of these investigations conclude that there are slightly altered DNA sequence motifs for different WRKY proteins, which are illustrated in Fig. 4.10. The last position in this DNA sequence can be variable, which results in: 5'-TTGACY-3', whereby Y is basically either C or T. A WRKY protein forming specific contacts to

the W-box sequence (5'-TTGACC-3') is illustrated in Fig. 4.11. It is still not completely evident why certain WRKY proteins favor a certain DNA sequence motif over another. Therefore, we examined protein-DNA complexes of selected WRKY proteins at the atomic level. We could identify amino acids at the binding interface of these protein-DNA complexes, which form specific interactions with bps of the DNA sequence. We performed *in silico* mutation studies to analyze our observations in more detail, especially for *At*WRKY11 and *At*WRKY50 and our findings could be confirmed by experimental binding affinity studies.

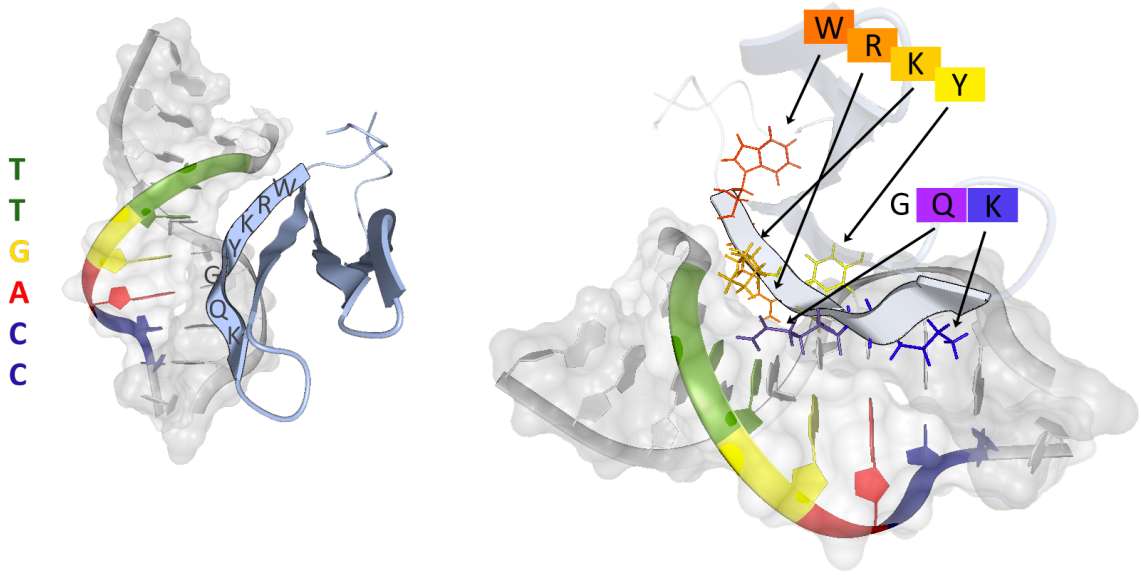


Figure 4.11: WRKY-DNA binding interface. The DBDs of WRKY proteins feature a β-strand that is formed by the following amino acids: WRKYGQK. This β-strand lies in the major groove of the DNA double helix and makes contacts with the DNA binding motif: 5'-TTGACC-3'.

WRKY proteins can be grouped into different categories due to their overall sequence similarities and the structural motif of their zinc ion coordinating residues. WRKY proteins that are part of group I comprise two structural WRKY domain subunits: one N-terminal and one C-terminal DNA-binding domain (nDBD and cDBD). WRKY proteins of group II and III feature just one WRKY DBD. It is assumed that this DBD is for most WRKY proteins more similar to the C-terminal DBD of group I proteins. Up to this date the complete structure of group I WRKY proteins or the N-terminal subunit has not been experimentally obtained and it is also not fully understood whether or not and if so how it interacts precisely with the DNA double helix. Studies identified that *At*WRKY33 nDBD binds to DNA, but how this is accomplished remains unclear. In order to shed light on the structural differences of the C- and N-terminal domain we modeled the structures of *At*WRKY33 cDBD and *At*WRKY33 nDBD and built protein-DNA complexes using these homology models.

4.3.2 Materials and Methods

Sliding along the DNA

To identify the correct placement of a WRKY protein along the DNA major groove we performed a specific procedure. The protein slides along the DNA major groove in 5'-3' direction of one strand and then in 5'-3' direction of the opposite strand in step size of one bp. The parsley promoter sequence (5'-TCAAAGTTGACCAATAAT-3') [248] serves as a reference DNA for modifying DNA bps. This sequence includes the previously described WRKY target DNA sequence, also known as the W-box (5'-TTGACC-'3). The DNA double helix of *At*WRKY1 cDBD bound to the TFBS of *Mm*GCM1a (present in 1odh) is mutated using the DNAMutator method in *BALL* (version 1.4.2) [114]. The DNAMutator method builds a new bp by keeping DNA backbone atoms, deleting nucleobase atoms, and inserting atoms of the new nucleobases in the same plane as the old ones. The AT bp at the 3' position of the modeled WRKY-DNA complex was deleted. The *At*WRKY1 cDBD structure was kept at the same

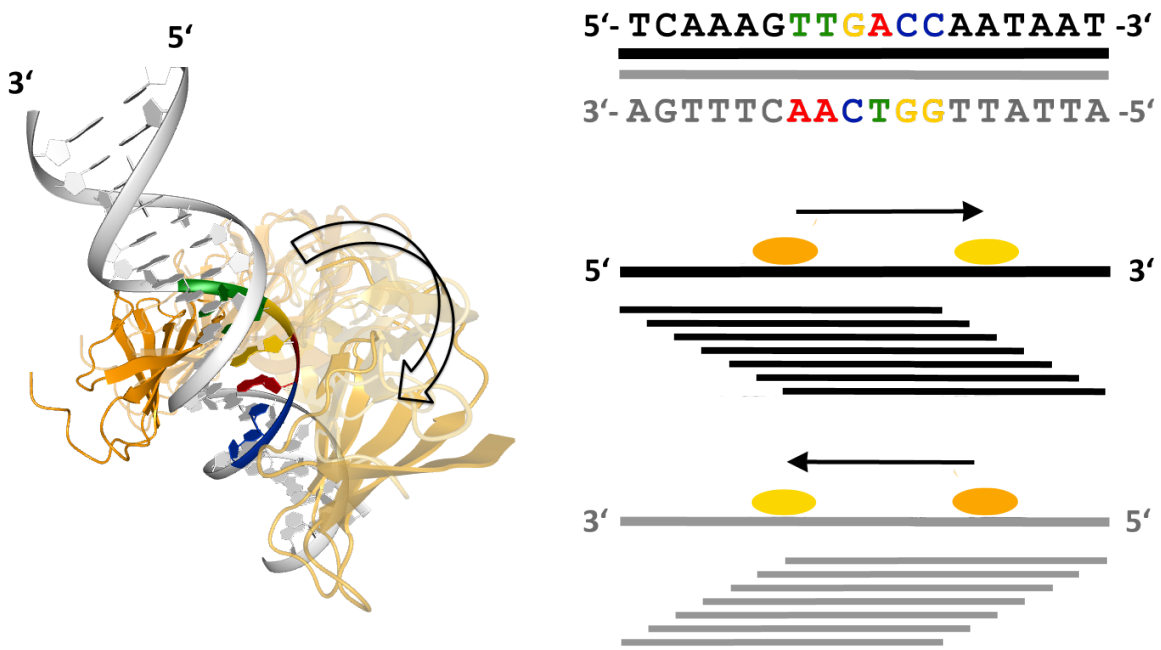


Figure 4.12: WRKY protein slides along DNA. Binding free energies between the *At*WRKY1 cDBD and fourteen different DNA sequences were estimated using the MM-PBSA approach. The standard deviations for each binding free energy are also determined. The highest binding free energies are determined for sequence 3 (5'-AAAGTTGACCAA-3') and 10 (5'-TATTGGTCAACT-3', antisense strand: 5'-AGTTGACCAATA-3'). The position of the TF relative to the W-box motif is the same for both structures, only the sequence direction differs.

relative position the the 12 bp long DNA double helix, thereby always placed closest to the sequence positions five through ten. The first complex has the DNA sequence: 5'-TCAAAGTTGACC-'3, the second: 5'-CAAAGTTGACCA-3', that means the W-box motif shifts from the end to the beginning

of the sequence or the protein slides along the DNA double helix in 5'-3' direction, as illustrated in Fig. 4.12. When the WRKY protein slides along the opposite strand direction, again in 5'-3' direction, the eighth complex has the DNA sequence: 5'-ATTATTGGTCAA-3', the ninth complex: 5'-TTATTGGTCAAC-3', and so on. The antisense DNA strand of the eighth complex has the DNA sequence: 5'-TTGACCAATAAT-3', the night complex: 5'-GTTGACCAATAA-3', which is illustrated as gray lines in Fig. 4.12.

Molecular dynamics and binding free energies

In order to identify the most probable protein-DNA complex we performed MD simulations for all of the 14 different protein-DNA complexes and calculated their binding affinity with the MM-PBSA approach, provided by *AMBER 11* [39]. The simulations were performed using a similar protocol as described in Section 4.2.3.2 with one slight addition: in order to be able to calculate binding affinities for the same binding interfaces the protein structures were constrained during the production run. 23 β-sheet C_α atoms (residue ids: 25-47, 61-64, 72-76, and 86-100) were hold fix with an harmonic force constant of 1 kcal/(mol·Å^2). The last step of the MD simulation is 20 ns long and we calculated binding free energies over the last 5 ns, after the complexes reached an equilibrium state. The *AMBER* MM-PBSA perl script was used to extract 26 complexes evenly spread over the last 5 ns of the production run. Standard parameters (listed in Table B.8 in the appendix) were used to calculate the binding affinities for each complex.

Simple contact count

Interactions between protein and DNA atoms can be analyzed in a very simple way by counting contacts between these two biomolecules. The distance cutoff was set to 2.5 Å. Additionally, contacts were only counted if the distance was greater than 1 Å, since close contacts are not favorable for binding. It was also distinguished between DNA backbone and DNA nucleobase contacts to identify the number of specific interactions. All interactions were analyzed over a certain time span of an MD simulation production run from 15 to 20 ns. Snapshots of protein-DNA complexes were taken every other picosecond which results in 2500 complexes for each protein-DNA structure. Additionally, we only count contacts that were lasting for at least 80% of the 5 ns time span.

Homology modeling

In order to be able to study the binding interfaces of different WRKY proteins with DNA, homology modeling was performed to obtain other WRKY protein structures. Amongst others we modeled protein structures for *At*WRKY11, *At*WRKY50, *At*WRKY33 cDBD, and *At*WRKY33 nDBD with Prime (version 3.0, build v301111) [129, 130] using *At*WRKY1 cDBD (PDB id: 2ayd) as template structure for all four homology models. The multiple sequence alignment used for the respective sequence alignment is illustrated in Fig. 4.1. The sequences of WRKY proteins are highly conserved, which is reflected by high sequence identities (calculated by *ALIGN* [228]) between the template

sequence of *At*WRKY1 cDBD (PDB id: 2ayd) and the other sequences, see Table 4.1. Validation of the homology models with *Anolea* [195–197], *ProSA* [268, 303], and *what-check* [121] confirm an excellent quality of the obtained models (validation results listed in Table B.10 and shown in Fig. B.2 in the appendix). In addition to these statistical analysis tools, all modeled protein structures were simulated in explicit solvent for 20 ns using a standard MD simulation protocol for proteins in *AMBER* [39]. Throughout the simulation the proteins were highly stable and no major conformational changes were observed. In order to gain TF-DNA complexes, the modeled protein structures were superimposed onto our previously modeled WRKY-DNA complex structure. The *At*WRKY1 cDBD protein bound to the DNA sequence: 5'-AAAGTTGACCAA-3'.

Table 4.1: WRKY homology models. Homology modeling was performed using *At*WRKY1 cDBD as template to obtain structures of four WRKY DBDs.

Protein	Structure	Locus name	UniProt ID	Sequence	Identity
WRKY1 cDBD	PDB id: 2ayd	*At*2g04880	Q95137	Ser 293 - Gly 368	-
WRKY11	homology model	*At*4g31550	Q9SV15	Val 232 - Met 308	50.6 %
WRKY50	homology model	*At*5g26170	Q8VWQ5	Gly 99 - Asn 173	50.0 %
WRKY33 cDBD	homology model	*At*2g38470	Q8S8P5	Pro 348 - Ala 423	67.1 %
WRKY33 nDBD	homology model	*At*2g38470	Q8S8P5	Asn 170 - Ser 244	46.1 %

4.3.3 Results

4.3.3.1 Structural analysis of *At*WRKY1 cDBD-DNA structures

The binding free energies between the *At*WRKY1 cDBD protein and different DNA sequences are shown in Fig. 4.13 (and in Table B.9 in the appendix). We identified two DNA sequences with similar low binding free energies, whereas all other DNA sequences seem to have higher binding free energies to the *At*WRKY1 cDBD protein. The 5'-AAAGTTGACCAA-3' and 5'-TATTGGTCAACT-3' DNA sequences achieve the lowest binding free energies. The latter DNA sequence comprises the W-box motif at exactly the same position as the 5'-AAAGTTGACCAA-3' sequence only oppositely oriented, which means that the WRKY protein binds to the same sequence motif (5'-TTGACC-3'), but in the opposite direction. Since both MM-PBSA energy values are similar we can not identify a preferred binding orientation, when relying on these values alone. We try to discriminate the binding preference by determining the contacts which were formed during the same time in the production run. The overall number of contacts between WRKYGQK residues with the W-box motif is eleven as opposed to eight comparing the 5'-AAAGTTGACCAA-3' and the 5'-TATTGGTCAACT-3' sequence. Additionally, we could confirm the proposed contacts of Duan *et al.* [71], where they identified six specific contacts. Five of the six specific contacts to the W-box agree for the 5'-AAAGTTGACCAA-3'

sequence. The 5'-TATTGGTCAACT-3' sequence could only form one out of six. Therefore, we chose the 5'-AAAGTTGACCAA-3' DNA sequence as preferred binding motif for WRKY proteins.

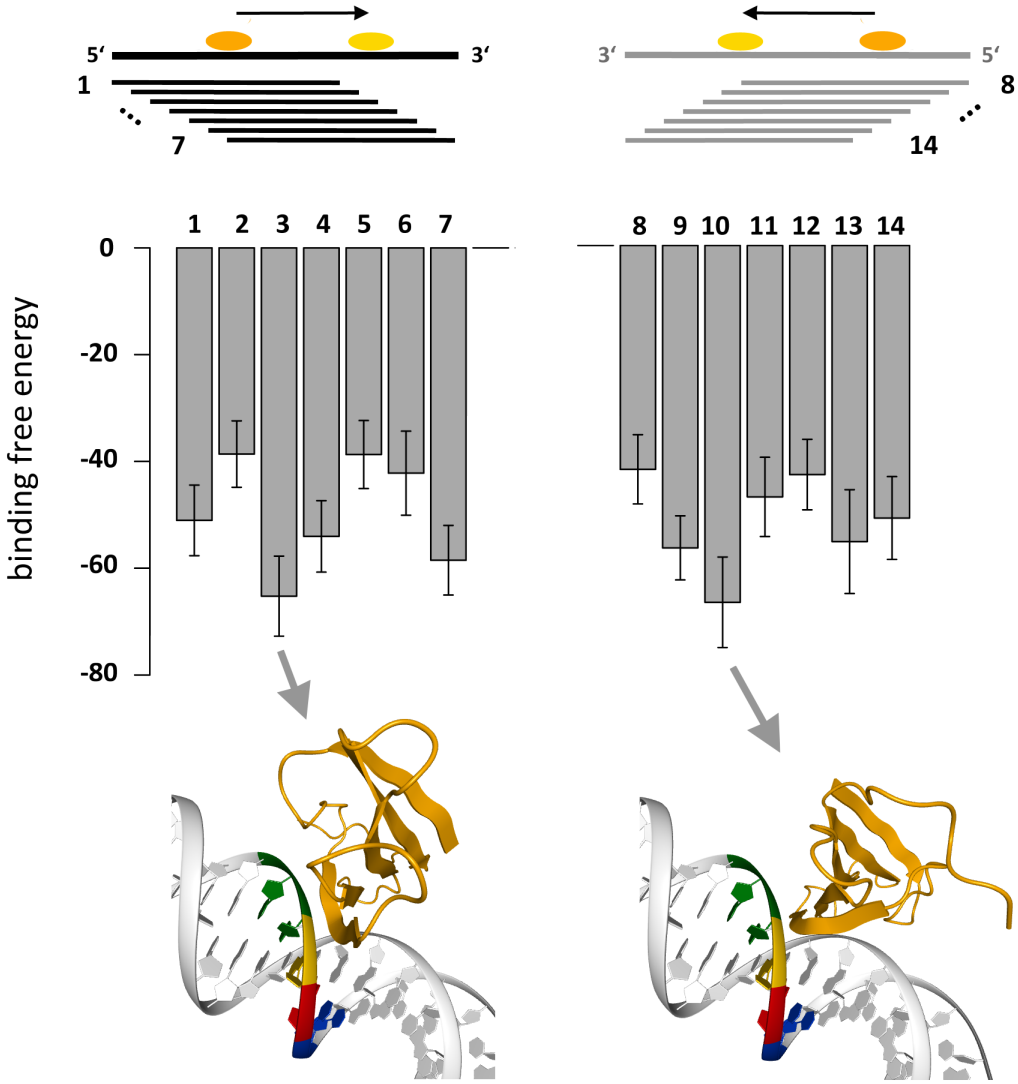


Figure 4.13: Binding free energies between *At*WRKY1 and DNA. Binding free energies between the *At*WRKY1 cDBD and fourteen different DNA sequences were estimated using the MM-PBSA approach. The standard deviations for each binding free energy are also determined.

4.3.3.2 Binding specificities of *At*WRKY11 and *At*WRKY50

In order to be able to identify binding specificities of WRKY proteins at the binding interface, we mapped the homology models of *At*WRKY11 and *At*WRKY50 onto the *At*WRKY1 cDBD protein bound to the 5'-AAAGTTGACCAA-3' sequence. After modeling the two complex structures, they were simulated with the same MD simulation protocol as the *At*WRKY1 cDBD-DNA complexes for

identifying the correct binding site. *At*WRKY11 is a highly specific WRKY protein binding almost exclusively to the 5'-TTGACC-3' motif [31, 50], whereas *At*WRKY50 apparently binds to varying motifs with similar affinities [31]. *At*WRKY50 features a WRKYGKK binding sequence as compared to the WRKYGQK sequence for most WRKY proteins, such as *At*WRKY11. DNA-binding studies showed that WRKY proteins bind to a conserved 5'-TTGACY-3' binding consensus. In contrast, only the clade of *At*WRKY50 is able to bind to the 5'-TWGACY-3' as well [31]. Since, the glutamine in *At*WRKY11 and the lysine in *At*WRKY50 have been the only positions in the binding site which differ drastically in their amino acid properties between these two proteins, we analyzed the proteins at this position in more detail. We mutated the wild-type *At*WRKY11 WRKYGQK motif to WRKYGKK and the *At*WRKY50 WRKYGKK motif to WRKYGQK and simulated these mutated protein-DNA complexes as well. During the MD simulations the mutated amino acid side chains were able to gain the preferred orientation. Comparing the wild-type and mutated *At*WRKY50-DNA complexes we

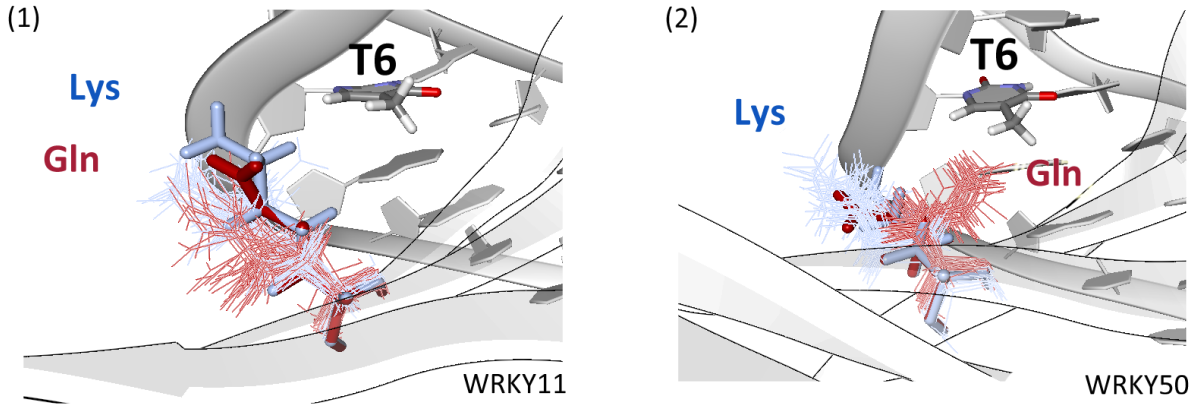


Figure 4.14: The *At*WRKY11 and *At*WRKY50 protein bind to DNA. In (1) the wild-type *At*WRKY11 protein and the glutamine (red) and the mutated lysine (blue) are shown. Several snapshots over the last 5 ns of the production run show that both residues form contacts with the DNA backbone in this simulation. In (2) the wild-type *At*WRKY50 protein and the lysine (blue) and the mutated glutamine (red) are shown. Several snapshots over the last 5 ns of the production run indicate that the glutamine forms contacts with a thymine and the lysine with the negatively charged DNA backbone.

observed the following: The wild-type lysine, which displays a local positive charge at its side chain, forms contacts preferably with the acidic DNA backbone atoms, whereas the mutated glutamine residue interacts with the thymine at position six of our 5'-AAAGTTGACCAA-3' sequence. This might indicate that lysine interacting with the DNA backbone confers sequence variability at this site, whereas glutamine at this position favors specific contacts with thymine. Unfortunately, we could not observe the opposite effect for *At*WRKY11. In this complex both, the glutamine and the mutated lysine form contacts with the DNA backbone. However, since the overall structure of *At*WRKY11 compared to *At*WRKY50 looks a little bit different and the position of the amino

acid is not completely identical it seems as if the DNA backbone is in the way and the side chains, especially the glutamine, cannot move towards the thymine during MD simulation. In binding affinity experiments our proposition was confirmed and it could be shown that mutated *At*WRKY50 binding became tighter and loses its ability to bind to the more degenerate 5'-TWGACY-3' consensus.

4.3.3.3 Structural details of *At*WRKY33 cDBD- and nDBD-DNA complexes

We simulated both *At*WRKY33 cDBD and *At*WRKY33 nDBD in complex with DNA. The overall structure of both proteins is similar although the sequence similarity is not as high as between *At*WRKY33 cDBD and other WRKY proteins. We could show that the binding interface of *At*WRKY33 cDBD and *At*WRKY33 nDBD is also highly similar (Fig. 4.15). This demonstrates that the N-terminal domain of WRKY proteins might interact with the DNA in a similar way as the C-terminal domain. However, the N-terminal domain is probably not important for specificity. We determined binding free energies for both *At*WRKY33 cDBD and nDBD to the 5'-AAAGTTGACCAA-3' DNA sequence. *At*WRKY33 cDBD has an even lower binding free energy than *At*WRKY1 cDBD.

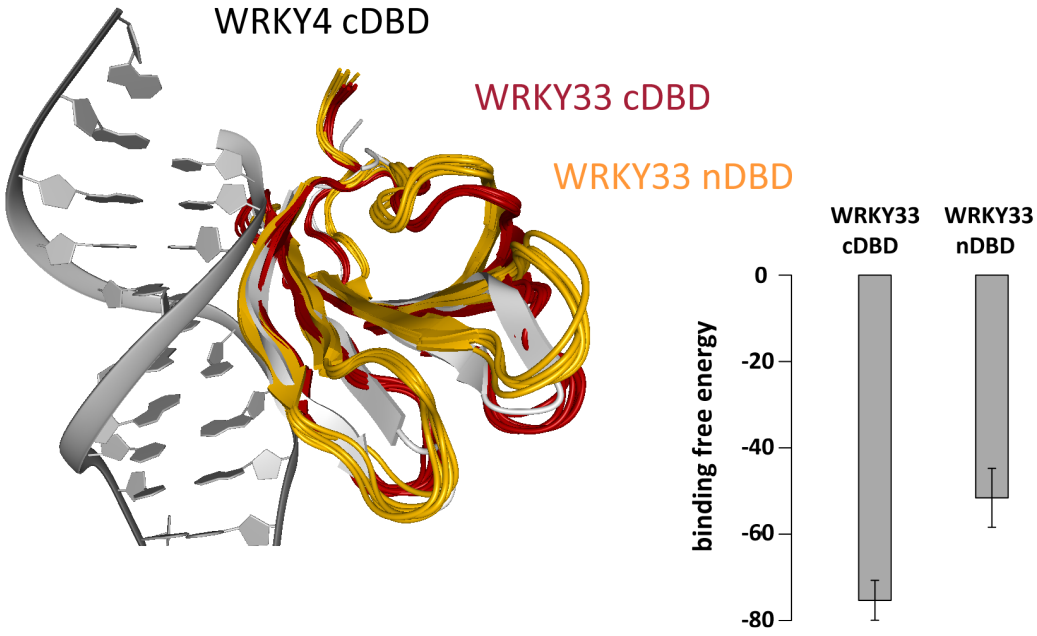


Figure 4.15: Binding of cDBD and nDBD to DNA. The *At*WRKY33 cDBD (red) and *At*WRKY33 nDBD (yellow) are superimposed with the NMR structure of *At*WRKY4 cDBD (gray) bound to the DNA. Only loop regions differ between all three proteins and the overall structure seems to be stable over the production run of the MD simulation since different snapshots of the *At*WRKY33 proteins were superimposed.

4.3.4 Discussion

Duan *et al.* [71] introduced the first model of how WRKY proteins might bind to DNA and which interactions are formed. They proposed amino acids important for binding and specificity at the binding interface. A better model was introduced by Yamasaki *et al.* [309] which could obtain a *At*WRKY4 cDBD-DNA complex by NMR. Since the complex was solved they could study interactions at atomic detail. Even though we modeled the WRKY-DNA structures without knowledge of this only recently solved complex structure we obtained a similar binding interface using molecular modeling approaches. By MD simulations and calculating binding free energies we could even identify the correct DNA binding sequence. Investigating in more detail the interactions in our modeled protein-DNA complex we could identify five specific contacts between amino acids and bases, which were also found by Yamasaki *et al.* [309]. As we are now able to simulate *At*WRKY1 cDBD protein-DNA interactions, it will be possible to gain a deeper insight into the binding mechanism of other WRKY proteins like *At*WRKY11, *At*WRKY50, *At*WRKY33 cDBD, and *At*WRKY33 nDBD.

In contrast to previous reports new experimental data provides evidence that there are WRKY proteins capable of binding a new aberrant W-box consensus, which might explain some diversity within the WRKY protein family. Binding specificity studies in the laboratory confirmed our assumption that the change of lysine to glutamine in *At*WRKY50 affects the binding specificity. We could show with molecular modeling techniques that the glutamine side chain favors the specific interaction with thymine in contrast to lysine which favors contacts with the DNA backbone. Studying the interface without inducing flexibility would not have been able to identify this specific interaction.

However, we could only predict reasonable binding free energies with MM-PBSA when more than one mutation was induced. Mutating single amino acids and estimating relative binding free energies had no clear effect since all other contributions to the binding affinity are too large and the standard deviation is mostly too high for distinctions between two similar values. With MM-PBSA approaches the overall biophysical contributions to binding free energies are sufficiently modeled, but small changes might not be accounted for. For single mutations at the binding interface methods like thermodynamic integration are probably better suited. However, we could introduce explicit water molecules and counter ions in our simulations which models the solvent more truthfully. These additional factors can further be evaluated and could give insight to water molecule distribution at the protein-DNA binding interface.

As mentioned previously, both WRKY and GCM1 feature a novel binding motif [53]. We also observed that most specific contacts with the DNA are formed between thymine bases and residue side chains as opposed to H-bonds, which is the common interaction type for TFs. Typically, TFs exhibit a distinctive H-bond pattern at the binding interface. Zif268 is an example for such a specific H-bond pattern when four H-bonds are established between a three bp long DNA sequence and an α-helix. However, WRKY proteins interact mostly with thymine, which might also affect the structure of DNA. Analysis of the GCM-DNA complex structure revealed a possible Hoogsteen base pairing. The WRKY binding motif includes also a potential Hoogsteen-dinucleotide: TpG at the second and third bp of the W-box sequence 5'-TTGACC-3'. Therefore, we postulate that not only the DNA

sequence motif, but also the local DNA shape is important for specific interactions between WRKY proteins and DNA. This was described for other TFs before [247].

Our data also implies the existence of different affinities of WRKY proteins to the W-box, which gives evidence for a competition-based model of WRKY proteins for W-box sites *in vivo*. This allows better understanding of how the integration of specific response by related WRKY transcription factors is achieved. Additionally, we could show with our model that the N-terminal domain has a similar structure compared to the C-terminal domain. However, it should be noted that the N-terminal structure was modeled using a C-terminal DBD as template. Using molecular dynamics we could demonstrate that both protein-DNA complexes are stable and the N-terminal structure is similar to the C-terminal structure after simulation.

5 Studying flexibility of protein-ligand complexes

[E]nzymes change shape upon interaction with their substrates. [...] The induced-fit hypothesis was amply confirmed by thousands of atomic-resolution structures of proteins and protein-substrate complexes [...]. More recently, modern nuclear magnetic resonance techniques have shown that structural fluctuations can occur even in the absence of interactions with substrates. [59]

Frederick W. Dahlquist, 2007
remembering Daniel E.
Koshland Jr., who introduced
the induced-fit effect

In the previous chapter we highlighted the relevance of modeling flexibility at protein-DNA complex interfaces. In the following we will examine flexibility of selected protein-ligand complexes. Proteins can undergo structural alterations upon binding to small molecules, therefore it is important to cover protein dynamics in drug discovery. In the first part of this chapter, we present an approach that can generate and identify possible protein conformations when only a single protein structure is available. The resulting three-dimensional protein structures represent the conformational space of a protein and can be used as input structures for docking. In the second part, we introduce a fast and reliable MD re-scoring protocol for docked ligand poses. With this approach we show that including flexibility in scoring techniques improves binding free energy predictions.

5.1 Introduction

The ultimate goal in rational drug design is to find a small, potent, and selective compound that binds to the target of interest. Computational techniques in combination with experimental approaches are at the heart of this drug discovery process. When the target protein binding site is identified and potential ligand molecules are available, structure-based drug design can be used to tackle the following questions computationally: What is the most suitable ligand with the best binding affinity? What does the structure of the protein-ligand complex look like? What is the energy of the interaction between the protein and ligand? Molecular docking is an approach which can address these problems. Docking methods place a ligand in the protein binding pocket and predict a binding affinity (ΔG) between the protein and the ligand. One of the main difficulties is to find the optimal trade-off between speed and accuracy. During drug discovery procedures fast docking methods are usually employed in a first instance. Thereby, large compound libraries are screened *in silico* to identify hit structures that potentially bind to the target. During this virtual high-throughput screening step small molecules are placed rapidly into binding pockets of target proteins. Subsequently, more advanced and accurate protein-ligand docking techniques are performed to eliminate non-binders and to detect the most promising active compounds. Compound structures with good binding characteristics then serve as first drug candidates (lead structures). These lead structures are further optimized and tested mostly *in vitro*, toward improving selectivity, binding affinity, bioavailability, and pharmacological properties.

We focus on two different structure-based modeling tasks in the two parts of this chapter. In the first part, we focus on finding binding mode(s) (and predicting the affinity) of ligands that are known to bind to a flexible protein. One task of a successful docking run is the generation of an accurate ligand conformation (pose), since determining a correct ligand pose is essential for predicting the true binding affinity. This step is highly dependent on the quality and accuracy of the protein structure in which the ligand is placed. Experimental methods, such as XRD and NMR, provide three-dimensional protein structures for docking. Recent advances in both methods make high-resolution structures available. However, it is difficult to capture long-range protein motions with these experimental methods. NMR techniques produce an ensemble of protein structures which represent, at least partially, the conformational space of flexible proteins. However, for some, especially large proteins it is difficult to determine NMR structures. Common XRD techniques disregard movements of proteins. Thus, only one single, high-quality image of a protein structure under a certain condition is provided. It is possible to obtain different protein conformations by applying the crystallization step several times using varying conditions. More advanced XRD techniques exist, e.g., time-resolved XRD, which enables to examine protein dynamics. Despite recent efforts, there are still limitations and obstacles to overcome when determining protein crystal structures, especially for flexible proteins. Typically, the more rigid and stable a protein structure, the easier and more successful is the crystallization process. Thus, docking methods which rely solely on experimentally determined protein structures are restricted to a limited number of available conformations. Ligands which bind to protein conformations, not experimentally obtained yet, are therefore most likely not detected by common molecular docking

methods. Most proteins can adopt their binding pocket upon binding of ligands. Most state-of-the-art docking methods [89, 136, 194, 207] therefore offer side chain flexibility within binding pockets to account for small induced-fit effects. More advanced computational methods exist [41, 51, 147, 152, 191, 223] which try to represent protein backbone movements, thereby trying to capture the whole conformational space of proteins. However, protein flexibility and especially large conformational alterations are difficult to include in molecular docking.

Protein-ligand docking methods typically do not only neglect large conformational changes of proteins, but also use approximate scoring functions in order to perform docking in reasonable time frames. Since all natural physiological conditions cannot be described completely by mathematical functions yet, one must draw on approximations in any case. Nonetheless, there exist advanced scoring functions that yield excellent results and place true active compounds at the top of resulting docking hit lists. In the second part of this chapter we discuss advances and drawbacks of state-of-the-art scoring functions. Additionally, we introduce an advanced MD simulation and binding affinity prediction protocol which accounts for flexibility while scoring ligands. We could demonstrate that re-scoring similar compounds while additionally accounting for flexibility discriminates between good and inferior binders and thus improves the rank-ordering of docking hit lists.

5.2 Representing major movements in protein-ligand docking

In this section we describe an approach consisting of three steps that produces a description of the conformational space of a protein. In a first step, we perform MD simulations to induce large protein backbone movements and overall flexibility to protein structures. In a second step, we conduct a principal component analysis (PCA) to reduce dimensionality. This was done in order to be able to cluster protein structures and identify distinct protein conformations more efficiently. These distinct protein conformations can then be used as input structures for molecular docking.

5.2.1 Introduction

As described in the general introduction to this chapter, molecular docking is a common technique to model protein-ligand complexes. State-of-the-art docking methods try to include flexibility of protein structures in order to be able to place ligands more accurately in protein binding pockets. When only one single rigid protein receptor conformation is considered, docking predicts an incorrect ligand binding pose for about 50-70% of all ligands [281]. Therefore, more advanced docking strategies were developed over the last years. Techniques which account for protein receptor flexibility vary from methods that provide local receptor flexibility in regions close to the binding site, to approaches that ensure even large protein backbone movements.

Methods which model local structural alterations are either so-called "soft docking" methods [134] or approaches which allow for flexibility of selected side-chains in the binding pocket [165]. "Soft docking" methods cope with steric clashes upon ligand binding by providing two different approaches. When force field-based scoring functions are used vdWs energies are commonly modeled by the Lennard-Jones

potential. The Lennard-Jones potential highly penalizes small distances between ligand and protein atoms. That implies that a true active compound which is too large for the rigid protein binding pocket is rejected. One solution in this case is to replace the Lennard-Jones potential by another term which tolerates small inter-atomic distances. It is also possible to reduce vdWs radii of protein atoms to enlarge the binding pocket artificially. Another approach to model local structural alterations is to account for flexibility of selected side chains in the protein binding pocket by using discrete rotamer libraries [7, 165, 310]. Rotatable bonds of binding site residues are first identified and then possible low-energy conformations of side chains, so-called rotamers, are applied. Especially, when docking larger ligands into protein binding pockets, that were refined with respect to smaller ligands during crystallization, these local structural alterations improve the placement step in molecular docking. The induced-fit effect can also be mimicked after docking by local energy minimization techniques that account for protein flexibility [40, 63, 194, 280].

Some proteins are very flexible and undergo domain movements that cannot be captured by flexible side chains or other local structural adjustments alone. Databases exist which accumulate possible protein conformations and known protein movements [84, 95]. Although, large backbone motions are computationally expensive to be realized by docking methods, some approaches exist that model global protein flexibility, because in certain case studies it is inevitable to include large structural alterations. One approach to model global flexibility is using multiple protein structures and docking the ligand into every structural conformer [5, 170]. These protein conformations can be obtained experimentally from XRD or NMR experiments [60, 170] or computationally from Monte Carlo or MD simulations [32, 37, 188, 230, 306]. FlexE [51] uses another possible method to represent flexibility. It uses a set of experimentally obtained protein conformations. FlexE creates an average structure using the regions of all proteins that comprise a similar structure and are considered rigid. Flexible parts of the protein are alternately exchanged while the average, rigid structure part remains fixed. Thus, also novel protein conformations are generated, which might represent a relevant conformation preferred by the ligand. However, these generated hybrid protein conformations might not exist in reality. Protein flexibility can also be represented when docking into an ensemble-average energy grid calculated from multiple protein conformations [152, 223]. However, these artificially constructed average structures might not be inhibited by true active ligands. Finally, low-frequency normal modes representing large protein motions can be identified, which can be included in docking algorithms [41, 147, 191]. During the docking procedure variables in the algorithm can account for the extent to which the protein is structurally altered along these normal modes.

We apply MD simulations to induce flexibility to protein structures. The difference to Monte Carlo simulations or normal mode analyses is that MD simulations produce structures that are generated continuously over a certain period of time. Therefore, we can additionally capture the movements of proteins and intermediate structures between two specific protein conformations. Large computational resources make it feasible to produce trajectories over several nanoseconds. Recent advances in MD simulation techniques enable even trajectories up to millisecond time scales [70]. Protein conformations stored in these trajectories represent the conformational space of a protein and distinct

protein structures can be identified, which may be able to interact with different ligands. We introduce a workflow that detects significant protein conformations and present results gained by applying this workflow to three proteins: aldose reductase (AR), dihydrofolate reductase (DHFR), and HIV-1 protease.

AR is an interesting drug target due to the fact that it is involved in the development of diabetes complications. AR catalyzes the reduction of glucose to sorbitol in cells. High glucose concentration leads to accumulation of sorbitol in cells, which results in osmotic damage and retino- and neuropathies. Different inhibitors for AR, as zenarestat [149], tolrestat [274], and IDD594 [125], have been developed. These inhibitors bind to different structural conformations of AR. All of them differ from the AR structure in complex with nicotin amideadenine dinucleotide phosphate ($NADP^+$), the so-called "holo-conformation" by Sotriffer *et al.* [269].

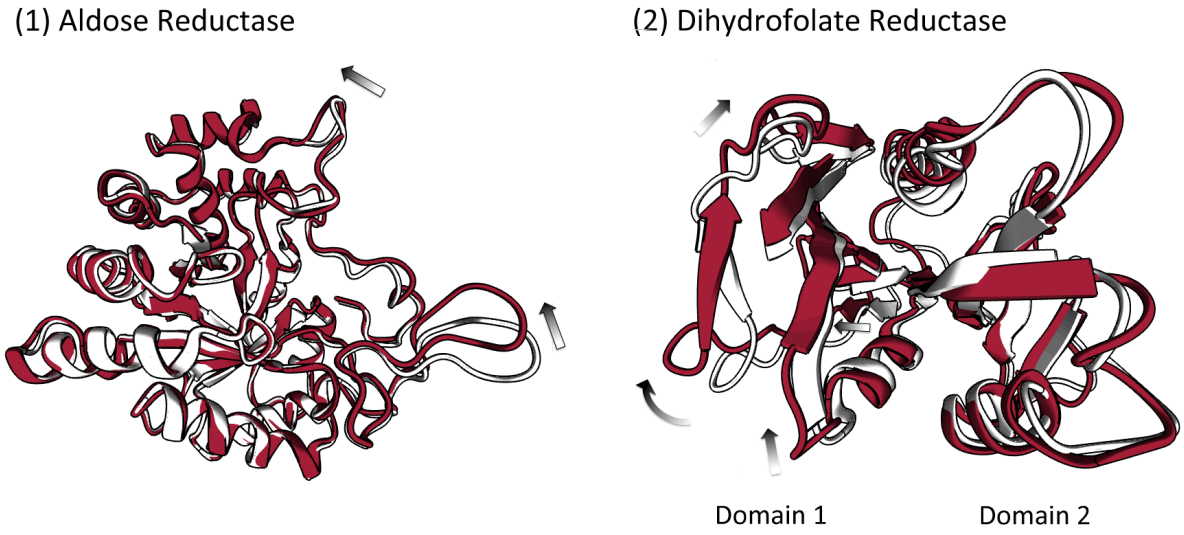


Figure 5.1: Flexibility of aldose reductase (AR) and dihydrofolate reductase (DHFR). (1) AR is predominantly flexible at two loop regions. Two protein crystal structures (PDB id: 2acr and 1iei) are superimposed and colored white and red, respectively. The C_α RMSD value of these two protein conformations is 0.8 Å. (2) DHFR undergoes shear motion with respect to its two domains. Two protein crystal structures (PDB id: 1pdb and 1yho) are superimposed and colored white and red, respectively. The C_α RMSD value between 1pdb and 1yho is 2.6 Å.

DHFR is an enzyme which regulates the amount of tetrahydrofolate in cells, which is essential for purine and thymidylate synthesis. Its central role in DNA precursor synthesis has made DHFR a popular target of anti-cancer chemotherapy, since its inhibition can limit the growth and proliferation of cells. It is known to undergo shear movement of its two domains and is therefore studied in this thesis.

HIV-1 protease is responsible for processing viral polypeptide precursors in HIV and thereby continues to be one of the primary targets of drug discovery efforts against AIDS. Consistent structural differences are present between the bound and the free states of the protein. When a ligand is bound to HIV protease it assumes a "closed" conformation, whereas in the unbound state it adopts a "semi-

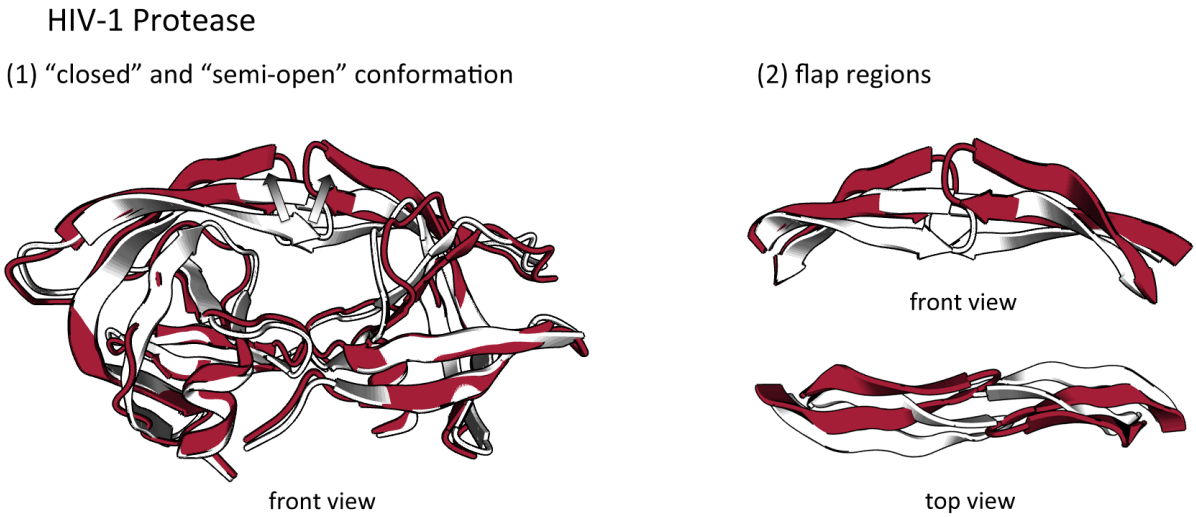


Figure 5.2: Flexibility of HIV-1 protease. (1) Two protein crystal structures in the "closed" (PDB id: 1hvr) and "semi-open" (PDB id: 1hhp) conformation are superimposed, and colored white and red, respectively. HIV-1 protease shows flexibility mainly in its flap regions. (2) Close-up of the front and top view of the flap regions. In the top view of the flap regions the change of the orientation of the flap tip regions is illustrated.

open" conformation [124]. Understanding the issues that govern HIV protease to open and close has implications for elucidating the detailed mechanism of this enzyme and the design of new therapeutic agents. These could be allosteric inhibitors, which interfere with the opening motion and thereby with the enzymatic function.

We introduce a workflow that comprises MD simulations, principal component analysis (PCA), and clustering in order to be able to cover the protein's conformational space as complete as possible starting from one protein crystal structure. We study different MD simulation protocols at different temperatures and up to 100 ns simulation time. Then PCA is applied to the resulting MD simulation trajectories in different coordinate spaces (C_α, backbone, dihedral, and rotamer space). This is accomplished to reduce the dimensionality of the data.

5.2.2 Materials and Methods

Proteins and protein-ligand complexes

We study three proteins in the following: AR, DHFR, and HIV-1 protease. For AR and DHFR a single protein structure was chosen in each case as input for MD simulations. Two protein crystal structures were selected for HIV-1 protease comprising two different protein conformations. We picked these crystal structures, because of their quality and to be able to compare our results to other studies [124, 269]. The overall protein quality was determined by analyzing the results of *Anolea* [195–197], *ProSA* [268, 303], *what-check* [121], and H^{++} [6, 101].

The protein crystal structure of human AR (PDB id: 2acr [105]) is refined at a resolution of 1.8 Å,

experimentally obtained as described in [304], and is in complex with cacodylate and $NADP^+$. The protein is a member of the triose phosphate isomerase (TIM) barrel class of proteins and contains eight β-strands. These strands form a barrel in the interior of the protein [105]. The cofactor $NADP^+$

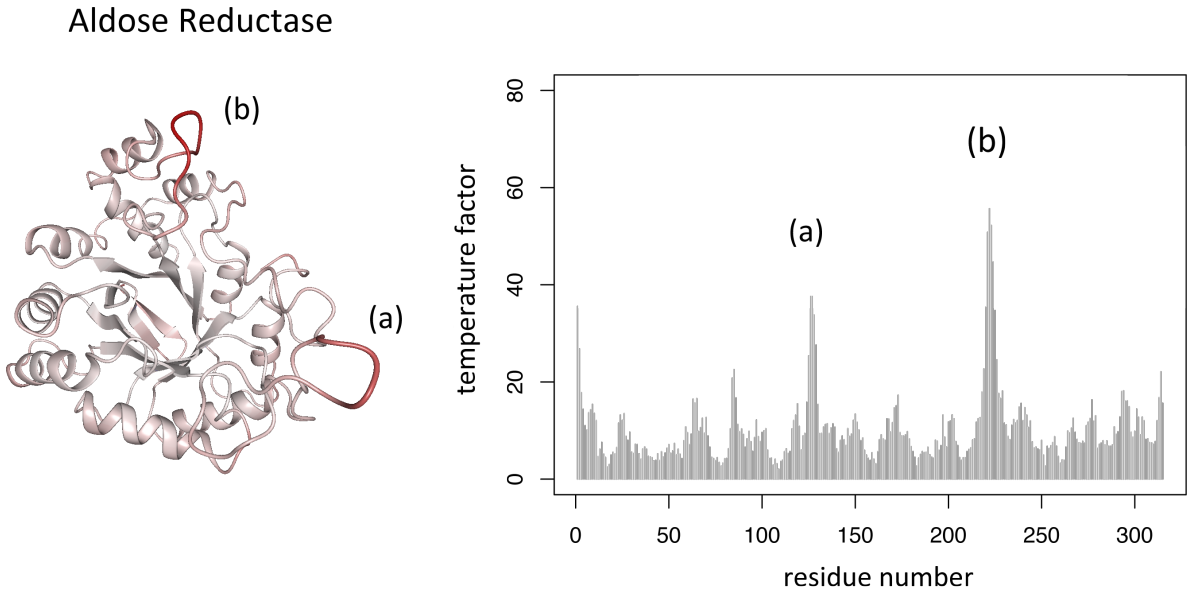


Figure 5.3: Temperature factors of aldose reductase (AR). The protein structure of AR (PDB id: 2acr) is colored with respect to temperature factors taken from the PDB structure file. For each residue of AR the temperature factor is illustrated and the two loop regions that are flexible can be clearly identified.

is bound to a long loop region, which comprises 20 residues (residues 210 to 230 in 2acr), see loop (b) in Fig. 5.3. This loop undergoes large conformational changes upon binding of $NADP^+$ [105]. Cacodylate binds to the active site of AR, but shows no inhibitory effects and at a certain pH value only weak activation effects. Inhibitors like citrate and glucose 6-phosphate bind to a similar binding site, while the binding pocket undergoes structural alterations. The structure of AR in 2acr exhibits a "holo-conformation", since it is similar to 1ads [304], the structure described as "holo-conformation" by Sotriffer *et al.* [269]. Before performing MD simulations, we deleted all co-factors as well as all water molecules. The 2acr structure was checked for missing residues and the following residues were renamed to fulfill *AMBER* naming conventions and protonation states determined by H^{++}: 41 to HID, 46, 83, 110, 187, 240, and 306 to HIE, and 163 to HIP. Aside from the 2acr PDB structure, which serves as input for MD simulations, we extracted 47 protein structures from the PDB listed in Table C.1 in the appendix. All of them were mapped onto the 2acr PDB structure using RMSD-minimizing superposition (as implemented by the atom bijection method in *BALL* [114]). If necessary we also mutated protein residues with *BALL* to feature the same protein sequence as for 2acr. We defined the following residues for mapping as input in the C_α atom pair list accounting for the lowest temperature factor values (see Fig. 5.3) in 2acr: 15-19, 40-44, 73-78, 105-109, 155-160, 180-185, 205-209, and 258-262. Temperature factors indicate the relative vibrational motion of different parts

of proteins. Atoms with low values belong to a part of the structure which is well ordered. Atoms with large temperature factors generally belong to a part that is flexible. In Fig. 5.3 high temperature factors are present in the two loop regions of AR.

The protein crystal structure (PDB id: 1pdb [52]) of human DHFR is a refined three-dimensional structure at resolution of 2.2 Å and represents an apo structure. DHFR features a shear interface with

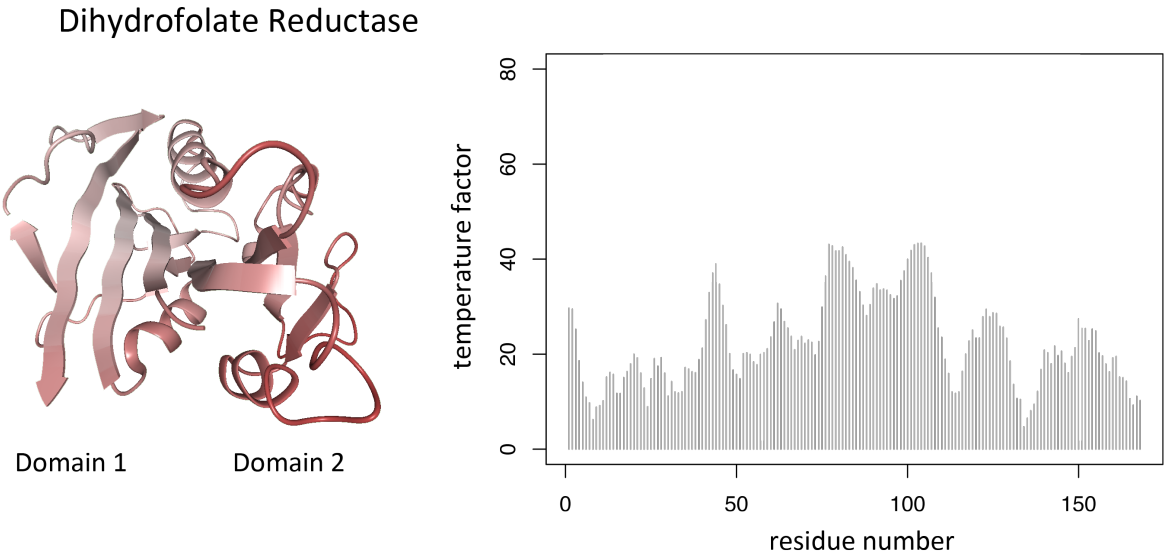


Figure 5.4: Temperature factors of dihydrofolate reductase (DHFR). The protein structure of DHFR (PDB id: 1pdb) is colored with respect to temperature factors available from the PDB structure file. The highest temperature factors are present at loop regions.

hinges and can obtain different conformations upon ligand binding with respect to its two domains. The first domain consists of the following residues in 1pdb: 1-39 and 116-186. The second domain includes residues 40-116. We deleted all water molecules and renamed the following residues according to *AMBER* naming conventions to represent protonation states determined by H^{++}: 87 and 130 to HIE and 127 to HIP. Apart from the 1pdb protein structure of DHFR, which we used as input structure for MD simulations, we obtained 33 PDB structures similar to 1pdb (listed in Table C.2 in the appendix). We mapped these structures onto 1pdb and mutated amino acids according to the residues in 1pdb. C_α atoms of the following residues were used as input for the atom pair list when superposing the structures by RMSD-minimization: 71-75, 88-90, and 93-98. Since 1pdb is refined at lower resolution as 2acr, temperature factors (see Fig. 5.4) give no clear indication whether or not the structure is very flexible in particular parts. We can only observe that the highest temperature factors are present in all loop regions.

HIV-1 protease is a homodimer formed by two subunits of 99 amino acids each. Both of them feature a large and highly mobile region, the so-called flap region (residues 43-58 and 142-157 in 1hvr and 1hhp). The protein crystal structure of HIV-1 protease in the "closed" conformation (PDB id: 1hvr [161]) is refined at 1.8 Å resolution and is in complex with a non-peptide cyclic urea inhibitor.

Both domains, represented by chain A and B, are present in this structure and the flaps are pulled in toward the active site (residues: Asp 25, Thr 26, Gly 27). The two flap regions have the highest temperature factors and are colored red in Fig. 5.5. The apo structure features the "semi-open"

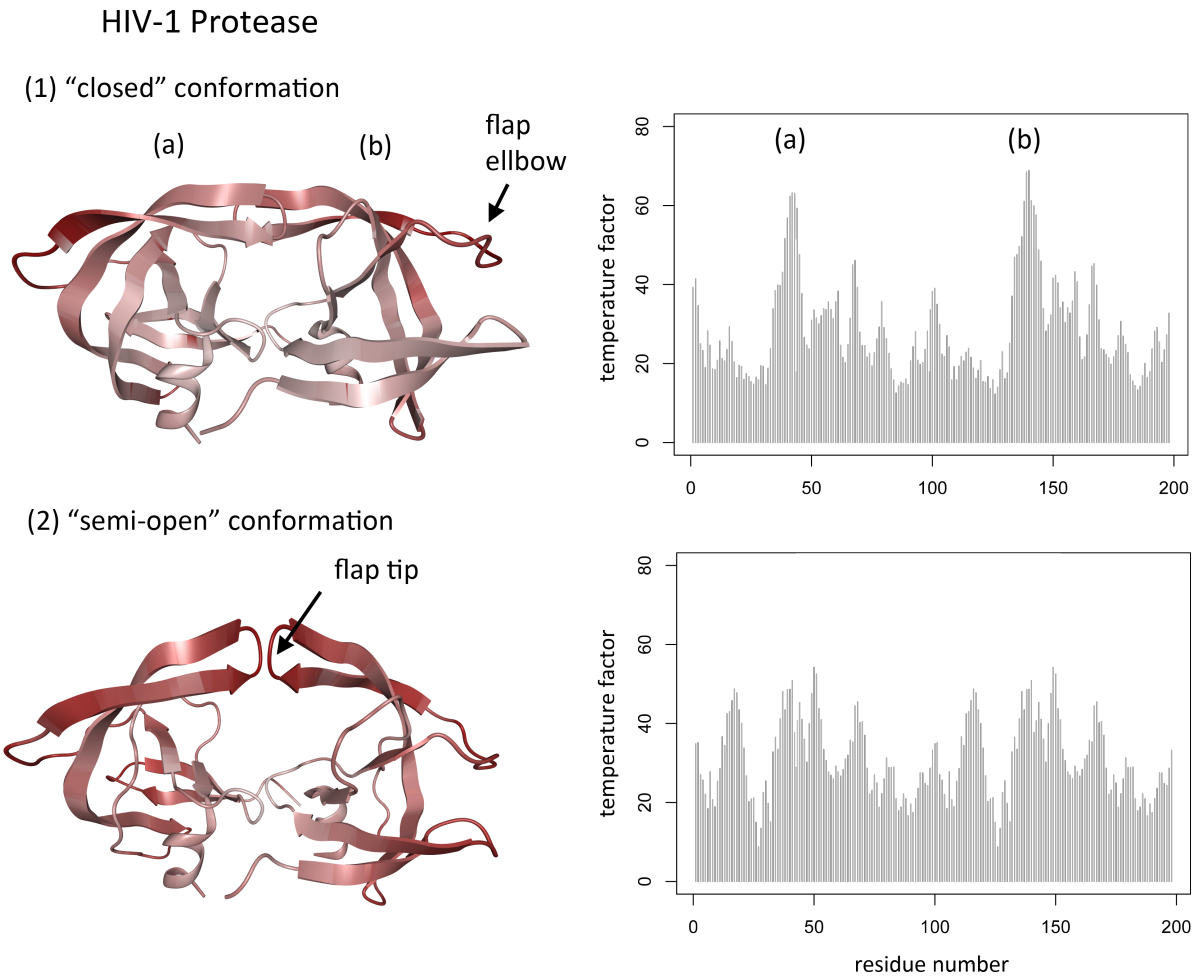


Figure 5.5: Temperature factors of HIV-1 protease. The protein structures of HIV-1 protease (1) "closed" conformation (PDB id: 1hvr) and (2) "semi-open" conformation (PDB id: 1hhp) are colored with respect to temperature factors taken from PDB files. In both conformations the flap regions, especially the flap elbow regions, show highest temperature factors, however, they are more clearly identified in the "closed" conformation. In the "semi-open" conformation the flap tips present higher temperature factors as in the "closed" conformation.

conformation (PDB id: 1hhp [270]) and is refined at 2.7 Å resolution. A second domain (chain B) is not present in this structure and needs to be obtained by applying the transformation matrix on the protein atoms of chain A. The temperature factors of the "semi-open" conformation (see Fig. 5.5) indicate the flap regions, as well as other loop regions as flexible. We deleted the inhibitor present in the 1hvr structure, changed the CSO residues to CYS residues, and renamed histidine 69 to HIP in both domains of 1hvr and 1hhp according to *AMBER* naming conventions to represent protonation states determined by H^{++}. Apart from the two protein structures of HIV-1 protease we obtained

seven further protein structures from the PDB which possess both chains, listed in Table C.3 in the appendix. If necessary we mutated amino acids of these PDB structures to feature the same sequence as in 1hvr. All proteins are then mapped onto the 1hvr PDB structure by RMSD-minimizing superposition (bijection method) using C_α atoms of the following residues: 9-15, 18-25, 108-114, and 117-124. These residues share the lowest temperature factors in the "closed" HIV-1 protease structure.

Protein structure generation with molecular dynamics

LEaP processes the input PDB files to *AMBER* topology and coordinate files. We used the ff99SBildn force field [171] and corresponding parameters for the MD simulations in *AMBER 11* [39]. The following numbers correspond to each of the protein structures in this order: AR, DHFR and both HIV-1 protease proteins (2acr/1pdb/1hvr/1hhp). 0/1/9/9 Cl^- ions were added to neutralize the system and 59/42/51/50 additional Na^+ and 59/42/51/50 Cl^- ions were added to obtain a salt concentration of 0.2 mol/l, while the protein was placed in an octahedral water box of TIP3P water molecules with at least 15 Å distance to the boundaries of the box.

The MD simulation protocols are similar for all proteins, see Fig. 5.6 for an illustration (details are listed in Tables C.4 and C.5 in the appendix). The system is minimized in four steps (step A-D) releasing gradually more and more atoms from spatial constraints. Then the system is heated up from 100 to 300, 315, 350, or 400 K during a constant volume simulation (NVT-MD) and relaxed at this temperature to an equilibrium state in step E. This results in four different MD simulation setups for each temperature. The seed for the random number generator was set to 209,858. While heating up the system all protein atoms were constrained. These constraints were gradually released in terms of the strength of the force constant and the number of atoms in five subsequent steps (steps F-J). In all of these 100 ps long constant pressure simulation (NPT-MD) steps the center-of-mass-motion is removed every 1,000 steps [42] to avoid energy drains [49, 107]. When the final temperature is reached the system is kept at this temperature using a Langevin thermostat with a collision frequency of 2.0 ps^{-1}. The particle mesh Ewald method (PME) [61] is used to treat long range electrostatic interactions. SHAKE [250] is applied to constrain bond lengths involving bonds to hydrogen atoms, therefore a time step of 2 fs is sufficient. Subsequently, the production run (step K) is executed in which the system is simulated over 100 ns.

Principal component analysis

PCA is applied to reduce the number of dimensions. Internally, the data obtained by the production run of each MD simulation is represented by an MxN matrix, where M is the number of dimensions (e.g., Cartesian coordinates or dihedrals) and N the number of snapshots stored in column vectors. A snapshot, which represents a protein conformation at a certain point in time, was extracted every 25 ps from the production run trajectories. This results in 5,000 snapshots for the 100 ns long production runs. The length of the column vectors and in this case also the dimensionality, depends on the space the PCA is performed in. The protein structures stored in the trajectories are preprocessed to perform

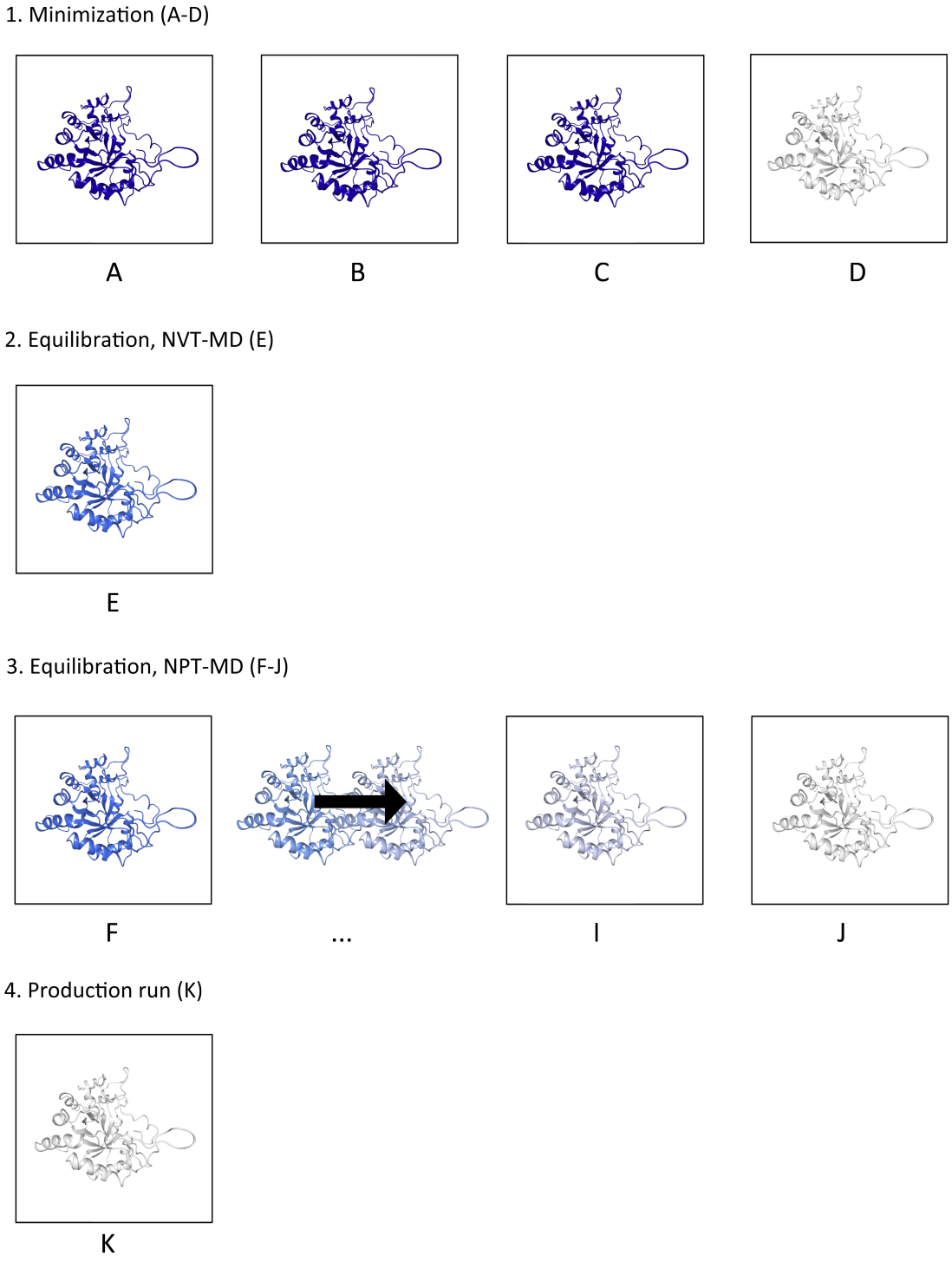


Figure 5.6: MD simulation procotol. Initially, the system is minimized in four steps (A-D). Then it is equilibrated at constant volume (NVT-MD), shown in step E. Steps F to J indicate the equilibration at constant pressure (NPT-MD). Finally, the production run is performed in step K. The constrained atoms are indicated in blue, whereby the lighter the color blue the lesser constraint are the atoms. However, since no H-atoms are represented in detail this is only an approximate illustration.

meaningful PCA analyses. Movements of both N- and C-termini are high and would add noise to interior structural protein dynamics. We truncate these terminal ends to be able to study relevant structural changes. Five residues were deleted from the N-termini of DHFR and of each domain of both HIV-1 protease structures and ten residues were deleted from the N-terminal of AR. Five residues were deleted from the C-termini of AR, DHFR, and of each domain of both HIV-1 protease structures. This results in 300, 176, and 178 residues representing AR, DHFR, and HIV-1 protease, respectively. The PCA is performed in the following structural spaces: C_α, backbone, dihedral, and rotamer. When PCA is applied in C_α space, the coordinates (x-, y-, and z-positions) of all C_α atom of each snapshot are used. This results for instance for AR in 900 dimensions. In backbone space, all backbone atom coordinates of all residues are considered for calculation, which results in 3,600 dimensions for AR. Dihedral and rotamer space are represented by internal coordinates instead of Cartesian coordinates. The dihedral space is constructed by taking Φ and Ψ angles of all residues into account. This results in a dimensionality of 1,200. Rotamer space includes all angles, which means all χ angles of the side chains and all Φ and Ψ angles of the backbone. For AR this results in 3,600 dimensions.

We use a singular value decomposition method for the PCA, which is provided by methods implemented in *BALL*. With this method we try to identify the atoms or angles of proteins which contribute most to structural changes. Thus, we reduce the dimensionality of the data, in this case, the structural space with respect to significant atoms and angles. The output of the PCA is a matrix containing the principal components of every snapshot. Each row corresponds to one principal component. The first principal component is given by the first row of the matrix, the second principal component by the second row of the matrix and so on. The first principal component points into the direction of the largest variance of the dataset, the second into the direction of the second largest variance perpendicular to the first principal component, and so on. The matrix of all principal components represents the data in the new coordinate system, which is spanned by eigenvectors of the covariance matrix. Each eigenvector corresponds to an eigenvalue, which represents the fraction of variance being explained when projecting the data on the corresponding eigenvector.

Clustering and cluster representatives

Different data clustering algorithms (k-means [192], PAM [141], fuzzy C-means [24], and spectral clustering [212], provided by R) were thoroughly studied in [201]. In this study [201], details about the results of different clustering algorithms are discussed. Since we could not observe large differences and k-means performs well with respect to quality and run-time we employ k-means clustering in the following.

In principle, the goal of k-means is to partition a data set X with n observations into k groups (clusters) by assigning each observation to a group with the nearest mean. Let $X = \{x_1, ..., x_n\}$ be the data set of n observations. Each observation has to be assigned to one of the k clusters with the centers $C = \{c_1, ..., c_k\}$ that the within-cluster sum of squares $\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$ (where μ_i represents the cluster center of the cluster C_i) is minimized. The implementation of the k-means

algorithm in *R* is included in the package "stats". In this *R* package four different k-means algorithms are included: Hartigan and Wong [106], Lloyd [175], Forgy [87], and McQueen [192]. All algorithms were tested, but no major differences could be detected for our problem. Therefore, we chose to present in this work results produced by the default algorithm in R, which is the k-means approach of Hartigan and Wong [106].

The input for the clustering is an MxN matrix, where M denotes the number of principal components representing 95% of the variance in the data. And N, the length of each principal component, is equal to 5,000. We chose 50 for the number of clusters, since we determined 50 as the best value for the number of clusters with respect to performance and run-time in [201], and performed 100 runs with random start configurations. In addition to the 50 k-means cluster centers, we obtained two more data points of each cluster that have the largest distance to one another within this cluster, yielding 150 cluster representatives.

Since we performed 100 independent cluster runs we obtain different results and therefore have to choose the best result. We determined the best result by calculating intersections of volumes between convex hulls. Convex hulls were determined with the 3D polyhedral surface method as implemented

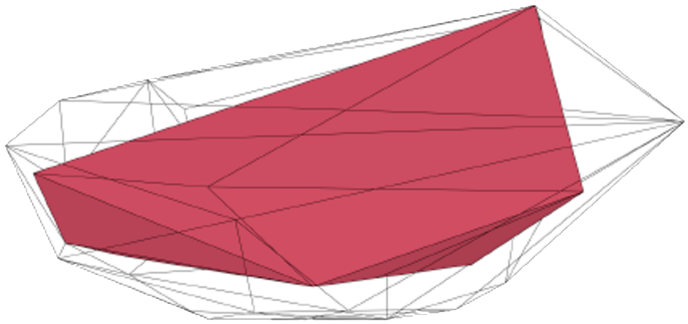


Figure 5.7: Convex hulls of MD simulation data and cluster representatives. As illustration for the intersection of volumes between two convex hulls the convex hull of MD simulation data is shown as wire frame and the volume of the convex hull of cluster representatives in red. This figure is taken from [201].

in CGAL [112]. First, a convex hull was calculated using MD simulation data: 5,000 snapshots of the trajectory, represented by their first three principal components. Then, intersections were calculated between this convex hull and each convex hull spanned by 150 cluster representatives described by their first three principal components. The clustering result with the highest volume was chosen, since it covers the largest fraction of the original space.

To obtain a smaller subset of the 150 cluster centers we performed a hierarchical clustering in R. The input matrix for this clustering is an RMSD distance matrix. C_α RMSD values are calculated using the distance method applied on corresponding C_α atoms of the previously superimposed structures between all 150 cluster representatives. The hierarchical tree was cut to obtain 5, 10, 15, ..., or 150

protein structures to be able to determine the minimal value necessary to represent the structural space spanned by PDB structures. The structures with minimal C_α RMSD values to all other structures in the cluster were chosen as new representatives. This was performed for all 16 options for each protein: molecular dynamics at 300, 315, 350, or 400 K in combination with PCA in C_α, backbone, dihedral, or rotamer space.

Docking

We performed a docking study using *Glide* (version v58025) [89] and docked four ligands into different AR structures. Six AR conformations were chosen (2acr, 1ads, 1el3, 1iei, 2pdl, and 2pdk) and were first prepared with the Preparation Wizard (version v40025) of the Schrödinger Molecular Modeling Platform. The ligands present in these structures (IDD384 (1el3), zenarestat (1iei), tolrestat (2pdl), and sorbinil (2pdk)) were processed with LigPrep (version v40025). All protein structures have been mapped onto 2acr as described previously. Therefore, we could determine the common center of all ligand atoms (x = 17.754 Å, y = 24.569 Å, and z = 61.086 Å). This point was used as center for creating the grid box. Each side of the three-dimensional grid box is 30 Å long. *Glide* SP was used for docking using standard parameters.

The same docking procedure was conducted using the 16 small subsets of cluster representatives of AR as input conformations. All four ligands were docked into all cluster representatives.

5.2.3 Results

We show that with our method we can cover the structural space of available protein crystal structures starting from a single protein conformation. Thus, we can generate and identify one structure that is highly similar to each of the known PDB structures. We tested our method with AR, DHFR, and HIV-1 protease structures. In order to examine the validity of our protein conformations as input structures for docking methods, we docked small ligand molecules into our AR structures.

5.2.3.1 Conformational diversity

Some proteins can undergo major structural changes, which often correspond to their function. However, it is not possible to determine all three-dimensional protein conformations for these flexible proteins experimentally. Only a single structure might be available from which different conformations can then be generated and identified with our method. We used MD to induce flexibility to one protein structure and PCA and clustering to identify the most distinct conformations. We tested our method with AR, DHFR, and HIV-1 protease which are all important drug targets. We chose these targets, since there are already different relevant conformations available in the PDB. Our goal is to discover all relevant structures starting from a single structure. In Fig. 5.8 (1) the conformational diversity of the respective PDB structures is illustrated. We chose one crystal structure for each protein (AR: 2acr, DHFR: 1pdb, HIV-1 protease "closed" form: 1hvr, and HIV-1 protease "semi-open" form: 1hhp) and compared its structure to all available protein conformations of the same protein in

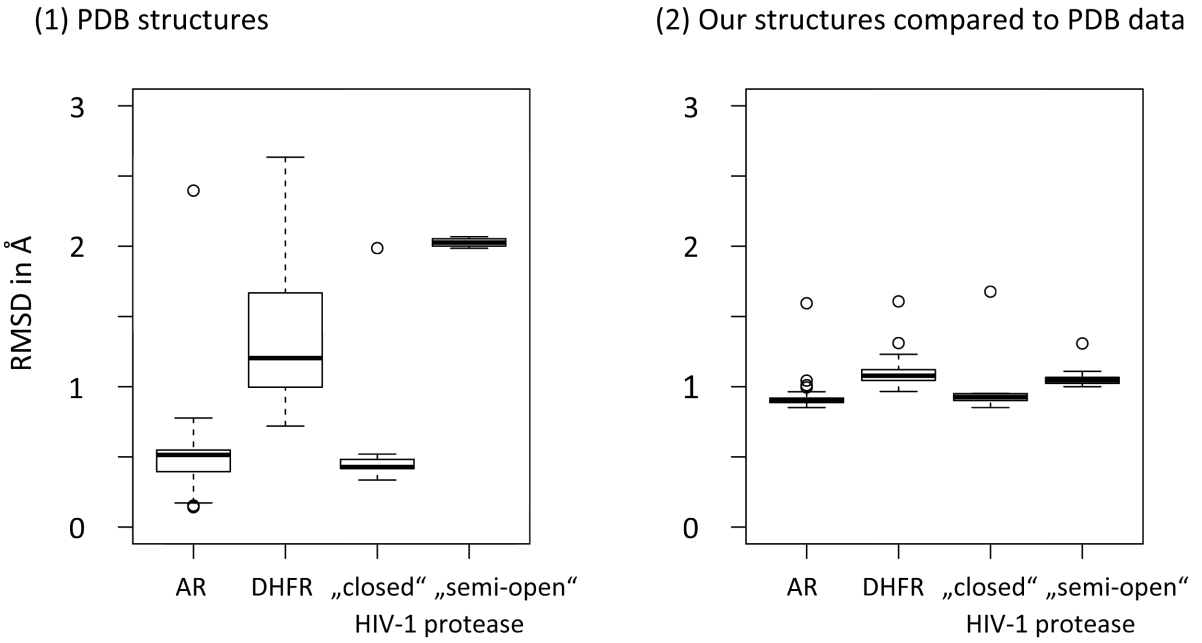


Figure 5.8: Structural diversity of PDB structures and optimal generated protein conformations. In (1) we compare all available PDB structures of each protein (AR, DHFR, HIV-1 "closed" and "semi-open" form) to one protein PDB structure of the same protein (2acr, 1pdb, 1hvr, and 1hhp). These four PDB structures serve as starting structures for our method which is used to generate different conformations. In (2) the lowest RMSDs between each available PDB structure and its most similar conformation sampled by our method are presented.

the PDB. The higher the RMSD, the higher is the structural difference between the protein conformations. It is evident from Fig. 5.8 (1) that there exists structural diversity in the PDB structures. At least one PDB crystal structure is different to each of our selected PDB structures by 2 Å. In Fig. 5.8 (2) the overall result of our study is presented. We show that we can sample all relevant protein conformations present in the PDB within an RMSD of 1.7 Å starting from one protein crystal structure. For most conformations the RMSD is even below 1.5 Å. For AR and HIV-1 protease ("closed" and "semi-open" form) the RMSD clusters around 1 Å for all generated structures except one.

We generated different AR conformations by applying our method to the 2acr structure. Most C_α RMSD values calculated between each relevant PDB structure and the most similar structure sampled by our method are below 1 Å, except three outliers. However, even the protein conformation 1xgd that comprises a completely different second loop conformation, see Fig.5.1 (1), (C_α RMSD between 2acr and 1xgd: 2.4 Å) is almost obtained with our approach. The structure we identified to be most similar to 1xgd has an RMSD of 1.7 Å which is the highest we observe for AR.

DHFR comprises a well-defined shear motion. Two structural conformations of this protein, 1pdb and 1yho, are illustrated in Fig. 5.1 (2), whereby their C_α RMSD difference is 2.6 Å. In Fig. 5.8 (2) we illustrate the C_α RMSDs between each DHFR PDB structure and the most similar protein conformation we generated. The structure we identified to be most similar to 1yho has an RMSD value

of 1.6 Å. Compared to all other PDB structures 1xgd comprises the most dissimilar conformation. All other RMSD values cluster around 1.1 Å.

In the PDB two different conformations of HIV-1 protease are available as crystal structures. Both conformations, the "closed" and "semi-open" form, are illustrated in Fig. 5.2 which have an RMSD of 2.0 Å. In Fig. 5.8 (1) RMSD values are calculated between one representative of the "closed" form (PDB id: 1hvr) and all HIV-1 protease crystal structures available in the PDB. We have only one "semi-open" conformation in our data set, which results in an outlier at 2.0 Å. When calculating RMSD values between this conformation and all other structures in the data set, the RMSD values cluster all around 2 Å. When starting from the "closed" HIV-1 protease structure we could not determine a perfect "semi-open" conformation. The RMSD between the "semi-open" conformation and our best generated structure starting from the "closed" conformation is 1.7 Å. However, when starting from the "semi-open" HIV-1 protease conformation we can generate and identify structures that are highly resembling to the "closed" conformation (RMSD of 1.0 Å).

Overall we can generate and identify for each of the available PDB structures one structural conformation that is highly similar (RMSD < 1.5 Å), except for three structures. Even these three conformations have an RMSD value below 1.7 Å. For obtaining these structures we used four different MD simulations protocols conducting the production run at four temperatures (300, 315, 350, and 400 K). Additionally, we tested different PCA spaces to reduce the dimensionality. Which combination of the two methods performs best is analyzed in the following. In Fig. 5.9 generated structures with

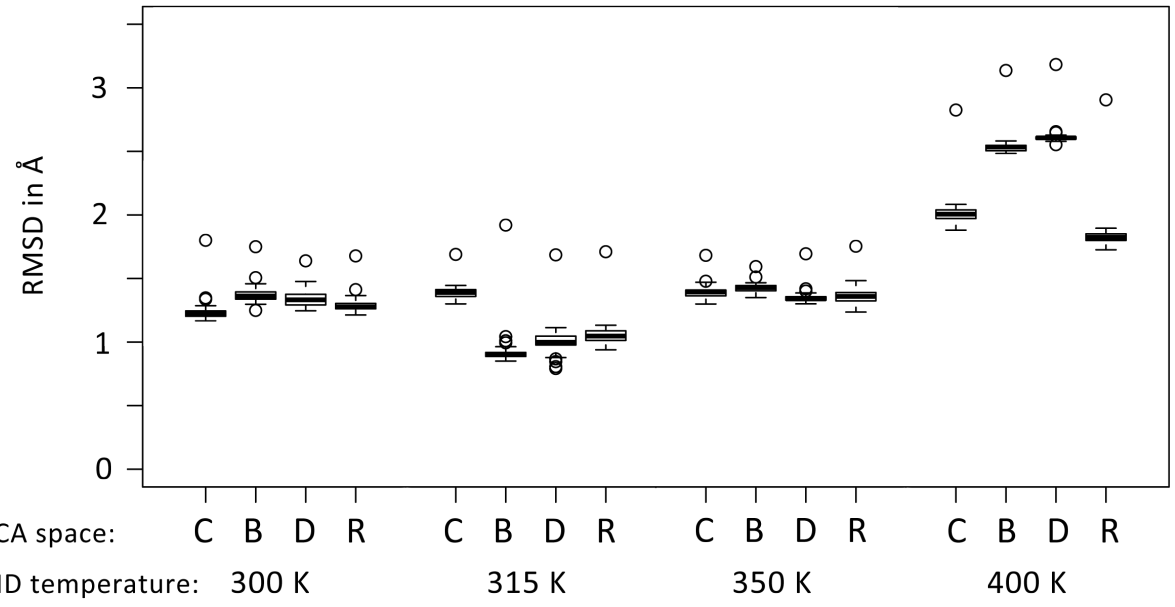


Figure 5.9: Structural diversity of AR for different MD simulations and PCA. Different AR protein conformations were generated with MD at 300, 315, 350, and 400 K and identified by clustering on PCA results performed in C_α (C), backbone (B), backbone dihedral (D), and rotamer (R) space.

minimial RMSD values to each of the known PDB structures are illustrated for AR for each PCA space (C_α, backbone, backbone dihedral, and rotamer space) and each temperature. It seems as if the MD simulation at 315 K generates the most similar structures to the PDB structure of AR in combination with PCA in backbone, backbone dihedral, or rotamer space. However, we observe one outlier, since we could not perfectly generate the structure of 1xgd. In case of 1xgd, simulating 2acr at 350 K and performing PCA in backbone space yields the most similar structure for this AR conformation. Simulating 2acr at 400 K generates structures that are not close to any relevant PDB structure conformations. For DHFR and both HIV-1 protease conformations these results differ, see Fig. C.1, Fig. C.2, and Fig. C.3 in the appendix. The best "closed" HIV-1 protease conformation we identified starting from the "semi-open" conformation was generated when simulating 1hhp at 400 K. In contrast to AR, HIV-1 protease needs to undergo large backbone motion to generate this conformation. This might only be possible at higher temperatures. Smaller structural alterations can already be achieved at 300 K, but larger protein dynamics like shear movements in DHFR or opening and closing of HIV-1 protease flap regions require higher temperature to be successfully sampled.

For AR we show how much of the conformational space generated by each MD simulation is covered by our identified 150 cluster representatives emerging by clustering in different PCA spaces. We performed 100 k-means cluster runs and determined the best clustering result by calculating the intersection between volumes as described in the materials and methods section. The clustering result which yields the maximal coverage is indicated as a bullet in Fig. 5.10. The best coverage is obtained

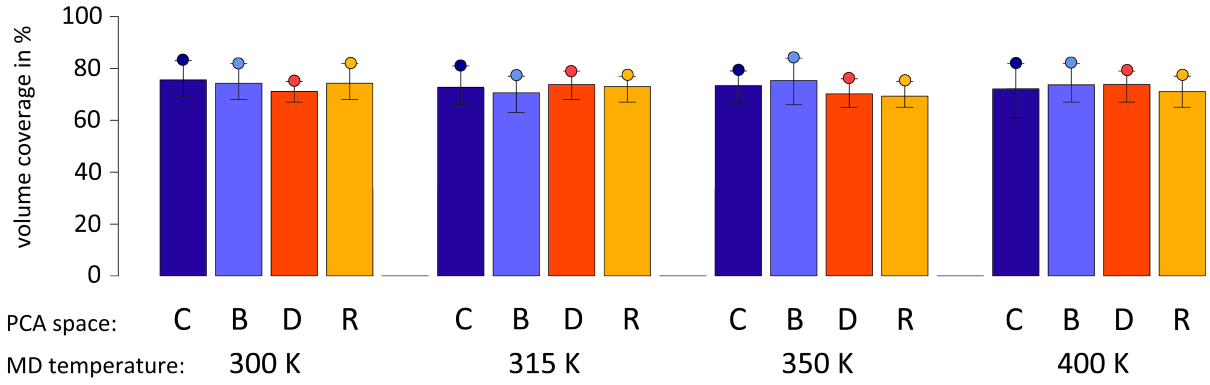


Figure 5.10: Conformational space of AR covered by cluster representatives. The volume coverage represents the space covered by 150 cluster representatives of the space spanned by all 5,000 protein conformations obtained with MD simulations. The best clustering result for each MD simulation trajectory at 300, 315, 350, and 400 K and PCA performed in C_α (C), backbone (B), dihedral backbone (D), and rotamer (R) space is indicated as a bullet point.

when using C_α space regardless of the temperature during the MD simulation. The clustering performs comparably well in all PCA spaces, with backbone dihedral covering the least volume, but still covering 75% of the volume. How well 150 cluster representatives cover the conformational space generated by MD simulations of DHFR and HIV-1 protease is presented in Fig. C.4, Fig. C.5, and Fig. C.6 in the appendix.

In all previous results we used 20 identified structures selected from our 150 cluster representatives. This number was determined by comparing the minimal RMSD values calculated between available PDB structures and our generated structures. The threshold to describe a structure as similar was set to 1.5 Å. In Fig. 5.11 the percentage of PDB structures which are covered by our identified protein conformations with an RMSD value below 1.5 Å is illustrated for different numbers of sampled AR structures simulated at 300 K. When we take 20 structures 96% of all PDB crystal structures are highly similar to our conformations regardless of the space the PCA was performed in. When PCA is performed in backbone or rotamer space we can observe that more structures are necessary to represent all PDB protein conformations.

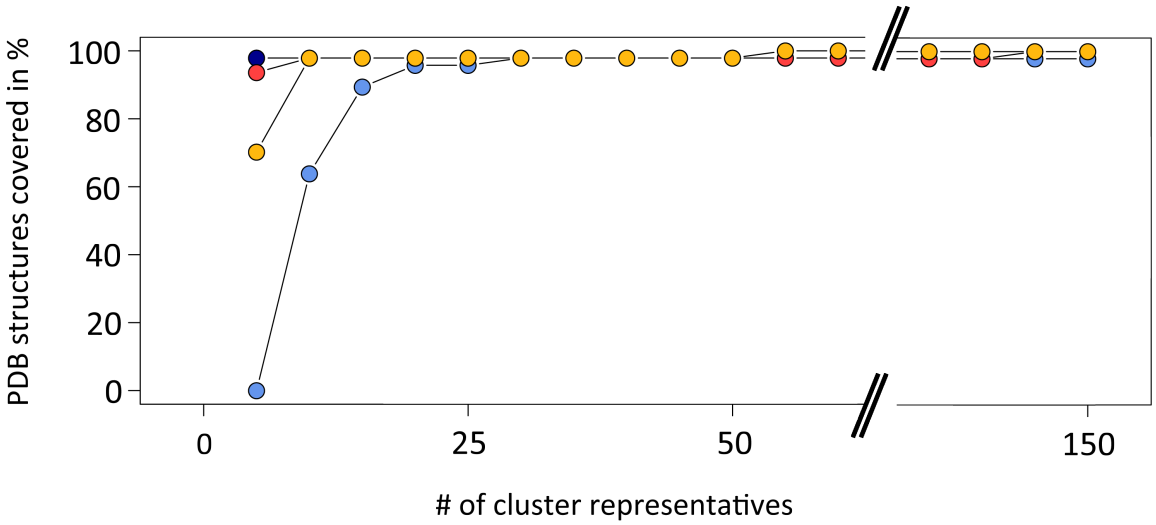


Figure 5.11: Optimal number of generated AR structures. C_α and backbone space are colored in dark and light blue and dihedral backbone and rotamer space are colored in red and yellow, respectively. The MD simulation was performed at 300 K. Different numbers of representatives for determining the best hierarchical clustering results were tested. A PDB structure is defined as similar to a generated protein conformation when the RMSD between the two structures is below 1.5 Å.

Details on inducing flexibility using molecular dynamics and identifying protein conformations that are similar to relevant PDB structures are described in Appendix C. Our approach is described in chronological order exemplary for AR. Results of PCA and clustering are illustrated as well as 20 AR conformations sampled with molecular dynamics at 300 K and PCA in C_α and backbone dihedral space.

5.2.3.2 Recurrent conformational changes

In Fig. 5.12 we illustrate the structural differences for existing PDB crystal structures of HIV-1 protease. HIV-1 protease is known for dynamics predominantly in its flap regions [124] that are mainly responsible for diverging C_α RMSD values. An inhibitor is bound to the active site in the "closed" HIV-1 protease structure. Interestingly, we can generate HIV-1 protease conformations similar to the "closed" structure when starting from the "semi-open" conformation. This suggests that the bound inhibitor might not be responsible for the "closed" conformation. The most similar structure we identified, when starting from the "semi-open" conformation, has a C_α RMSD value of 1.0 Å compared to the "closed" HIV-1 protease PDB structure, see Fig. 5.12. We can observe the opening of the flap

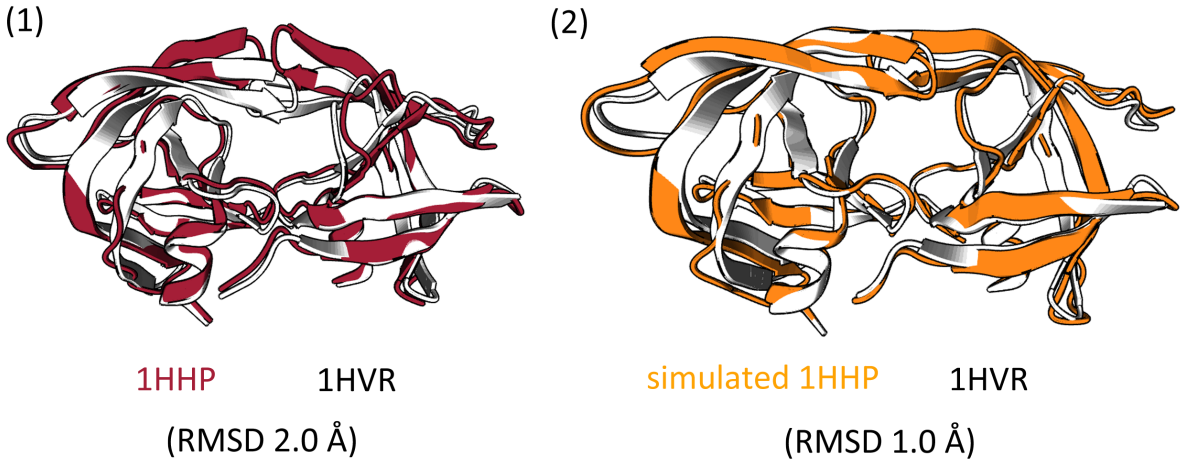


Figure 5.12: HIV-1 protease conformations. (1) The most different PDB structure conformations representing the "closed" (PDB id: 1hvr, white) and "semi-open" (PDB id: 1hhp, red) conformation of HIV-1 protease. (2) Our most similar identified "closed" conformation (yellow) generated starting from the "semi-open" HIV-1 protease structure.

regions as described by Hornak *et al.* [124], see Fig. 5.13. The trend of flap tip dynamics occurring on a much more rapid timescale than full opening could also be reproduced by our explicit solvent simulation. However, we could not create the "fully-open" form of HIV-1 protease. As consistent with recent explicit solvent simulations we show that the "closed" form is present and can be generated from the "semi-open" conformation. This implies that no inhibitor needs to be bound to the active site of HIV-1 protease for the "closed" conformation to occur. Other studies described an irreversible opening of the flaps, which could not be observed here. However, these studies were conducted in implicit solvent simulations which might have an effect on highly flexible regions of proteins. In Fig. 5.13 we illustrate the opening and closing events of HIV-1 protease flap regions. This fact can be observed by molecular dynamics, since this method generates structural conformations are continuously.

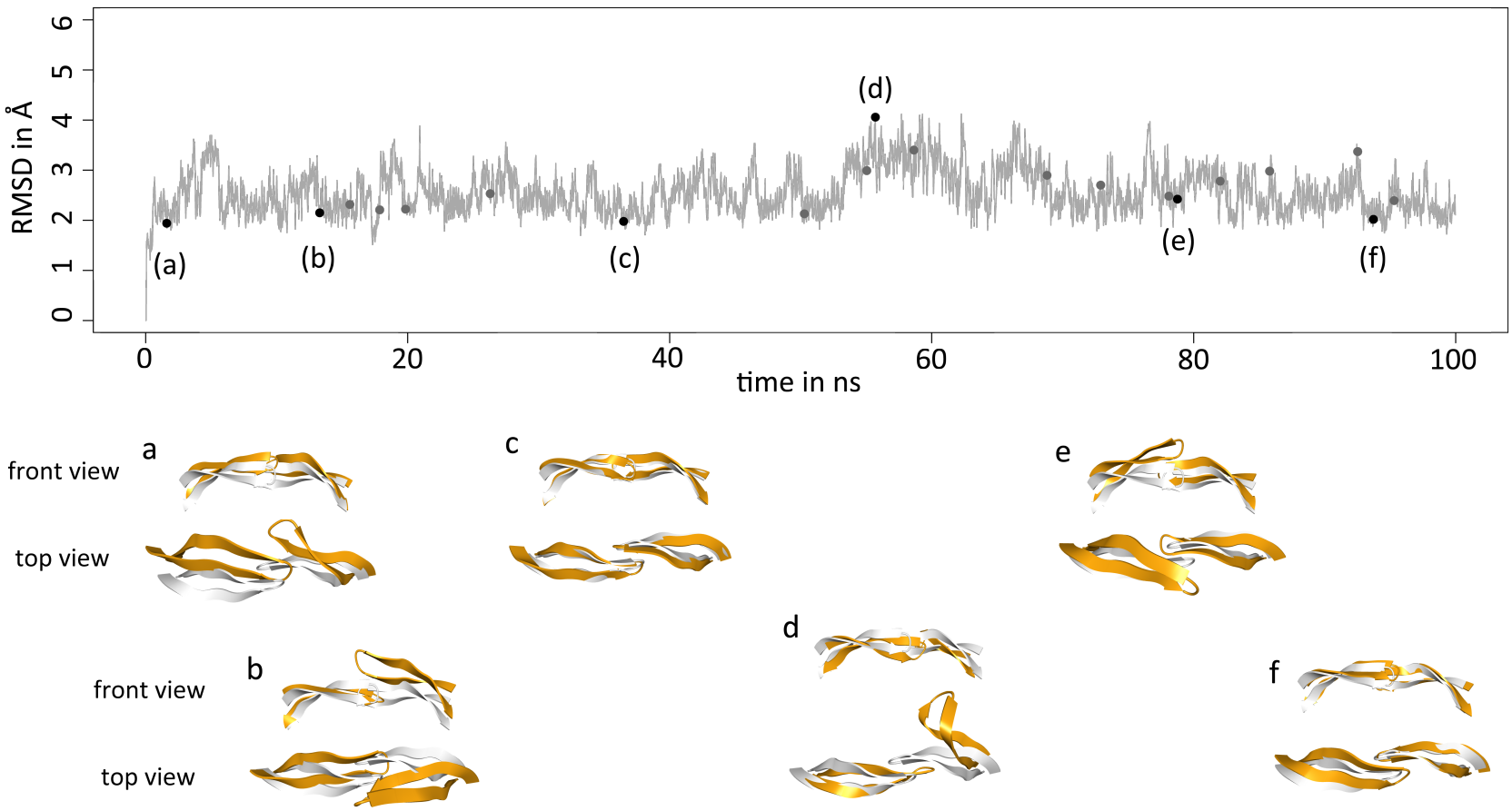


Figure 5.13: Opening and closing events of HIV-1 protease flap regions during MD simulation at 400K. The "semi-open" HIV-1 protease is simulated at 400 K. RMSD values between the starting structure of HIV-1 protease (PDB id: 1hhp) and conformations generated during the production run at 400 K are illustrated. The dots indicate 20 identified structures by PCA in dihedral rotamer space, whereby six are illustrated. In yellow the flap regions of 1hhp at different time points are shown superimposed with the "closed" HIV-1 protease conformation in white (PDB id: 1hvr).

5.2.3.3 Docking into structural representatives

Structural dynamics of proteins can be studied with MD simulations that generate conformations that describe the structural space. As stated by Hornak [122, 124] and Sotriffer [269] knowledge about conformational changes of proteins influence drug discovery processes and are important to consider, for instance in molecular docking. We could demonstrate that our approach can identify structurally diverse protein conformations. Especially, backbone movements of AR, DHFR, and HIV-1 protease can be clearly identified. The question arises if our approach is also applicable to identifying smaller differences, e.g., side chain alterations in ligand binding pockets. Exemplary, we study differences in binding pockets of AR using five PDB protein structures (2acr, 1iei, 1el3, 2pdk, and 2pdl). C_α RMSD values between 2acr and the PDB structures are comparably low, although they reveal different side

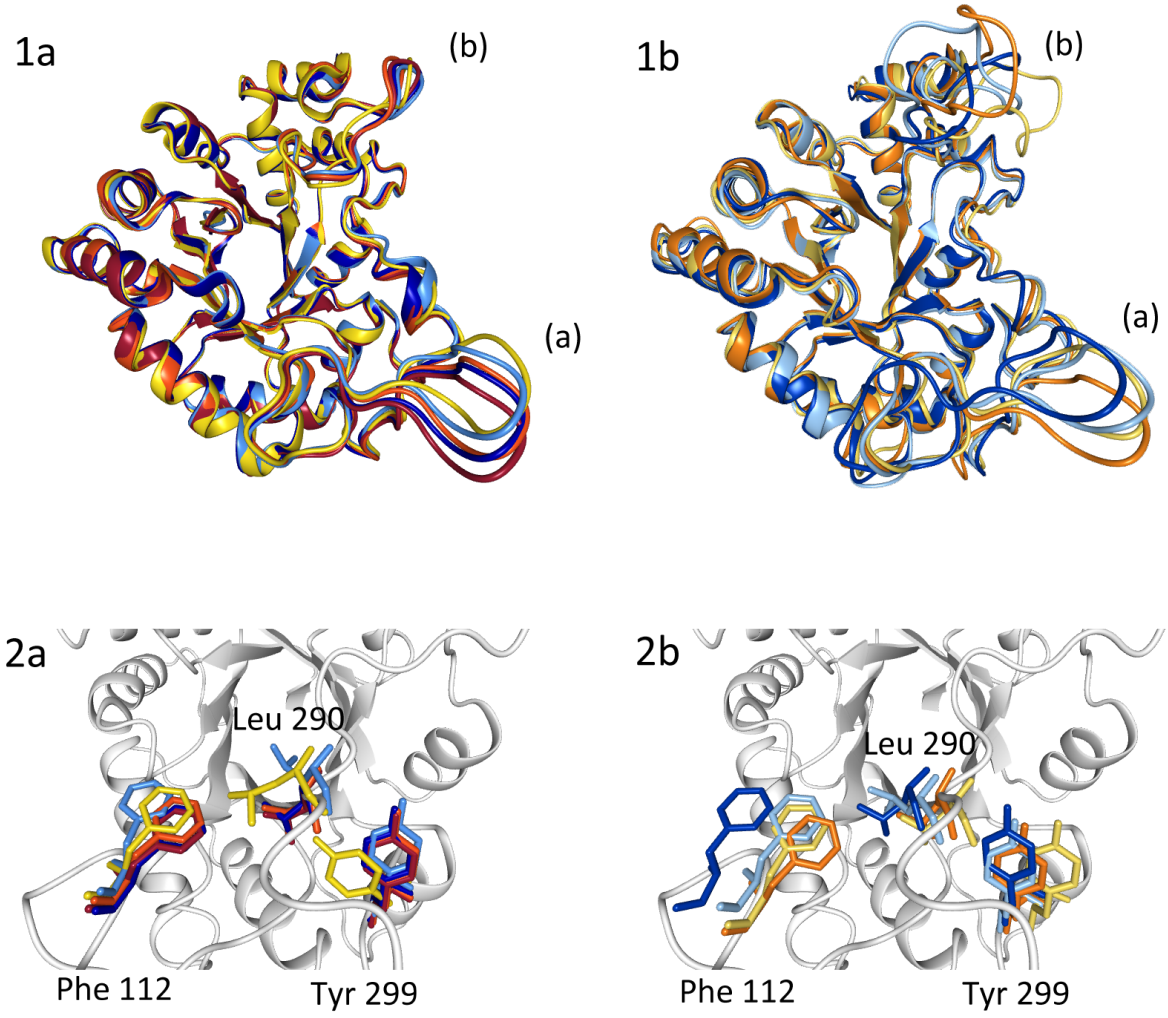


Figure 5.14: Structural differences of AR. In (1a) five PDB structures: 2acr (red), 1iei (yellow), 1el3 (orange), 2pdk (dark blue), and 2pdl (light blue) are superimposed. (1b) shows the overlay of structures (A-D) that were depicted by docking: A (dark blue), B (light blue), C (yellow), and D (orange). In (2a) PDB structures and (2b) our structures, three flexible active site side chains are illustrated and colored with respect to their corresponding PDB structures.

chain orientations in their active sites (see Fig. 5.14 2a). Loop (a) comprises different conformations, but other large backbone dynamics are not visible in PDB structures (Fig. 5.14 1a). RMSD values reveal slightly larger structural variations between 2acr and 1iei and between 2acr and 2pdl, see Table 5.1. RMSD values raise to 1.3 Å when considering only active site atoms (residues: 38, 69, 100, 101, 112, 208, 209, 288, 289, 290, and 299) and including also all side chain atoms to the calculation. These small local structural changes can also be observed in Fig. 5.14 2a. Side chains described as being flexible (Phe 112, Leu 290, and Tyr 299) [269] can clearly be identified when superimposing different PDB structures. These three side chains are probably responsible for increased binding pocket RMSD values. RMSD values between our identified structures and 2acr are overall higher. As for the PDB structure, the structure zenarestat bound to with lowest affinity (structure A) yields an RMSD value of 1.9 Å compared to 1.5 and 1.4 Å for the other structures. Vast binding pocket variations between 2acr and structure A and D can be observed through respective RMSD values. The structures with lowest binding affinities (Emodel scores) for each ligand (A-D) are illustrated in

Table 5.1: AR input docking structures. RMSD values between 2acr and four PDB structures comprising different conformations are calculated for all C_α, binding pocket (BP) C_α atoms, and all atoms in the binding pocket (BP). The ligands present in the PDB files are used for docking. RMSD values for corresponding structures depicted by docking (ligand with lowest binding affinity).

PDB id	ligand	C_α	C_α BP	all atom BP	structures	C_α	C_α BP	all atom BP
1iei	zenarestat	0.8	0.9	1.3	**A**	1.9	4.0	3.9
1el3	IDD384	0.3	0.2	0.6	**B**	1.5	2.5	2.7
2pdk	sorbinil	0.4	0.3	0.7	**C**	1.4	2.2	2.6
2pdl	tolrestat	0.6	0.9	1.3	**D**	1.5	4.5	5.4

Fig. 5.14 1b and 2b. Both loop regions are highly flexible and different conformations can be observed for our structures. Compared to PDB structures our AR conformations show significant differences in loop regions a and b. Loop (b) tremendously changed its structure in A and D causing high RMSD values, since residues of this loop are shifted up and to the right in Fig. 5.14 1b. These large backbone motions alter the position of the side chains present in the ligand binding pocket (see Fig. 5.14) 2b. Structure A, in which sorbinil docked in with lowest affinity has Phe 112 shifted more to the left compared to the other structures. In structure C (bound with tolrestat) Leu 290 is oriented into the active site region, which results in displacement of the ligand during docking. In Fig. 5.15 (B) the original ligand is placed differently than the docked ligand. Whereas for sorbinil, which is the smallest ligand we tested, both ligands overlay nicely (A). Tyr 299 might be important for π-stacking with the aromatic ring of zenarestat. Since the tyrosine is shifted to the right in our structure zenarestat is placed differently. IDD384 is also not perfectly docked, see Fig. 5.15 D. Interestingly, our docked ligand is located deeper inside the binding pocked compared to the experimentally determined ligand position. Docking results reveal for all four ligands (sorbinil (A), tolrestat (B), zenarestat (C), and

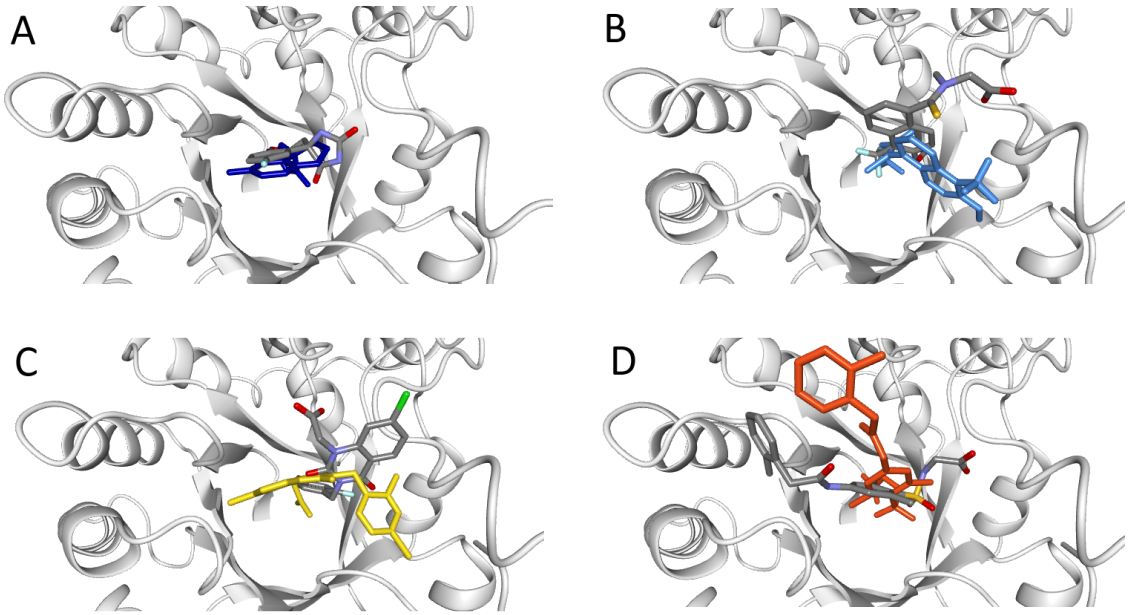


Figure 5.15: Ligand docking poses. PDB structures of docked ligands (colored according to element symbols) are superimposed with original PDB structures. The ligands in the PDB structures are colored corresponding to the PDB structure coloring. (A) sorbinil (dark blue), (B) tolrestat (light blue), (C) zenarestat (yellow), and (D) IDD384 (orange).

IDD384 (D)) that in each case the ligand was not perfectly docked into the binding pocket. Even small changes in side chain conformations cause problems during docking. The backbone dynamics of AR during MD simulation are probably too high and, therefore overly affect the structure of the ligand binding pocket. Generally, we observe that structures taken from the beginning of the MD simulation trajectory, with smaller overall backbone changes, are preferred by all ligands during docking. These small side chain alterations of AR can probably be obtained with much shorter simulations. We identified AR conformations where loop (a) was folded completely over the active site, thereby blocking the binding pocket. These conformations are not suitable for docking, however give insight of large backbone dynamics of AR.

5.2.4 Discussion

Advances in MD simulation software development yield more and more accurate results for representing flexibility of biological macromolecules. The progress of acquiring more refined parameter sets allow longer simulations and at the same time reduce the risk of creating artifacts. The study performed by Hornak *et al.* [124] used an implicit solvent approach. They discussed that using this artificially low solvent viscosity permitted observations of large structural changes on a computationally tractable timescale. Nowadays, similar simulations can easily be conducted in explicit solvent [252], which then can also provide useful insights into the role of water during dynamics. We performed our simulations in explicit solvent and could demonstrate that they perform well for all tested protein

structures. Other approaches could also be used to generate diverse protein conformations. These conformations could then be used as input for PCA analyses. With MD simulations we could not only create diverse protein structures, but could also observe that the opening in HIV-1 protease flap regions is reversible. Compared to other studies [233, 252] our simulations of HIV-1 protease generates less dynamic structures. Reasons for this might be different force fields, simulation protocols, solvent models, and maybe even structural variations of the starting structure. Despite the fact that we could not observe the "fully-open" structure of HIV-1 protease, we demonstrate that our approach yields structural conformations similar to all PDB structures for the respective proteins.

In this study PCA was used to reduce the dimensionality of the data set in order to be able to perform clustering. Other, more advanced methods than linear dimensionality reduction techniques, such as PCA exist which might yield preferably results. PCA was first introduced in the analysis of MD simulation trajectories by Ichiye and Karplus [128] and Garcia [93]. In some cases, results gained by linear PCA provide inadequate preservation of local structures [54] or separation of stable conformational states [62, 271]. Therefore, non-linear PCA approaches have been developed to overcome limitations of linear PCA, one promising example are diffusion maps [54, 79]. To gain somewhat reasonable PCA results we truncated proteins at their N- and C-termini after simulation. The terminal regions often show high flexibility during MD simulations. Especially, when these regions are not folded properly in distinct secondary structural elements they unfold and add noise to the data. Since PCA detects the highest variation in the data, most probably, these terminal dynamics would be described. Despite its drawbacks, in our cases PCA detects residues which have the greatest influence on structural changes and is therefore sufficient for this purpose. We could show that residues present in the first principal component comprising the highest influences in PCA are congruent with temperature factors extracted from original PDB files. We performed PCA also in dihedral space and showed for C_α and backbone dihedral space that we obtain similar protein conformations.

Clustering was performed on 95% of the original data points after reducing the dimensionality of the data set with PCA. Different clustering algorithms result in varying performances. We tested four clustering algorithms: k-means, spectral, fuzzy, and PAM. Results of this study are described in [201]. Although k-means has limitations and tends to produce clusters with similar size, we chose this method. This might lead to results as shown for the backbone dihedral clustering, where data points present in smaller clusters are merged into one cluster. Thus, distinct conformations present in high density regions are then not covered by k-means clustering and discarded for further evaluations. Therefore, algorithms such as centroid-linkage, centripedal, and Bayeseian algorithms which are able to create clusters with distinct shapes and sizes might be superior [263].

We docked four ligands into each of the 20 AR structures and chose the best protein-ligand complex with respect to the ligands' Emodel value provided by *Glide*. It is obvious that the docking fails for all ligands, except sorbinil and we were not able to gain perfect ligand binding poses. Structural alterations in the active site of our sampled AR structures are too large. Shorter MD simulations would be beneficial in this case, since small changes of side chains could already be generated with shorter production runs. When analyzing a particular protein binding pocket molecular dynamics should be

performed on protein-ligand complex structures. This would lead to more specific information of ligand binding modes and interactions between the protein and ligand. We also did not account for side chain flexibility during docking. Side chain flexibility during docking might resolve some steric clashes and might improve docking results. Nevertheless, we observed side chain movements in the binding pocket similar to the ones detected in PDB structures. It has to be noted that docking results presented here are obtained using AR conformations identified by C_α PCA. Therefore C_α atom and backbone dynamics are probably better represented in these structures compared to side chain alterations. One solution would be to perform PCA on certain atoms of the ligand binding pocket in order to identify smaller, specific, and local structural changes.

We can conclude that our method is better suited for sampling large backbone dynamics in protein structures than identifying small side chain alterations in protein binding pockets.

5.3 Rapid molecular dynamics simulation protocol for re-scoring docked ligand poses

This section describes an advanced protocol for re-scoring docked ligand poses using an implicit solvent MD simulation and MM-PB/GBSA. We include flexibility while re-scoring ligands, which have been placed in a protein binding pocket with the docking program *Glide*. While developing an implicit solvent MD simulation protocol and determining a parameter set for estimating binding free energies between proteins and ligands we attached great importance to improve speed as well as accuracy.

5.3.1 Introduction

Coupled with the rapidly growing number of experimentally solved protein structures, protein-ligand docking has become prominent in drug discovery procedures. Among the various structure-based computational methods adopted for identifying active compounds, the principal one is molecular docking [272]. When the structure of the target protein is available, protein-ligand docking can be used for screening large chemical compound libraries [219]. Scoring functions in docking programs try to model biophysical interactions in order to estimate binding affinities of protein-ligand complexes as accurately as possible. Despite huge efforts in developing improved scoring functions [27, 46, 74, 236, 240], a scoring function that can universally predict binding free energies between proteins and ligands remains an elusive goal [217, 292]. The underlying physico-chemical conditions at binding interfaces are still not completely understood. Given a particular scoring function good results can be obtained for a certain target protein, but addressing the whole range of different protein structures is not yet possible. Even if it would be able to capture all features present in nature, they could not be represented by present techniques due to storage and run-time limitations.

Notwithstanding important successes, docking continues to struggle with many methodological deficiencies. Since many compounds need to be screened in a timely fashion the accuracy in predicting their relative binding free energies is quite poor. A common conclusion when comparing different scoring functions is that they are accurate enough to yield substantial statistical enrichment over random or diverse compound selection, but are not suited to rank-order structurally related compounds [279]. Thus, computational screening results still suffer from false positives and false negatives and are not sufficiently accurate to rank compounds according to experimentally obtained binding affinities [102]. This implies that compounds which are ranked at the top after high-throughput *in silico* screening might not be true binders and some active inhibitors are possibly placed too low on resulting rank lists to be selected for further experiments within drug discovery procedures.

Approximations employed in such docking procedures are reduced protein flexibility, inadequate treatment of solvation, ambiguity in protonation states, and neglection of most entropic terms [102]. In addition to these approximations the overall simplistic nature of energy functions contributes to the inability of scoring functions to discriminate between compounds of similar chemical structure that differ by several log units in potency [160, 182]. Furthermore, the docking score is typically calculated using a single conformation even though natural binding free energies are ensemble properties.

An stepwise approach using scoring schemes with rough estimates in the beginning and more sophisticated functions in the end has been shown to be efficient [33, 34]. In the first steps fast, approximate, and general scoring functions are applied to a large number of ligands. Then final scoring functions usually attempt to include solvation and entropic effects while scoring a smaller amount of compounds. These are additionally more target specific. In the following, we introduce a rapid MM-PB/GBSA re-scoring protocol which estimates binding affinities while also including flexibility. It can discriminate weak and strong binders with high accuracy, especially when applied to similar ligands. Therefore, this re-scoring protocol can serve as one of the final scoring schemes.

5.3.2 Materials and Methods

Protein-ligand complexes

We studied a set of 18 ligands for the target protein urokinase. The protein urokinase is a serine protease that is related to several malignant diseases. Brown and Muchmore [34] used the same dataset for developing a high-throughput version of the MM-PBSA method, however they used protein-ligand complexes as in-house crystal structures. After extensive evaluation of all available PDB structures using *Anolea* [195–197], *ProSA* [268, 303], and *what-check* [121], we chose the urokinase protein structure (PDB id: 1owh [302]). The three-dimensional structures of the ligands were created with Maestro (version v90211) using the naphthalene-derived ring system of the co-crystallized ligand (PDB id: 1owh [302]) as scaffold. All 18 ligands have this scaffold in common consisting of a system of two conjugated six-membered rings.

Figure 5.16: Urokinase inhibitors. All 18 urokinase inhibitors are superimposed and show their initial position for docking. The naphthalene-derived ring system is marked with a box. This illustrates the similarity of all 18 urokinase inhibitors that differ only in their decorations.

Structural preprocessing and docking

The input protein-ligand complexes for the re-scoring protocol are generated using the docking program *Glide* (version v58025) [89, 90, 104]. Each ligand is docked into the urokinase protein structure (PDB id: 1owh). This structure was prepared using the Protein Preparation Wizard (version v40025) of the Schrödinger Molecular Modeling Platform. We assigned protonation states of the ligands manually based on $\mathrm{pK_a}$ values of the ionizable groups. The ligand present in the 1owh structure was used as reference to define a core constraint for the naphthalene-derived ring system. This constraint was used to gain ligand poses during docking which are similar to the ligand position in 1owh. A ligand pose was rejected when the ring system had an RMSD to the reference structure greater than 1.0 Å. The XP scoring function was used during docking with *Glide* (version v58025) [89, 90, 104]. The best determined ligand pose together with the protein structure was saved for each ligand.

Structural preprocessing for molecular dynamics simulation

The protein structure (PDB id: 1owh) was protonated according to the H^{++} [6, 101] output.All water molecules and other small molecules were deleted and residues were renamed to agree with *AMBER* naming conventions. The cysteine residues: 34, 55, 147, 180, 196, 217, 207, and 235 form disulfide bonds and were therefore renamed to CYX. Histidine residues are renamed to either HIE: 22, 96, 107, 177, 184, 249 or HIP: 54, 106, 257. The modified protein PDB file was then used as input for *LEaP*.

Antechamber [291] was used to assign partial charges and atom types to the ligand in order to prepare it for the MD simulation. The force field parameters for the small molecules are taken from the GAFF force field [290] and the partial charges are obtained using the am1-bcc method [131]. Finally, ligand atom coordinates and partial charges are stored in an output mol2 file. In specific frcmod and library *AMBER* files the corresponding ligand parameters for each ligand are saved, which are needed for *LEaP*.

LEaP generates topology and coordinate files of protein-ligand complexes for *AMBER* MD simulations and requires a PDB file for the protein and a mol2 file for the ligand geometry as input. Additionally, the frcmod and library file which contain ligand parameters need to be loaded. In this step the force field needs to be specified as well. We tested the ff99SB and the ff03 force fields. Since we use an implicit solvent model for our MD simulation PB radii need to be defined. We used the following radii: *PB radii bondi* [28]. These are the PB radii which are recommended to use when performing an implicit solvent MD simulation with the method of Mongan *et al.* [204]. It is essential to use the same implicit solvent model for both: MD simulations and MM-GBSA calculations.

MM-PB/GB parameters

Binding free energies can be determined with the MM-PB/GBSA method. This procedure can be divided into three parts. Molecular mechanical (MM) energy calculations describe the first part which are represented by a force field. The energy contributions calculated with MM are vdWs, electrostatic, and internal energies. We use a single trajectory approach in this work and therefore we can assume

that the internal energy of the bound and unbound protein-ligand complex is zero. The dielectric constant is the only parameter associated with the MM calculations for determining the electrostatic energy. The second contribution is described by the electrostatic contribution to the solvation free energy which is realized by the PB or the GB part of the MM-PB/GBSA method. The PB part is calculated with the pbsa program contained in *AMBER*, which represents a numerical solution to the PB equation [177, 179]. The GB part is calculated with GB models implemented in the *sander* program. The third part is described by the non-polar contribution of the solvation free energy (SA). This contribution is modeled as a linear dependence on the solvent accessible surface area (SASA). The linear dependence is determined by the following empirical parameters: surften and surfoff. The surface tension parameter (surften) can be set to 0.0072 or 0.00542 kcal/(mol·Å^2) and surfoff can be set accordingly to 0.0 or 0.92 kcal/mol. These values are taken from the *AMBER 10* manual [38].

The PB model is a dielectric continuum model. Thus, there is a dielectric boundary present in the system that separates the solute from the solvent. The solvent-excluded surface (SES) of the molecule usually defines this dielectric boundary. This surface is determined by the so-called rolling ball algorithm. A sphere is rolled over the surface of the molecule, thus the SES depends on the radius of the solvent probe sphere. The values for this radius are typically between 1.4 and 1.8 Å. One can also use 0.0 Åfor the radius which represents the vdW surface of the molecule [69].

For solving the PB equation the pbsa program in *AMBER* was used which is a finite-difference numerical solver [177, 179]. These methods use an equally spaced, three-dimensional grid to discretize the space. At each grid point the electrostatic potential is then calculated. The electrostatic potential present in between grid points is interpolated from the values at neighboring grid points. The accuracy of the results is affected by the spacing of the grid. When a large grid spacing is used the electrostatic potential is mostly approximated which results in poorly converged solvation energies. The electrostatic potential is smoothly represented when the grid spacing is small. However, this leads to higher computational run-times. Values between 0.3 and 1.0 Åyield sufficiently accurate results and have still a practicable run-time.

The PB and GB models are both dielectric continuum models. They describe a low dielectric solute surrounded by a high dielectric continuous medium with unlocalized charges. The external dielectric constant for aqueous solution is usually set to 80.0, the bulk dielectric constant of water. The internal dielectric constant represents the interior of the protein. However, a protein is not a uniform dielectric medium. Proteins have polar and non-polar regions featuring different dielectric constants. Typically, the internal dielectric constant is set to a value between 2.0 and 4.0. This is in accordance with internal dielectric constants for small organic molecules. A value of 4.0 for the internal dielectric constant reduces the solvation forces compared to a value of 2.0. In some studies internal dielectric constants are assumed to be higher and set to 12.0 or even 20.0 [9, 10, 258].

The value of the ionic strength is generally set to agree with experimental conditions. Physiological salt concentrations lie in the range between 0.1 and 0.2 M. Commonly, the value of the ionic strength is set to a value of 0.2 M. When the ionic strength is set to a non-zero value for the GB calculation a modified GB model, based on the Debye-Hückel limiting law for ion screening of interactions, is used.

There are several GB models available in *sander* for *AMBER 10* [38]. The GB model of Hawkins *et al.* [109, 110] with parameters of Tsui and Case [282, 283] has been most extensively tested. According the the *AMBER 10* manual [38] the GB model of Onufriev [220–222] is newer and is therefore recommended for most projects. Mongan *et al.* [204] presents the newest GB model.

5.3.3 Results

A thorough comparison of different *AMBER 10* [38] and Gromacs 4 [113] MD simulation protocols is presented in [256]. We tested explicit and implicit solvent models for MD simulations and also calculated binding affinities using PB and GB approaches. The best results, considering accuracy and run-time, yields the implicit solvent MD simulation protocol using *AMBER* in combination with MM-PB/GBSA methods and is described in the following.

5.3.3.1 Rapid molecular dynamics simulation protocol

Our advanced rapid MD simulation protocol is derived from the protocol presented by Brown and Muchmore [34]. In order to be able to compare our results to the previous study parameters were chosen identically. We used an implicit solvent protocol and tested the ff99SB [123] force field, which is a non-polarizable force field based on the force field of Cornell *et al.* [57] and the ff03 force field [72, 167]. The ff03 force field is a modified version of the ff99 force field [289], whereat main changes have been made primarily for the Φ and Ψ torsion angles. Since an implicit solvent model was chosen we tested several GB models, which all performed similarly. The results for the newest GB model developed by Mongan *et al.* [204] are presented here. Details about the MD simulation protocol can be found in Table E.1 in the appendix.

During all three steps of this protocol (minimization, relaxation, simulation) all hydrogen bonds were constrained using SHAKE [200, 250]. Furthermore, we used a time step of 2 fs and a distance cutoff for non-bonded interactions of 12.0 Å. During the GBSA simulation the surface area is calculated with the LCPO method [301]. The ionic strength is set to 0.2 M and the interior and exterior dielectric constants to 1.0 and 80.0, respectively. In the first step, the system is minimized by 500 steps of steepest-descent minimization.

The minimized system is subsequently equilibrated by gradually heating from 0 to 300 K over 6 ps with a time step of 2 fs. During the equilibration the the leapfrog Verlet algorithm [117] is used as numerical integrator. For temperature regulation the Langevin thermostat with a collision frequency of 2.5 ps^{-1} is used. All other parameters are defined as described for the minimization step.

The relaxed structure is used as input for 13 ps of MD simulation at constant temperature (NPT-MD). During the final 10 ps of this production run ten structural snapshots of the system are collected at equal time intervals. These snapshots represent the protein-ligand complexes at different points in time and are used as input structures for the MM-PB/GBSA approach. Thus, we include structural differences while determining binding free energies.

5.3.3.2 Optimized MM-PB/GBSA re-scoring parameter set

Binding free energies are calculated using the MM-PBSA and MM-GBSA method provided by *AMBER 10*. The final and optimized set for the MM-PBSA as well as for the MM-GBSA method are described in this section and are listed in Tables E.2 and E.3 in the appendix.

For the MM-PBSA calculation we used the *AMBER 10* pbsa program [177, 179]. The interior and exterior dielectric constants were set to default values. The distance between the grid points was 0.7 Å and the solvent probe radius was set to 0.0 Å which results in a vdW surface. The ionic strength is 0.2 M and parameters for the surface area calculation, surften and surfoff, are set to 0.0072 kcal/(mol·Å^2) and 0.0 kcal/mol, respectively.

Additional to the MM-PBSA parameter section the GB model needs to be specified for MM-GBSA calculations using the *AMBER 10* pbsa program. As described previously it is important to chose the same GB model as for the implicit solvent MD simulation protocol. Therefore, we used also the GB model described by Mongan *et al.* [204]. Other than that only parameter names differ between the MM-PBSA and the MM-GBSA sections.

5.3.3.3 Binding free energies for urokinase-ligand complexes

The experimentally determined -pK$_i$ values for the 18 urokinase inhibitors were taken from Brown and Muchmore [34]. We used these values to be able to measure the quality of our calculated binding free energies. Since we cannot use the in-house crystal structures of urokinase we docked the 18 ligands into the binding pocket of one urokinase structure (PDB id: 1owh) available in the PDB. When applying a core constraint to the naphthalene-derived ring of all ligands the docking method determines good ligand poses with the *Glide* scores ranging from -4.36 to -9.73. However, the ranking of the urokinase inhibitors, derived from the *Glide* scores, does not correlate well with experimental data. The Pearson correlation coefficient between *Glide* scores and experimental -pK$_i$ values is 0.63. Spearman's rank correlation coefficient gives an estimate about the quality of the rank correlation and has a value of 0.58 for *Glide* compared to experimental values, which is only slightly better than random selection. This means, that the *Glide* scoring function did not perform well in discriminating weak and strong binders for this urokinase ligand set. Docking scores and values determined by MM-PBSA and MM-GBSA calculations are listed in Tables E.4 and E.5 in the appendix.

The MM-PB/GBSA scoring methods were applied to the initial structure of the MD simulation for each urokinase-ligand complex. Thus, the quality of the MM-PB/GBSA scoring functions can be assessed considering only a single structure. The binding free energies of the initial structures of the MD simulation, using the MM-PBSA method compared to experimental values are shown in Fig. 5.18. The Pearson correlation coefficient of the MM-PBSA energies is 0.83 for both force fields (ff99SB and ff03). This indicates that the MM-PBSA method performs better than the *Glide* scoring function in discriminating weak and strong binders. The Pearson correlation coefficient for the MM-GBSA energies (Fig. 5.18) is 0.54 and 0.55 for the ff99SB and the ff03 force field, respectively. Thus, the MM-GBSA method yields even worse binding free energies compared to *Glide*. Spearman's rank

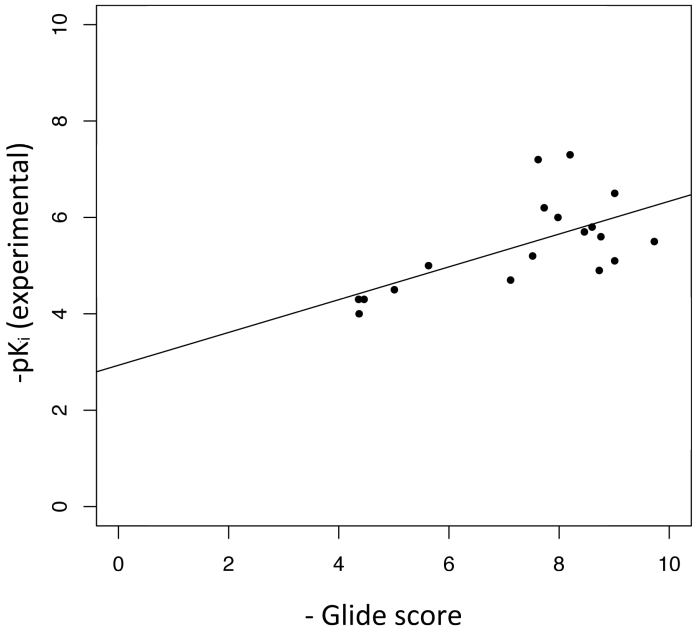


Figure 5.17: Correlation between docking scores and experimental binding free energies. Binding free energies for all 18 urokinase inhibitors estimated with an experimental binding assay are compared to binding scores determined by the docking program *Glide*. The correlation between binding free energies is 0.63 and the rank correlation is 0.58.

correlations are for the MM-PBSA method 0.75 and 0.79 and for the MM-GBSA method 0.49 and 0.50 for the ff99SB and ff03 force fields, respectively. This means that the MM-PBSA approach for each tested force field discriminates weak and strong binders comparably well even when structural alterations are not accounted for.

The best performing MM-PBSA parameter set combined with the rapid MD simulation protocol yields a Pearson correlation coefficient and a Spearman's rank correlation of 0.92, see Fig. 5.18. This implies that including flexibility during scoring improves binding free energy predictions. Pearson's correlation coefficient for the ff99SB and the ff03 force field for the MM-PBSA method is 0.92 and 0.81, respectively. Thus, the MM-PBSA method performs outstandingly well when additionally accounting for flexibility. The Pearson correlation for the MM-GBSA methods using ff99SB and ff03 are 0.93 and 0.80. This implies that the ff99SB force field yields slightly better results. However, it should be mentioned when using MM-GBSA and accounting for flexibility the improvements in binding free energies are as well drastically.

Comparing our results with the rank correlation of Brown and Muchmore [34] we were successful in producing better results for all methods, force fields, and most parameter sets. The rank correlation for their implicit solvent approach is 0.78 and for the explicit solvent approach 0.90. Our rapid MD re-scoring protocol combined with the ff99SB force field yields even higher rank correlations compared to their explicit solvent approach. Another important property, additional to the good quality of our method, is the low run-time, which is comparable to the high-throughput protocol of Brown and Muchmore [34]. Our implicit solvent high-throughput protocol requires approximately 250 CPU

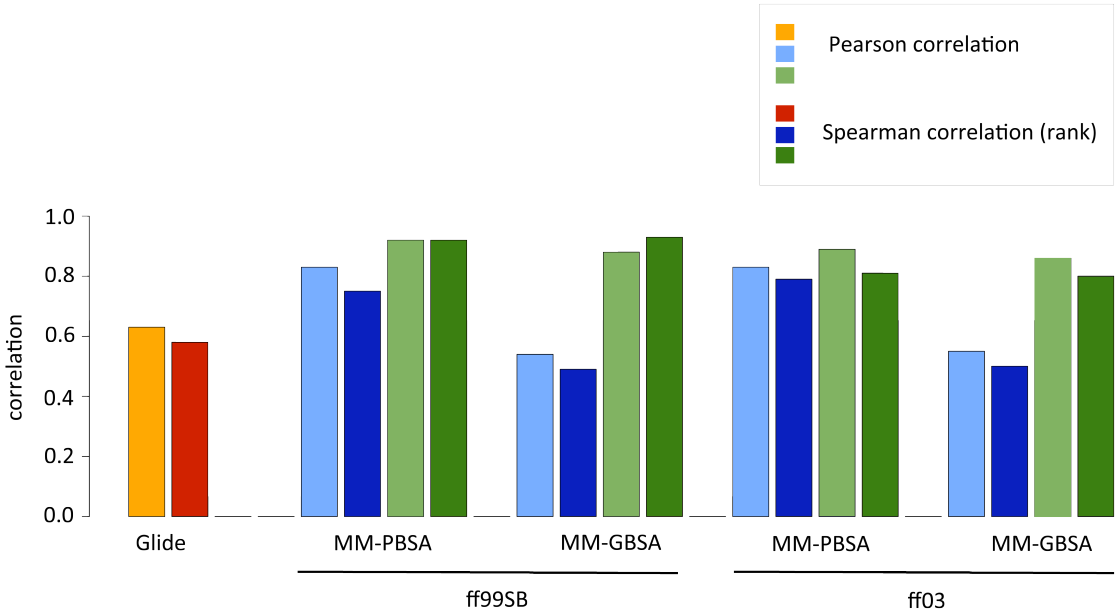


Figure 5.18: Correlation between *in silico* and *in vitro* binding free energies: with and without flexibility. Binding free energies for all 18 urokinase inhibitors estimated with an experimental binding assay are compared to four different force field-based approaches: MM-PBSA and MM-GBSA both tested with the ff99SB and ff03 force field. The correlation between *in silico* and *in vitro* binding free energies using the Pearson correlation and the rank ordering using Spearman's rank correlation were determined. Blue indicates that flexibility was not considered, while green indicates that flexibility was accounted for.

minutes per structure, serially executed on a Quad-Core AMD Opteron Processor 2354 running at 2.2 GHz. Brown and Muchmore [34] report a run-time on the order of 100 CPU minutes per structure, serially executed on a 3.0 GHz Xeon processor. Implicit solvent approaches have an overall faster run-time compared to explicit solvent approaches in these kind of experiments. In explicit solvent systems the water molecules need to be equilibrated first until an accurate production run can be started. This is not necessary for implicit solvent models, which typically converge faster. Moreover, we could demonstrate that we outperform the explicit solvent protocol in [34] in terms of accuracy.

5.3.4 Discussion

Overall we could show that our rapid MD simulation protocol combined with selected parameters for the MM-PB/GBSA methods performs better considering quality and equally with respect to run-time compared to similar methods. Moreover, we could demonstrate that including flexibility improves re-scoring results tremendously. In the following we discuss certain parameters for the MM-PB/GBSA methods.

Typically, in aqueous solutions the solvent probe radius is set to 1.4 Å, however we set the value to zero. This affects the calculation of the SES, which defines the dielectric boundary of the PB model. In a study, Dong *et al.* [69] presented good results when using a solvent probe radius of zero

in combination with setting the internal dielectric constant to 4.0. Dong *et al.* [69] refer to other studies [9, 10, 36] in which the following was reported. The desolvation cost of charges and the strengths of charge-charge interactions are consistently overestimated when using SES as dielectric boundary and a dielectric constant between 2.0 and 4.0. Antosiewicz *et al.* [9, 10] recommend to use a value of 20 for the solute dielectric constant. This would reduce the desolvation cost and weaken the charge-charge interactions. We used the solution of Dong *et al.* [69] to reduce the desolvation cost and weaken the charge-charge interactions and set the solvent probe radius to zero. Thus, a vdW surface is used as dielectric boundary.

As mentioned previously, the internal dielectric constant is usually set to a value between 2.0 and 4.0, a common value for organic molecules. However, the correct choice for this value has been controversially discussed by several groups. Antosiewicz *et al.* [9, 10] proposed a value of 20 for the protein dielectric constant. Schutz and Warshel [258] discuss this topic in detail. Several studies report [193, 262, 295, 296] that higher values of the internal dielectric constant give reasonable results for charge-charge interactions. However, using a different value than 1.0 for the PB energy calculation using the MM-PBSA method, while retaining a value of 1.0 for the MM energy calculations leads to inconsistencies [85, 86, 176]. To resolve this problem the dielectric constant in the MM energy calculation could also be increased. This leads to another issue. H-bonds are represented by electrostatic and vdW terms in common force fields. The energy contribution of H-bonds is decreased when using a higher value than 1.0 for the dielectric constant. To avoid inconsistencies we used a consistent value of 1.0 for interior dielectric constants and refer to Dong *et al.* [69].

The PB model was calculated with the pbsa program of the *AMBER* suite, which is a finite-difference numerical solver for the PB equation [177, 179]. The finite-difference method commonly uses a uniform, three-dimensional grid to discretize the space. The spacing of this grid strongly affects the computational cost and the accuracy of the solution. Going from a very fine grid spacing of 0.3 Å to a more coarse grid spacing of 0.7 Å we observe no loss in accuracy [256]. For all calculations in this thesis, a grid spacing of 0.7 Åper grid was used. This value provided the optimal tradeoff between accuracy and run-time.

How the ionic strengths affects MM-PBSA results was evaluated in [256]. Usually, the value of the ionic strength is set to agree with experimental data. In most studies, the ionic strength is set to values equal or less than 0.2 M. We tested values of 0.1, 0.2, 0.5, and 1.0 M, but only minor effects of different values on the correlation of the MM-PBSA energy estimates to experimental data could be observed. Thus, the ionic strength was set to 0.2 M, which is consistent with the salt concentration used in our MD simulations.

Furthermore, the effect of non-polar solvation free energy contributions on the results of the MM-PBSA method was studied. It was modeled by a linear dependence on the SASA. The SASA was calculated using *Molsurf*, which is an implementation of the algorithm described by Mike Connolly [55]. The linear dependence is described by the parameters surften and surfoff. The choice of these parameters is mostly empirical. In most studies, surften is set to 0.0072 kcal/(mol·Å^2) or 0.00542 kcal/(mol·Å^2) and surfoff is set to 0.0 kcal/mol or 0.92 kcal/mol, accordingly. We tested different values for surften

and surfoff, but no large effects on the overall binding free energy could be observed. However, using a surften value of 0.0072 kcal/(mol·Å^2) and a surfoff value of 0.0 kcal/mol showed the best results. Overall the non-polar contribution to the binding free energy is small compared to the electrostatic contribution.

We compared the *AMBER* force fields ff99SB [123] and ff03 [72]. The ff99SB force field [123] is a modification of the ff99 force field [289] with improved dihedral terms. The ff99 force field is based on the Cornell *et al.* force field [57], denoted as ff94 in *AMBER*. The ff03 force field [72] originates also from the ff94/ff99 force field. However, a fundamentally different concept to the derivation of partial atomic charges was used [123]. The ff99SB as well as the ff03 force field are well suited for proteins, nucleic acids, and organic molecules. Combining different force fields for structure generation and binding free energy calculations is not recommended [300]. Therefore, we applied the same force field for MD simulations and MM-PB/GBSA calculations.

The MM-PB/GBSA parameters were solely adapted for this urokinase-ligand test set and we did not demonstrate the applicability of our method to other protein-ligand complexes. However, all 18 urokinase ligands are highly similar, thus they are difficult to discriminate for most scoring functions. Therefore, this test set reflects the situation screening large compound libraries were scoring functions need to distinguish between strong and weak inhibitors having similar structures. We could show that our method can perform this task.

Compared to Brown and Muchmore [34] we performed docking to place the ligands in the protein target binding pocket instead of applying the re-scoring protocol on XRD protein-ligand complexes. Therefore, we can conclude that our re-scoring protocol performs well on docked protein-ligand structures. This situation is often faced in drug discovery procedures when millions of compounds are screened with crude scoring functions in a first instance. The best hits could then be re-scored using our rapid MD re-scoring protocol to discriminate between strong and weak binders. With this method it is possible to eliminate false positive compounds at the top of screening hit lists, especially when comparing structurally similar compounds. Additionally, more advanced PB solvers, such as *APBS* [12] for estimating binding free energies should be tested. These specific programs are able to determine binding free energies more accurately.

6 Conclusion

The same receptor can also signal through different intracellular pathways depending on the identity of the bound ligand. This is due to the relative flexibility of the receptor in the membrane, which allows different agonists or inverse agonists to stabilize different active or inactive forms. [172]

Brian K. Kobilka and
Robert J. Lefkowitz, 2012
won the Nobel Price in
Chemistry 2012 for their
studies on GPCRs

In the scientific background [172] on the Nobel Prize in Chemistry in 2012 awarded to Brian K. Kobilka and Robert J. Lefkowitz for their studies on G-protein-coupled receptors flexibility of proteins is acknowledged. Despite its significance, flexibility of biological macromolecules is often not optimal represented in computational structural modeling approaches. One method to display structural changes of biomolecules *in silico* is molecular dynamics. Recent advances in performance and scaling of MD simulation software facilitate long simulation times to account even for large conformational changes of proteins. We emphasize the valuable contribution of MD simulations by studying the flexibility of protein-DNA and protein-ligand complexes.

In contrast to animals, plants do not have the ability to avoid disadvantageous conditions by migration to more favorable locations. Due to their lack of flexibility, plants need to adapt to varying conditions by other factors, such as gene regulation. One of the largest classes of proteins that regulate gene expression in plants is the WRKY protein family. They help plants to overcome different stress situations and are involved in the regulation of developmental processes. 72 WRKY proteins are present in *Arabidopsis thaliana*, which all bind to the W-box sequence 'TTGACY'. We studied

DNA-binding specificities of different WRKY proteins at the atomic level. Since there is little structural data available, we developed a protocol for modeling arbitrary protein-DNA complexes. When performing such a modeling task it is beneficial to include as much structural knowledge as possible. Structures of closely related proteins in complex with DNA give a first idea on how the unknown protein-DNA complex is formed. Since, a related protein that was bound to DNA was available for WRKY proteins we could create a well-defined first WRKY-DNA complex model. Including flexibility after this modeling step refines the protein-DNA structure. We could show that our advanced MD simulation protocol developed for protein-DNA complexes is suitable for this type of refinement. Amino acids present at the binding interface change their orientation and interact with DNA base pairs during molecular dynamics. Especially, side chains of the second β-strand, known to interact with the W-box sequence form contacts with the DNA. Before the refinement step most of these interactions were not present and thus we could demonstrate the importance of accounting for flexibility when modeling protein-DNA complexes. Moreover, MD simulations of protein-DNA complexes reveal specific interactions formed between protein side chains and DNA base pairs. Interestingly, we could observe that most interactions are formed between amino acids and methyl groups of thymine nucleotides, which was also described by Yamasaki *et al.* [309]. Commonly, specific H-bond patterns are detected at protein-DNA complex interfaces, which could not be shown for WRKY proteins. It was reported in [53] that a closely related protein to the WRKY protein family features a novel mode of sequence-specific DNA recognition. We can observe this mode of DNA recognition also for WRKY proteins, which could explain the overrepresentation of interactions with thymine methyl groups.

WRKY proteins of group I possess two DNA-binding domains, whereby the N-terminal DNA-binding domain (nDBD) structure is not available in databases and also not the complete structure of group I proteins. Due to its sequence it was speculated that the nDBD features a similar structure compared to the C-terminal DNA-binding domain (cDBD). However, so far the binding to DNA of WRKY nDBDs could not be shown. Our refined WRKY33 nDBD-DNA complex illustrates for the first time that both domains bind to DNA in a likewise fashion. We could give a first structural indication that the nDBD might also interact with the DNA. Binding affinities between the W-box sequence and the WRKY33 cDBD and WRKY33 nDBD structures could be determined using our modeled complex structure. In these calculations we also account for flexibility. Our binding affinities are congruent with experimental results. These results [30] demonstrate for the first time that this domain is also involved in DNA binding. How both domains are connected by the linker region and how group I WRKY proteins bind to DNA in detail remains still elusive. However, our model illustrates a possible complex structure which can be used for further analyses.

We analyzed also how contacts between WRKY TFs and DNA account for specificity by comparing binding interfaces of *At*WKRY11 and *At*WRKY50 DBDs. We could identify an amino acid difference at the binding interface between *At*WRKY11 DBD and *AtWRKY*50 DBD and suggest a possible contact with the second thymine of the W-box sequence. This could also be experimentally identified. Hence, our *in silico* and *in vitro* muation studies share consistent results. We also performed MD simulations for wild type and mutated protein-DNA complexes in explicit solvent. Side chains at the

binding interface form contacts with DNA base pairs during the simulation, which were not present in modeled static protein-DNA complexes. We could also show different binding preferences of wild type and mutated side chains forming other contacts with the DNA. Since our results can explain experimental findings at the atomic level we expect our model to be valid. Without MD simulations structural changes of amino acid side chains at the binding interface could not have been displayed. Thus, MD simulations give valuable insights and refined our protein-DNA complex models. In our study we identified the GAC motif to be important for binding for WRKY proteins. This suggests a more degenerative DNA-binding motif for WRKY proteins as assumed so far (W-box sequence: 'TTGACC'). However, we could not detect specific contacts between GAC and protein side chains, except for a tyrosine residue which might interact with the methyl group of thymine. Specificity in the family of WRKY proteins is probably mediated by complex formation with other proteins and DNA shape recognition. We identified a possible Hoogsteen DNA bp in the DNA binding motif of WRKY proteins [30] which results in a specific DNA structure of the 'TTGACC' sequence.

Specific interactions between protein and small molecules (ligands) can only be modeled with highest accuracy by state-of-the-art molecular docking approaches when a high quality protein structure is available. Structural changes of the protein binding pocket as well as large protein backbone motions are challenging to address in molecular docking and are current research topics. Our approach can generate various protein conformations when a single protein structure is given and identifies relevant protein conformations. These protein conformations can then be used as input structures for docking. However, open questions remain. In this work we did not investigate topics, such as what the best strategy would be when docking into a set of different protein structures or what approach is the most promising to discriminate between good and inferior binders placed into different protein structures. Molecular docking strategies still perform best, when the ligand is placed in its originating crystal structure. Current research is performed to tackle problems such as docking into flexible proteins. When only a single protein structure is available one might want to examine the protein motion. As described for HIV-1 protease, a prominent target in anti-AIDS drug discovery, structural alterations should be accounted for [122, 124]. Our approach identifies all known and relevant HIV-1 protease conformations available in the PDB. When starting from the "semi-open" HIV-1 protease structure we can generate the "closed" conformation and also observe flap tip curling motions. When starting from the "closed" conformation, we observe structural alterations and structures similar to the "semi-open" conformation. However, all described known conformations could not clearly be identified, e.g., the "fully-open" structure of HIV-1 protease was not generated. DHFR, another important drug target in cancer research, is known for its shear motion. In this study, we did not examine the binding of inhibitors to different DHFR conformations. We only generated different structures starting from one protein conformation. We could show that we can identify at least one protein conformation that is highly similar to each of the available PDB structures. The same could be illustrated for AR. AR predominantly alters side chain orientations in its binding site. Some larger conformational changes could be observed, since our approach is better suited to identify structures with large protein backbone dynamics.

Besides studying large protein backbone motions, we also examined protein flexibility while estimating binding affinities of protein-ligand complexes. After computational molecular docking, protein-ligand complexes are commonly further investigated to improve the rank ordering of docking hit lists. Especially, when evaluating very similar compounds scoring functions of docking programs cannot discriminate between good and inferior binders. We performed the re-scoring using our approach with and without accounting for protein dynamics and could show that flexibility improves the rank-ordering compared to experimental binding affinities. Thus, we could demonstrate that an implicit solvent MD re-scoring protocol can be used after a docking study and is not only applicable to crystal structures as described by Brown and Muchmore [34]. It drastically improved the correlation to experimental binding affinities, when tested with highly resembling urokinase inhibitors.

Overall, we could demonstrate that modeling protein flexibility yields valuable insight into protein-DNA and protein-ligand complexes. From small side chain alterations at protein-DNA complex interfaces through to large backbone dynamics of protein structures—these tasks can be solved by molecular dynamics.

A Abbreviations

AR	aldose reductase
At	*Arabidopsis thaliana*
bHLH	basic helix-loop-helix
bZip	basic leucine zipper
bp	base pair
BP	binding pocket
cDBD	C-terminal DNA-binding domain
CG	conjugate gradient
DBD	DNA-binding domain
DHFR	dihydrofolate reductase
DNA	deoxyribonucleic acid
FF	force field
GB	Generalized Born
GCM	Glia cell missing
H-bond	hydrogen bond
HTH	helix-turn-helix
PB	Poisson Boltzmann
PCA	principal component analysis
PDB	protein databank
PME	particle mesh Ewald
PSSM	position-specific scoring matrix
QM	quantum mechanics
RNA	ribonucleic acid
RMSD	root-mean-square deviation
TIM	triose phosphate isomerase
TIP3P	transferable intermolecular potential three point
MC	Monte Carlo
MD	molecular dynamics
nDBD	N-terminal DNA-binding domain
MM	molecular mechanics
Mm	*Mus musculus*
MM-GBSA	Molecular Mechanics Generalized Born Surface Area
MM-PBSA	Molecular Mechanics Poisson Boltzmann Surface Area
$NADP^+$	nicotin amideadenine dinucleotide phosphate

NMA	normal mode analysis
NMR	nuclear magnetic resonance
PDB	protein databank
RNA	ribonucleic acid
SD	steepest descent
SASA	solvent accessible surface area
SES	solvent-excluded surface
TF	transcription factor
TFBS	transcription factor binding site
XRD	X-ray diffraction
ZF	zinc finger

B First Appendix

Table B.1: C_α atoms for mapping 2ayd onto 1odh. For superimposing two protein structures (PDB ids: 2ayd and 1odh) we used 33 C_α atoms. The numbers match the residue ids in the corresponding PDB structures. The C_α atoms were stored in a paired list and served as input for the atom bijection method in *BALL* [114].

PDB id	Residues 1 - 6	Residues 7 - 13	Residues 14 - 19	Residues 20 - 33
2ayd	Tyr 310 - Tyr 315	Ser 328 - Ser 334	Val 339 - Glu 344	Leu 352 - Met 365
1odh	Trp 59 - Thr 64	Leu 72 - Gly 78	Val 133 - Arg 138	Phe 143 - Arg 156

Table B.2: C_α atoms for calculating the RMSD between 2ayd and 1odh. For calculating the RMSD between two protein structures, the mapped 2ayd and the original 1odh structure, we used 26 C_α atoms. The numbers match the residue ids in the corresponding PDB structures. The C_α atoms were stored in a paired list and served as input for the atom bijection method in *BALL* [114].

PDB id	Residues 1 - 5	Residues 6 - 12	Residues 13 - 19	Residues 20 - 26
2ayd	Trp 312 - Gly 316	Arg 327 - Ser 333	Val 339 - Arg 345	Leu 352 - Glu 358
1odh	Met 61 - Asn 65	Ile 71 - Leu 77	Val 133 - His 139	Phe 143 - Lys 149

Table B.3: C_α atoms for calculating RMSDs between docked WRKY1-DNA complexes and 2lex. For calculating RMSDs between the docked (and our mapped 2ayd) WRKY1-DNA complexes and the 2lex structure, we used 23 C_α atoms. The numbers match the residue ids in the corresponding PDB structures.

PDB id	Residues 1 - 7	Residues 8 - 12	Residues 13 - 18	Residues 19 - 23
2ayd	Trp 312 - Lys 318	Arg 327 - Arg 331	Lys 340 - Arg 345	Ile 354 - Glu 358
2lex	Trp 414 - Lys 420	Arg 429 - Lys 433	Arg 424 - Arg 447	Val 456 - Glu 460

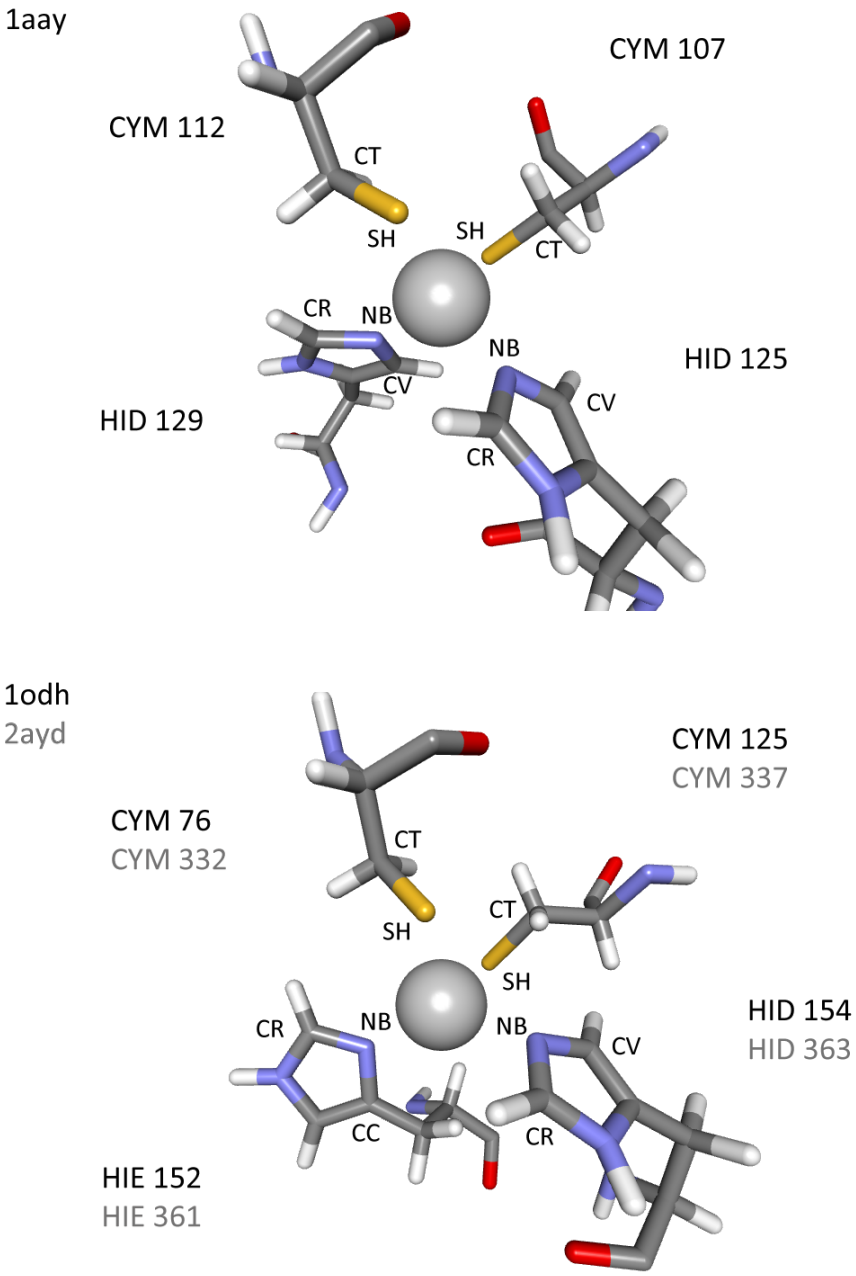


Figure B.1: Coordinated zinc ions. In Zif268 (PDB id: 1aay) the zinc ion is coordinated by two cysteine residues and two histidines residues, which are both protonated at their δ nitrogen atom (HID). In GCM1 (PDB id: 1odh) and all WRKY DBDs we studied in this thesis (PDB id: 2ayd) the zinc ion is coordinated by two cysteine and two histidine residues. One of the histidine residues is protonated at the δ (HID), the other one at the ϵ nitrogen (HIE) atom. The atom types which are relevant for the frcmod file, see Table B.5 are indicated.

Table B.4: Library file for zinc ion parameters. The library file is used as input file for *LEaP*. It defines the new name of the zinc ion for which specific parameters are stored in the frcmod file.

```
!!index array str
 "Z4"
!entry.Z4.unit.atoms table str name str type int typex int resx int flags int seq
 int elmnt dbl chg
 "Z4" "Z4" 0 1 196609 1 30 0.666
!entry.Z4.unit.atomspertinfo table str pname str ptype int ptypex int pelmnt dbl pchg
 "Z4" "Z4" 0 -1 0.488
!entry.Z4.unit.boundbox array dbl
 -1.000000
 1.570797
 0.0
 0.0
 0.0
!entry.Z4.unit.childsequence single int
 2
!entry.Z4.unit.connect array int
 0
 0
!entry.Z4.unit.hierarchy table str abovetype int abovex str belowtype int belowx
 "U" 0 "R" 1
 "R" 1 "A" 1
!entry.Z4.unit.name single str
 "Z4"
!entry.Z4.unit.positions table dbl x dbl y dbl z
 0.0 0.0 0.0
!entry.Z4.unit.residueconnect table int c1x int c2x int c3x int c4x int c5x int c6x
 0 0 0 0 0 0
!entry.Z4.unit.residues table str name int seq int childseq int startatomx str restype
 int imagingx
 "Z4" 1 2 1 "?" 0
!entry.Z4.unit.residuesPdbSequenceNumber array int
 0
!entry.Z4.unit.solventcap array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.Z4.unit.velocities table dbl x dbl y dbl z
 0.0 0.0 0.0
```

Table B.5: frcmod file for zinc ion parameters. The frcmod file is used as input file for *LEaP*. It defines missing torsion parameters for the zinc ion. Parameters were provided by Annette Hoöglund for the Zif268 (PDB id: 1aay) ZF structure based on parameters developed by Walker [288] and Ryde [251] and complemented by us if necessary.

```
MASS
Z4            65.39

BOND
Z4-NB        262.0    2.181
Z4-SH        342.0    2.293

ANGL
SH-Z4-SH      40.0    109.0
NB-Z4-NB      40.0    109.0
SH-Z4-NB      40.0    109.0
CR-NB-Z4      50.0    120.0
CV-NB-Z4      50.0    120.0
CC-NB-Z4      50.0    120.0
CT-SH-Z4      79.1    111.6

DIHE
CV-NB-Z4-SH      1     0.10    180.0    2.
CV-NB-Z4-NB      1     0.10    180.0    2.
CR-NB-Z4-SH      1     0.10    180.0    2.
CR-NB-Z4-NB      1     0.10    180.0    2.
CC-NB-Z4-SH      1     0.10    180.0    2.
CC-NB-Z4-NB      1     0.10    180.0    2.
CT-SH-Z4-SH      1     0.10      0.0    3.
CT-SH-Z4-NB      1     0.10      0.0    3.

IMPR
CR-CV-Z4-NB      1     50.0    180.0    2.
CV-CR-Z4-NB      1     50.0    180.0    2.
CC-CR-Z4-NB      1     50.0    180.0    2.
CR-CC-Z4-NB      1     50.0    180.0    2.
CC-CV-Z4-NB      1     50.0    180.0    2.
CV-CC-Z4-NB      1     50.0    180.0    2.

NONB
Z4            1.950    0.250                 Walker (1.85 0.06)
```

Table B.6: MD simulation protocol for protein-DNA complexes. Parameters for the minimization, relaxation, and production run steps are presented. The last three lines in the minimization step state how Cartesian constraints are applied to the system. The two lines which describe the weight and restraintmask are presented in Table B.7 for each step.

Steps A-D: minimization with Cartesian constraints

```
imin=1, maxcyc=5000, ncyc=4000,     invoke minimization
ntpr=5,                             print frequency
ntr=1,                              turn on Cartesian constraints
restraint_wt=20.0,                  weight for Cartesian constraint
restraintmask='1-103',              atoms in residues 1-103 constrained
```

Step E: relaxation with constant volume MD (NVT-MD)

```
imin=0, irest=0, ntx=1,             invoke MD simulation
nstlim=50000, dt=0.002,             run for 100 ps
ntpr=500, ntwx=500, ntwr=500,       output frequency
ntt=3, gamma_ln=2.0,                temperature control
temp0=100, tempi=300,               start/end temperature
ntb=1, ntp=0, ntc=2, ntf=2,         periodic boundary conditions, SHAKE, etc.
ig=209858,                          seed for random number generator
nrespa=2,                           slowly varying forces are evaluated every 2 steps
```

Steps F-P : relaxation with constant pressure MD (NPT-MD)

```
imin=0, irest=1, ntx=5,             restart MD simulation
nstlim=50000, dt=0.002,             run for 100 ps
ntpr=500, ntwx=500, ntwr=500,       output frequency
ntt=3, gamma_ln=2.0,                temperature control
temp0=300, tempi=300,               temperature control
ntp=1, taup=1.0,                    pressure control
ntb=2, ntc=2, ntf=2,                periodic boundary conditions, SHAKE, etc.
nrespa=2,                           slowly varying forces are evaluated every 2 steps
```

Step Q: production run

```
imin=0, irest=1, ntx=5,             restart MD
nstlim=10000000, dt=0.002,          run for 20 ns
ntpr=500, ntwx=500, ntwr=500,       output frequency
ntt=3, gamma_ln=2.0,                temperature control
temp0=300, tempi=300,               temperature control
ntp=1, taup=1.0,                    pressure control
ntb=2, ntc=2, ntf=2,                periodic boundary conditions, SHAKE, etc.
nrespa=2,                           slowly varying forces are evaluated every 2 steps
```

Table B.7: Constraints for each step in the MD simulation protocol. In each step, except step D, during our MD simulation protocol for protein-DNA complexes constraints are applied on the system. In addition to the `ntr` parameter, which is set to 1, these lines are added to the parameters shown in Table B.6 as illustrated for the first minimization.

Step A	
`restraint_wt=20.0,`	*weight for Cartesian constraint*
`restraintmask=':1-103',`	*DNA and protein atoms*
Step B	
`restraint_wt=20.0,`	*weight for Cartesian constraint*
`restraintmask='!@H=',`	*all atoms except hydrogen atoms*
Step C	
`restraint_wt=20.0,`	*weight for Cartesian constraint*
`restraintmask=':26-103@CA,C,O,N',`	*protein backbone atoms*
Step E	
`restraint_wt=5.0,`	*weight for Cartesian constraint*
`restraintmask=':1-26 | :27-103@CA,C,O,N',`	*DNA and protein backbone atoms*
Step F	
`restraint_wt=5.0,`	*weight for Cartesian constraint*
`restraintmask=':1-26 | :27-103@CA,C,O,N',`	*DNA and protein backbone atoms*
Step G	
`restraint_wt=4.5,`	*weight for Cartesian constraint*
`restraintmask=':1-26 | :27-103@CA,C,O,N',`	*DNA and protein backbone atoms*
Step H	
`restraint_wt=4.0,`	*weight for Cartesian constraint*
`restraintmask=':1-26 | :27-103@CA,C,O,N',`	*DNA and protein backbone atoms*
Step I-N	
`restraint_wt=x,`	*x = 3.5, ..., 1.0*
`restraintmask=':1-26 | :27-103@CA,C,O,N',`	*DNA and protein backbone atoms*
Step O	
`restraint_wt=1.0,`	*weight for Cartesian constraint*
`restraintmask=':1-26 | :27-103@CA,C,O,N',`	*DNA and protein backbone atoms*
Step P	
`restraint_wt=1.0,`	*weight for Cartesian constraint*
`restraintmask=':1,13,14,26',`	*flanking DNA base pairs*
Step Q	
`restraint_wt=1.0,`	*weight for Cartesian constraint*
`restraintmask=':1,13,14,26',`	*flanking DNA base pairs*

Table B.8: Parameters for the PB section. We determined binding free energies with the MM-PBSA approach provided by *AMBER* using the parameters described in the table.

PROC	2	*use Amber pbsa program*
REFE	0	*reference state for PB calculation*
INDI	1.0	*interior dielectric constant*
EXDI	80.0	*solvent dielectric constant*
SCALE	2	*lattice spacing in number of grids per Å*
LINIT	1000	*number of interactions with the linear PB equation*
PRBRAD	1.4	*solvent probe radius in Å*
ISTRING	200.0	*ionic strength in mM for PB solver*
RADIOPT	0	*use radii from prmtop files for PB calculation*
NPOPT	1	*restart MD*
INP	1	*use SASA to correlate the total non-polar solvation free energy*
SURFTEN	0.00542	*value to compute solvation free energy*
SURFOFF	0.92	*value to compute solvation free energy*

Table B.9: Binding free energies of *At*WRKY1 cDBD and *At*WRKY33. Binding free energies between 14 different DNA sequences and *At*WRKY1 cDBD are calcula

Protein		DNA sequence	DNA sequence (antisense)	ΔG
*At*WRKY1 cDBD	1	5'-TCAAAGTTGACC-3'	5'-GGTCAACTTTGA-3'	-51.04±6.61
*At*WRKY1 cDBD	2	5'-CAAAGTTGACCA-3'	5'-TGGTCAACTTTG-3'	-38.62±6.22
*At*WRKY1 cDBD	3	5'-AAAGTTGACCAA-3'	5'-TTGGTCAACTTT-3'	-65.23±7.50
*At*WRKY1 cDBD	4	5'-AAGTTGACCAAT-3'	5'-ATTGGTCAACTT-3'	-54.04±6.69
*At*WRKY1 cDBD	5	5'-AGTTGACCAATA-3'	5'-TATTGGTCAACT-3'	-38.71±6.37
*At*WRKY1 cDBD	6	5'-GTTGACCAATAA-3'	5'-TTATTGGTCAAC-3'	-42.20±7.87
*At*WRKY1 cDBD	7	5'-TTGACCAATAAT-3'	5'-ATTATTGGTCAA-3'	-58.50±6.52
*At*WRKY1 cDBD	8	5'-ATTATTGGTCAA-3'	5'-TTGACCAATAAT-3'	-41.93±6.49
*At*WRKY1 cDBD	9	5'-TTATTGGTCAAC-3'	5'-GTTGACCAATAA-3'	-56.64±6.00
*At*WRKY1 cDBD	10	5'-TATTGGTCAACT-3'	5'-AGTTGACCAATA-3'	-66.83±8.48
*At*WRKY1 cDBD	11	5'-ATTGGTCAACTT-3'	5'-AAGTTGACCAAT-3'	-47.09±7.42
*At*WRKY1 cDBD	12	5'-TTGGTCAACTTT-3'	5'-AAAGTTGACCAA-3'	-42.91±6.60
*At*WRKY1 cDBD	13	5'-TGGTCAACTTTG-3'	5'-CAAAGTTGACCA-3'	-55.46±9.72
*At*WRKY1 cDBD	14	5'-GGTCAACTTTGA-3'	5'-TCAAAGTTGACC-3'	-51.05±7.76
*At*WRKY33 cDBD	4	5'-AAGTTGACCAAT-3'	5'-ATTGGTCAACTT-3'	-75.32±4.63
*At*WRKY33 nDBD	4	5'-AAGTTGACCAAT-3'	5'-ATTGGTCAACTT-3'	-51.58±6.82

Table B.10: WRKY homology model validation. Homology modeling was performed using *At*WRKY1 cDBD (PDB id: 2ayd) as template to obtain structures of four WRKY DBDs. Anolea, ProSA, and what-check validation results are shown.

WRKY DBDs	**W1 c**	**W11**	**W50**	**W33 c**	**W33 n**
Anolea	-1.89	-1.68	-1.67	-1.42	-1.55
ProSA	-3.37	-3.01	-4.05	-3.31	3.00
what-check					
Structure Z-scores, positive is better than average					
2nd generation packing quality	-0.712	-1.486	-1.330	-1.224	-1.347
Ramachandran plot appearance	1.847	0.194	1.970	2.102	0.920
chi-1/chi-2 rotamer normality	1.057	-0.128	-1.177	-0.006	-1.231
Backbone conformation	-0.857	-3.903	-1.230	-1.673	-3.221
RMS Z-scores, should be close to 1.0					
Bond lengths	0.624	0.693	0.700	0.660	0.705
Bond angles	0.839	1.133	0.974	0.974	1.031
Omega angle restraints	1.181	1.279	1.189	1.181	1.199
Side chain planarity	1.429	1.099	1.112	1.230	1.038
Improper dihedral distribution	0.730	0.890	0.761	0.740	0.736
B-factor distribution	0.502	0.335	0.339	0.339	0.337
Inside/Outside distribution	1.139	1.173	1.124	1.175	1.110

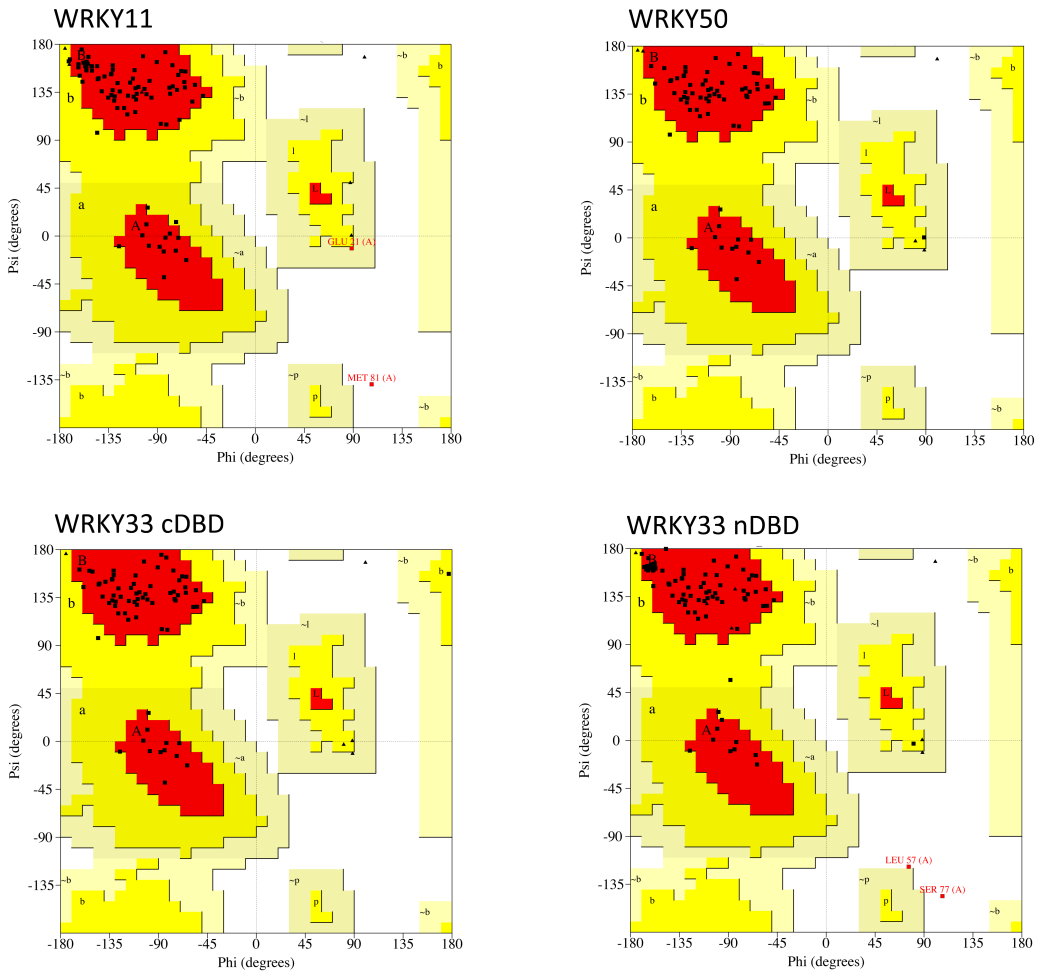


Figure B.2: WRKY homology model validation. Ramachandran plots generated with what-check show the excellent quality of all four homology models.

C Second Appendix

Table C.1: Aldose reductase PDB structures. 47 aldose reductase (AR) XRD structures were obtained from the PDB (sequence similarity to 2acr > 95% and no amino acid differences compared to residues 1-315 of 2acr.)

1ADS, 1Z8A, 2F2K, 2IKJ, 2PD5, 2PDJ, 2PDQ, 3LBO,
1EF3, 2ACQ, 2FZ8, 2INZ, 2PD9, 2PDK, 2PDU, 3RX2,
1EL3, 2FZ9, 2IQD, 2PDB, 2PDL, 2PDW, 3RX3, 1IEI,
2ACS, 2HVN, 2IS7, 2PDC, 2PDM, 2PDY, 3RX4, 1USO,
2DUX, 2HVO, 2NVC, 2PDH, 2PDN, 2R24, 3S3G, 1XGD,
2DUZ, 2IKH, 2NVD, 2PDI, 2PDP, 3DN5, 3U2C

Table C.2: Dihydrofolate reductase PDB structures. 33 dihydrofolate reductase (DHFR) XRD structures were obtained from the PDB (sequence similarity to 1pdb > 95% and no amino acid differences compared to residues 1-186 of 1pdb.)

1DHF, 1MVT, 1YHO, 2W3B, 3NTZ, 3NXX, 1HFR, 1OHJ,
1S3U, 2C2S, 2W3M, 3NUO, 3NXY, 1KMS, 1OHK, 1S3V,
2C2T, 3FS6, 3NXR, 3NZD, 1KMV, 1PD8, 1S3W, 2DHF,
3GHW, 3NXT, 3S7A, 1MVS, 1PD9, 1U72, 2W3A, 3GYF,
3NXV

Table C.3: HIV-1 protease PDB structures. 8 HIV-1 protease XRD structures were obtained from the PDB (sequence similarity to 1hvr is 100%)

1DMP, 1HVH, 1HWR, 1QBS, 1QBU, 1HHP, 1QBR, 1QBT

Table C.4: MD simulation protocol for flexible proteins. Parameters for the minimization, relaxation, and production run steps are presented for aldose reductase (AR) simulated at 300 K. The last three lines in the minimization step state how Cartesian constraints are applied to the system. The two lines which describe the weight and restraintmask are presented in Table B.7 for each step. MD simulation protocols for dihydrofolate reductase and HIV-1 protease are similar and differ in the number of constraint atoms.

Steps A-D: minimization with Cartesian constraints

`imin=1, maxcyc=5000, ncyc=4000,`	*invoke minimization*
`ntpr=5,`	*print frequency*
`ntr=1,`	*turn on Cartesian constraints*
`restraint_wt=20.0,`	*weight for Cartesian constraint*
`restraintmask='1-315',`	*atoms in residues 1-103 constrained*

Step E: relaxation with constant volume MD (NVT-MD)

`imin=0, irest=0, ntx=1,`	*invoke MD simulation*
`nstlim=50000, dt=0.002,`	*run for 100 ps*
`ntpr=500, ntwx=500, ntwr=500,`	*output frequency*
`ntt=3, gamma_ln=2.0,`	*temperature control*
`tempi=100, temp0=300,`	*start/end temperature*
`ntb=1, ntp=0, ntc=2, ntf=2,`	*periodic boundary conditions, SHAKE, etc.*
`ig=209858,`	*seed for random number generator*
`nrespa=2,`	*slowly varying forces are evaluated every 2 steps*

Steps F-J: relaxation with constant pressure MD (NPT-MD)

`imin=0, irest=1, ntx=5,`	*restart MD simulation*
`nstlim=50000, dt=0.002,`	*run for 100 ps*
`ntpr=500, ntwx=500, ntwr=500,`	*output frequency*
`ntt=3, gamma_ln=2.0,`	*temperature control*
`temp0=300`	*temperature control*
`ntp=1, taup=1.0,`	*pressure control*
`ntb=2, ntc=2, ntf=2,`	*periodic boundary conditions, SHAKE, etc.*
`nrespa=1,`	*slowly varying forces are evaluated every step*

Step K: production run

`imin=0, irest=1, ntx=5,`	*restart MD*
`nstlim=50000000, dt=0.002,`	*run for 100 ns*
`ntpr=1000, ntwx=1000, ntwr=1000,`	*output frequency*
`ntt=3, gamma_ln=2.0,`	*temperature control*
`temp0=300,`	*temperature control*
`ntp=1, taup=1.0,`	*pressure control*
`ntb=2, ntc=2, ntf=2,`	*periodic boundary conditions, SHAKE, etc.*
`nrespa=1,`	*slowly varying forces are evaluated every step*
`ioutfm=1,`	*output trajectory written in binary format*
`iwrap=1,`	*coordinates are wrapped into a primary box*

Table C.5: Constraints for each step in the MD simulation protocol for flexible proteins. In each step, except step D, J, and K, during our MD simulation protocol constraints are applied on the system. In addition to the `ntr` parameter, which is set to 1, these lines are added to the parameters shown in Table C.4 as illustrated for the first minimization step.

Step A

`restraint_wt=20.0,`	*weight for Cartesian constraint*
`restraintmask='':1-315'',`	*all protein atoms*

Step B

`restraint_wt=20.0,`	*weight for Cartesian constraint*
`restraintmask='!@H='',`	*all atoms except hydrogen atoms*

Step C

`restraint_wt=20.0,`	*weight for Cartesian constraint*
`restraintmask='':1-315@CA,C,O,N'',`	*protein backbone atoms*

Step E

`restraint_wt=2.0,`	*weight for Cartesian constraint*
`restraintmask='':1:315'',`	*all protein atoms*

Step F

`restraint_wt=2.0,`	*weight for Cartesian constraint*
`restraintmask='':1-315@CA,C,O,N'',`	*protein backbone atoms*

Step G

`restraint_wt=1.5,`	*weight for Cartesian constraint*
`restraintmask='':1-315@CA,C,O,N'',`	*protein backbone atoms*

Step H

`restraint_wt=1.0,`	*weight for Cartesian constraint*
`restraintmask='':1-315@CA,C,O,N'',`	*protein backbone atoms*

Step I

`restraint_wt=0.5,`	*weight for Cartesian constraint*
`restraintmask='':1-315@CA,C,O,N'',`	*protein backbone atoms*

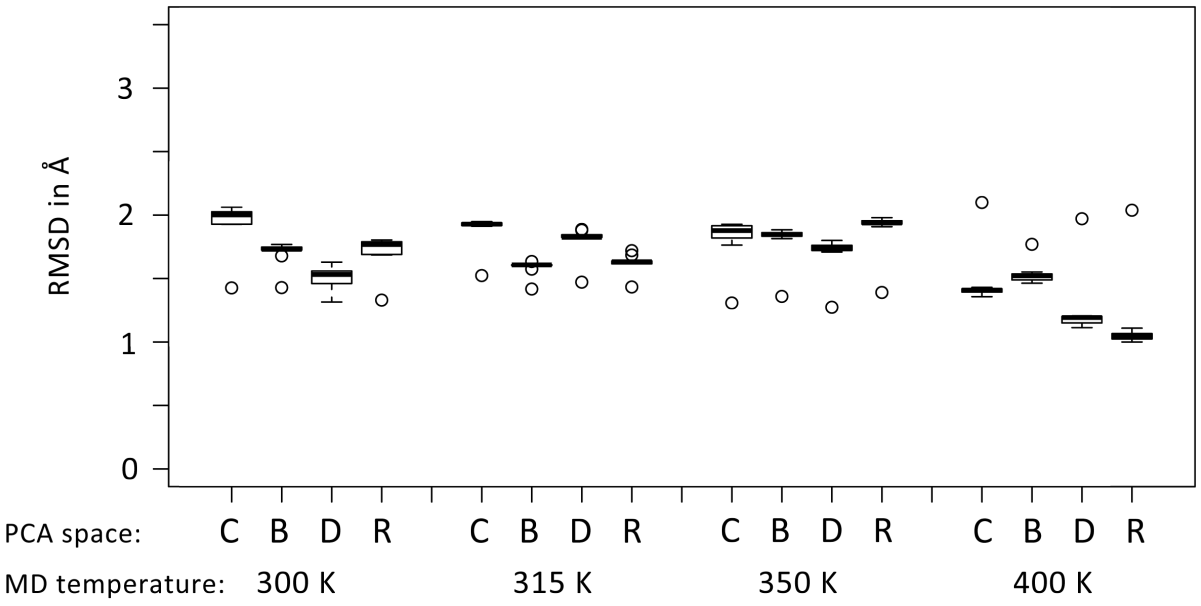


Figure C.1: Structural diversity of DHFR for different MD simulations and PCA. Different DHFR protein conformations were generated with MD at 300, 315, 350, and 400 K and identified by clustering on PCA results performed in C_α (C), backbone (B), backbone dihedral (D), and rotamer (R) space.

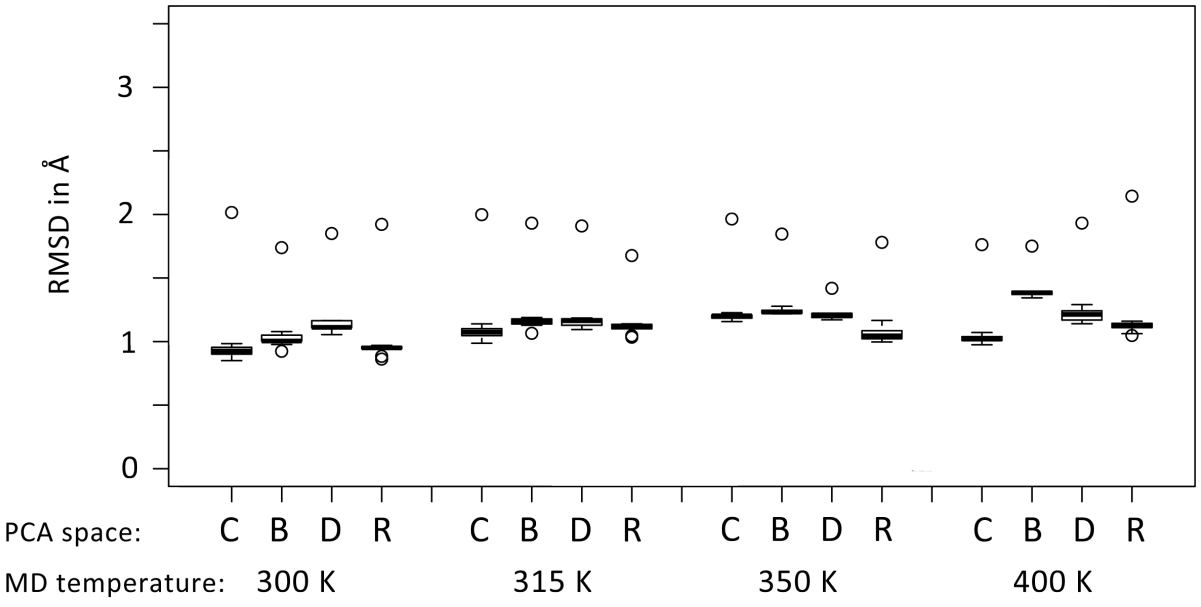


Figure C.2: Structural diversity of "closed" HIV-1 protease for different MD simulations and PCA. Different HIV-1 protease protein conformations were generated with MD at 300, 315, 350, and 400 K and identified by clustering on PCA results performed in C_α (C), backbone (B), backbone dihedral (D), and rotamer (R) space.

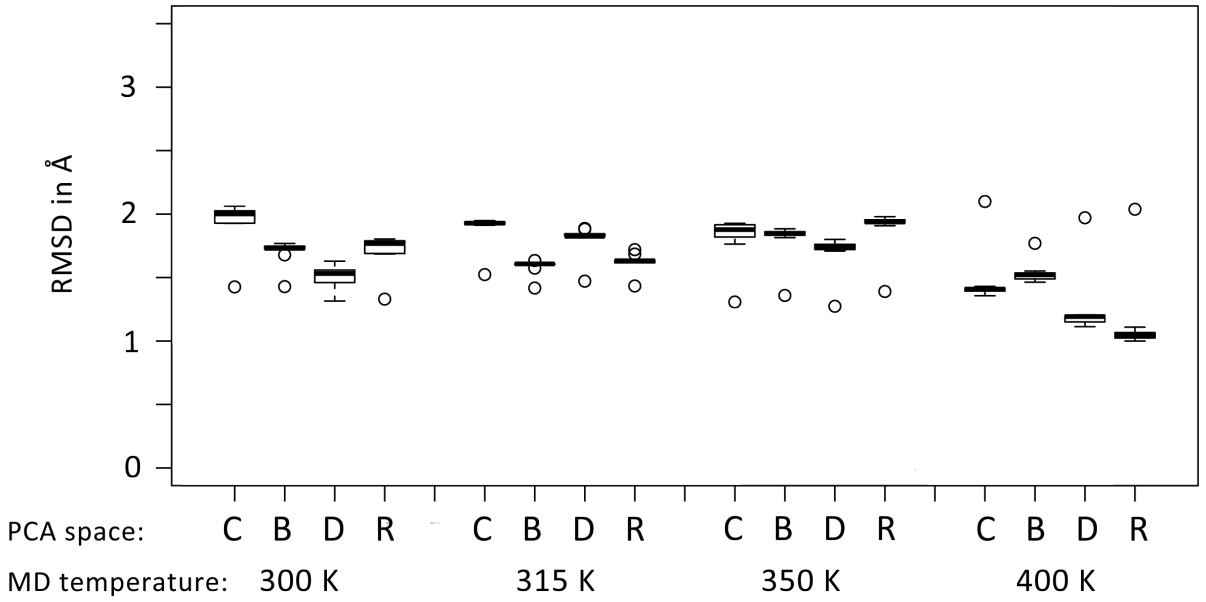


Figure C.3: Structural diversity of "semi-open" HIV-1 protease for different MD simulations and PCA. Different HIV-1 protease protein conformations were generated with MD at 300, 315, 350, and 400 K and identified by clustering on PCA results performed in C_α (C), backbone (B), backbone dihedral (D), and rotamer (R) space.

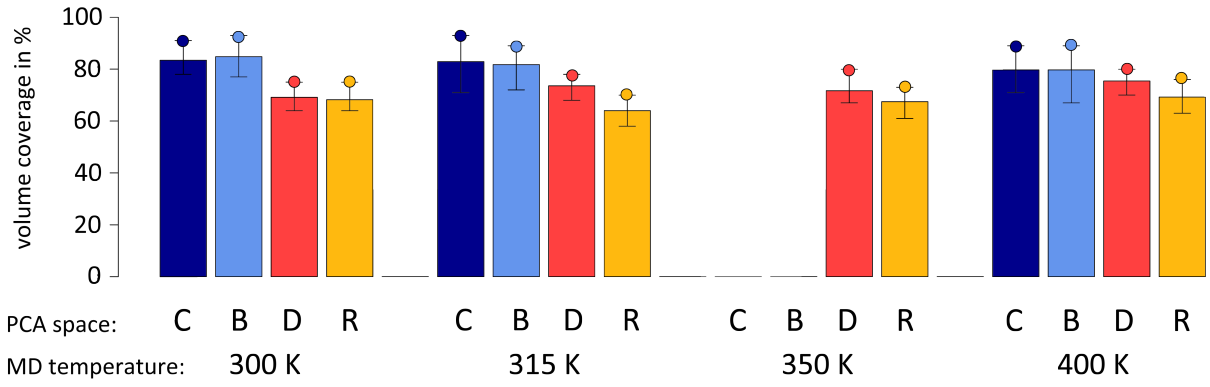


Figure C.4: Conformational space of DHFR covered by cluster representatives. The volume coverage represents the space covered by 150 cluster representatives of the space spanned by all 5,000 protein conformations obtained with MD simulations. The best clustering result for each MD simulation trajectory at 300, 315, 350, and 400 K and PCA performed in C_α (C), backbone (B), dihedral backbone (D), and rotamer (R) space is indicated as a bullet point.

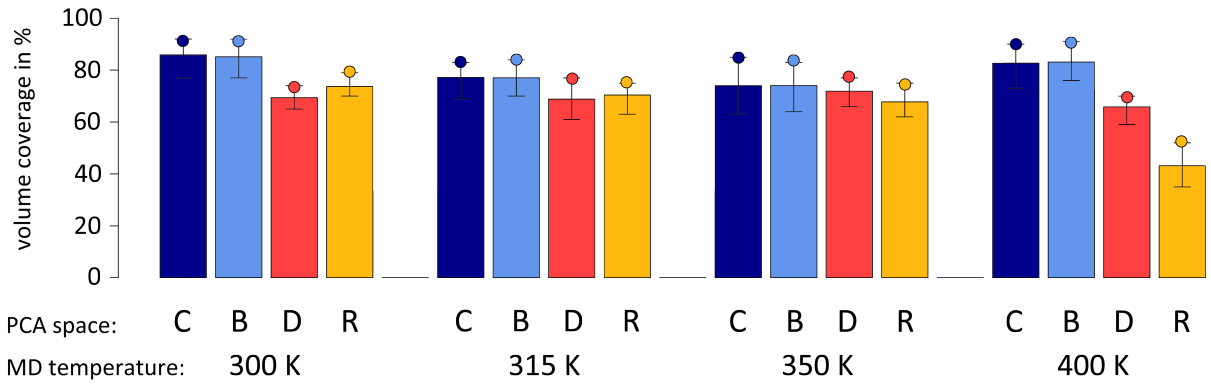


Figure C.5: Conformational space of "closed" HIV-1 protease covered by cluster representatives. The volume coverage represents the space covered by 150 cluster representatives of the space spanned by all 5,000 protein conformations obtained with MD simulations. The best clustering result for each MD simulation trajectory at 300, 315, 350, and 400 K and PCA performed in C_α (C), backbone (B), dihedral backbone (D), and rotamer (R) space is indicated as a bullet point.

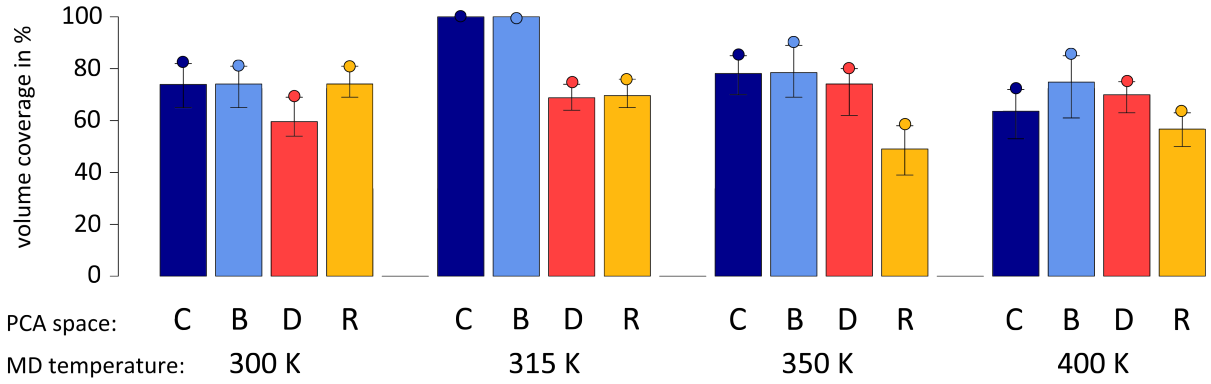


Figure C.6: Conformational space of "semi-open" HIV-1 protease covered by cluster representatives. The volume coverage represents the space covered by 150 cluster representatives of the space spanned by all 5,000 protein conformations obtained with MD simulations. The best clustering result for each MD simulation trajectory at 300, 315, 350, and 400 K and PCA performed in C_α (C), backbone (B), dihedral backbone (D), and rotamer (R) space is indicated as a bullet point.

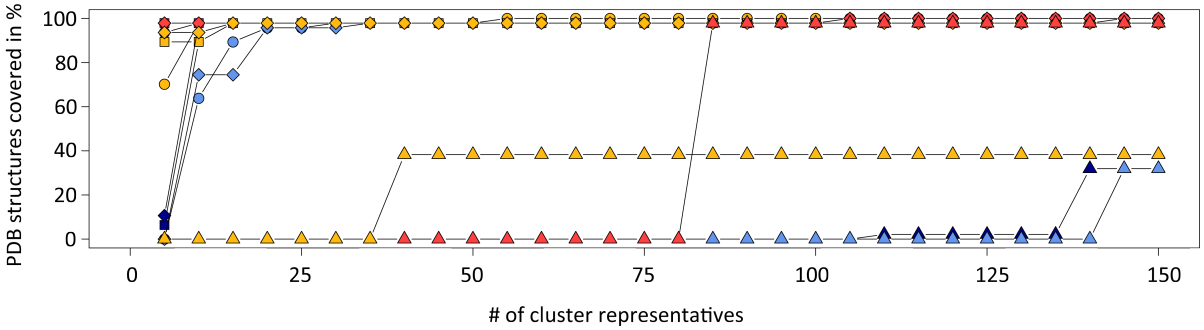


Figure C.7: Optimal number of generated AR structures. C_α and backbone space are colored in dark and light blue and dihedral backbone and rotamer space are colored in red and yellow, respectively. The MD simulation was performed at 300 (circles), 315 (square), 350 (diamond), and 400 K (triangle). Different numbers of representatives for determining the best hierarchical clustering results were tested. The RMSD threshold was set below 1.5 Å to determine whether a PDB structure is similar to a generated protein conformation.

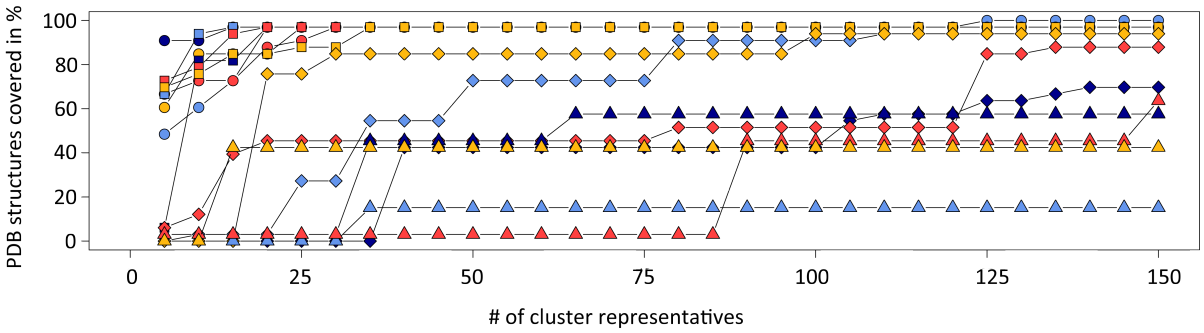


Figure C.8: Optimal number of generated DHFR structures. C_α and backbone space are colored in dark and light blue and dihedral backbone and rotamer space are colored in red and yellow, respectively. The MD simulation was performed at 300 (circles), 315 (square), 350 (diamond), and 400 K (triangle). Different numbers of representatives for determining the best hierarchical clustering results were tested. The RMSD threshold was set below 1.5 Å to determine whether a PDB structure is similar to a generated protein conformation.

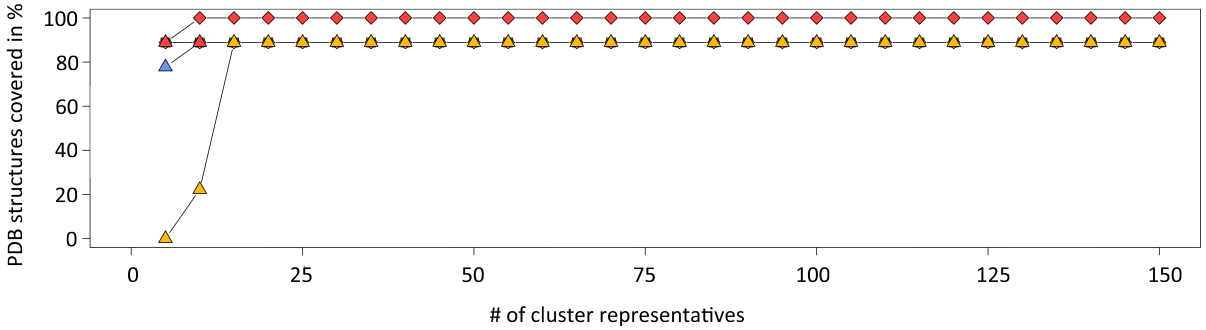


Figure C.9: Optimal number of generated "closed" HIV-1 protease structures. C_α and backbone space are colored in dark and light blue and dihedral backbone and rotamer space are colored in red and yellow, respectively. The MD simulation was performed at 300 (circles), 315 (square), 350 (diamond), and 400 K (triangle). Different numbers of representatives for determining the best hierarchical clustering results were tested. The RMSD threshold was set below 1.5 Å to determine whether a PDB structure is similar to a generated protein conformation.

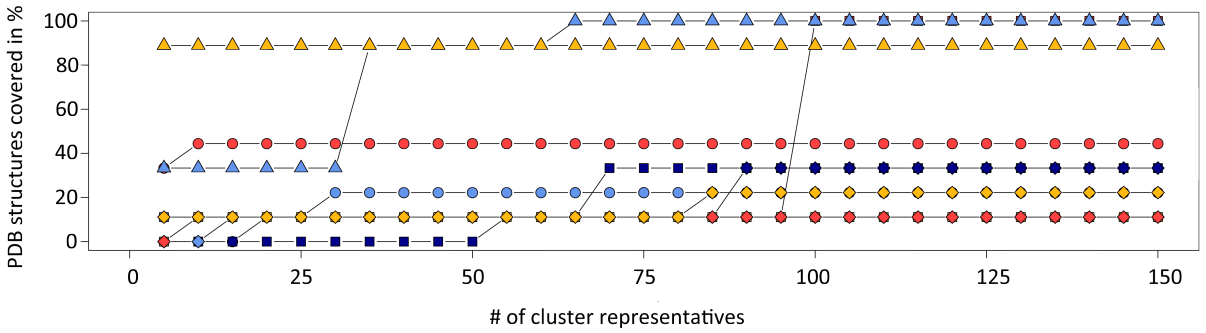


Figure C.10: Optimal number of generated "semi-open" HIV-1 protease structures. C_α and backbone space are colored in dark and light blue and dihedral backbone and rotamer space are colored in red and yellow, respectively. The MD simulation was performed at 300 (circles), 315 (square), 350 (diamond), and 400 K (triangle). Different numbers of representatives for determining the best hierarchical clustering results were tested. The RMSD threshold was set below 1.5 Å to determine whether a PDB structure is similar to a generated protein conformation.

D Third Appendix

Inducing flexibility using molecular dynamics

Providing an MD simulation protocol that induces flexibility to protein structures and additionally accounts for a certain amount of stability is crucial for this study. To be able to perform a stable production run, we introduce several equilibration steps in which constraints on atoms are gradually reduced. As a result we obtain a protein structure that is relaxed. This structure is then used as input for the production run of the MD simulation.

Root-mean-square-deviation (RMSD) values between the relaxed input structure and snapshots along the production run describe the overall dynamics of a protein structure. It is generally assumed that an overall stability is obtained when RMSD values are below 3 Å. This is the case for AR simulated at 300, 315, and 350 K. When simulating AR at 400 K, RMSD values rise continuously during the time interval of 5 to 40 ns. Although, the RMSD values reach sometimes even 8 Å in the second half of the production run (50-100 ns), they always return to 4 Å. No steady increase during this period is observed, which indicates that the protein structure of AR is sufficiently stable despite its motion. As

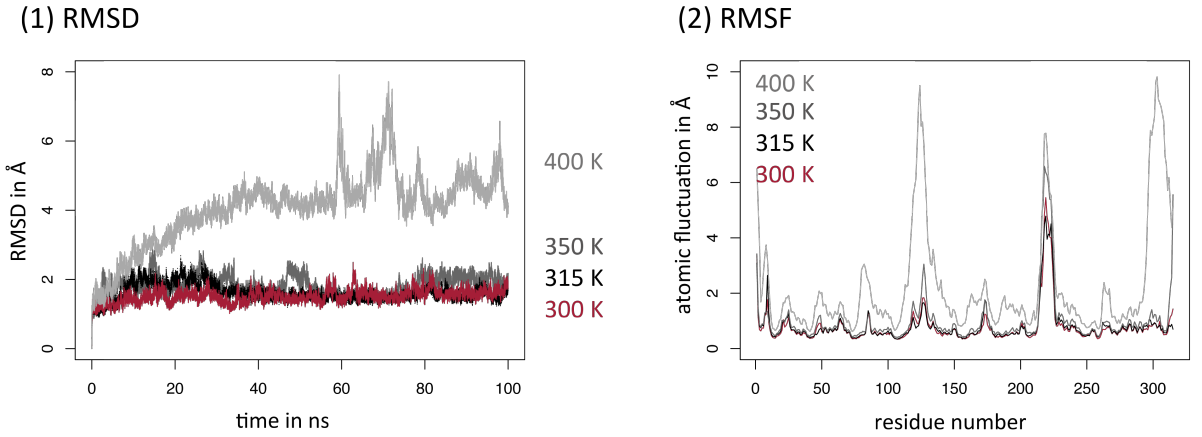


Figure D.1: RMSD and RMSF values of aldose reductase. RMSD C_α atom values during 100 ns simulation of AR (PDB id: 2acr) at four different temperatures: 300, 315, 350, and 400 K (from left to right). The higher the temperature is during the simulation, the more structural dynamics can be observed.

expected, higher temperatures yield larger structural changes. Although RMSD plots indicate overall structural stability or flexibility, they do not define which part of the protein features the highest flexibility. Therefore, we identified flexible protein regions with root-mean-square-fluctuation (RMSF) values. RMSF values are calculated for each residue, in our case for each C_α atom over the entire length of the production run. Higher RMSF values correspond to larger fluctuations of the C_α atom with respect to its initial position. It is again obvious that at higher temperatures the protein is more flexible. Regions which do not show a lot of motion at 300 K, for example residues 110 to 145 of AR, show large structural changes at 400 K.

The N- and C-terminal ends also show high flexibility. Therefore, AR residues 1-5 and 305-315 are excluded in the following to be able to perform reasonable PCA analyses. The dynamics of the termini would add too much noise to the interior structural changes of the protein. The RMSD and RMSF predictions and all following calculations use mapped protein structures as described in the materials and methods section.

Reducing conformational space of flexible proteins

We applied PCA in four different protein representation spaces: two Cartesian coordinate spaces and two dihedral spaces, to reduce the conformational space of the protein. In the following results of the C_α and backbone dihedral space are investigated. In PCA the first principal component accounts for the highest variance. The eigenvalues of the first ten principal components are illustrated in Fig. D.2. Generally, we observe that in C_α space the first principal component accounts for about 70% of the entire variance in all cases. Additionally, almost 100% of the variance can be described by the first

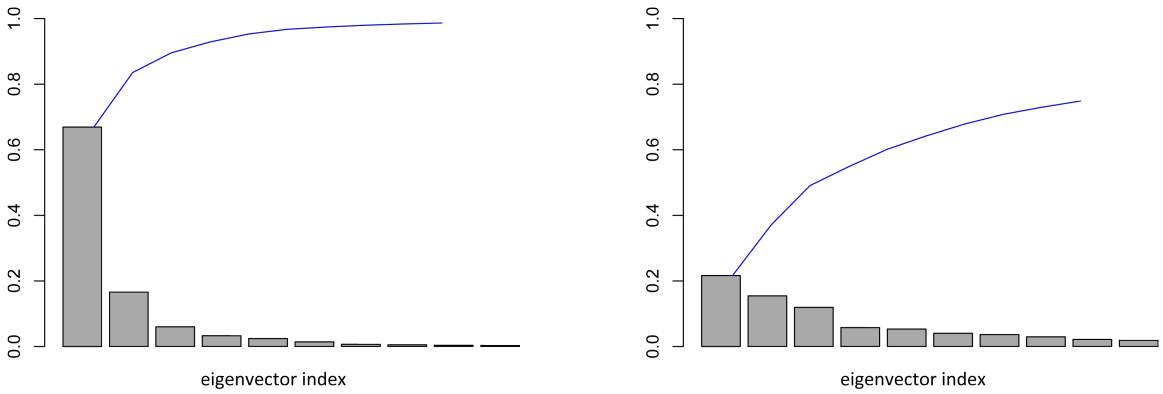


Figure D.2: Variance of principal components. The variance of the first ten principal components is shown for AR in C_α atom space and backbone dihedral space. The blue line indicates the summation of all variance values up to this step. For C_α atom space 95% of the data is described with the first three principal components, whereas in backbone dihedral space only 60% is described.

three principal components, at least in C_α space. Whereas, in dihedral spaces the first principal components do not account as much for the variance. We assume that at 300 K in C_α space the direction of the motion and also residues responsible for these structural changes can be more clearly defined. Whereas, in dihedral spaces more angles contribute to the description of the same structural alteration.

In Fig. D.3 the contribution of each residue to the variance of the first principal component in C_α and dihedral backbone space (Φ and Ψ angles) is illustrated. In C_α space the residues with highest contributing factors are badly congruent with the temperature factors of the high-quality structures, see Fig. 5.3. The two loop regions, (a) and (b), can clearly be identified in C_α space. The highest temperature factors were observed for residues 124-130 (a) and 218-228 (b) for 2acr. For AR the residue numbers are now shifted compared to the residues numbers in Fig. 5.3, since we truncated ten residues at the N-terminal of AR. Therefore, we reveal for the residues with highest contributing factors in C_α space numbers between 110 and 120 for loop (a) and starting at residue 205 up to 215 for loop (b). These regions were also identified as highly flexible by RMSF values shown in Fig. D.1. For the backbone dihedral space the plot looks slightly different. Higher peaks are distributed indicating certain amino acid backbone angles changed at these positions. We can identify specific amino acids

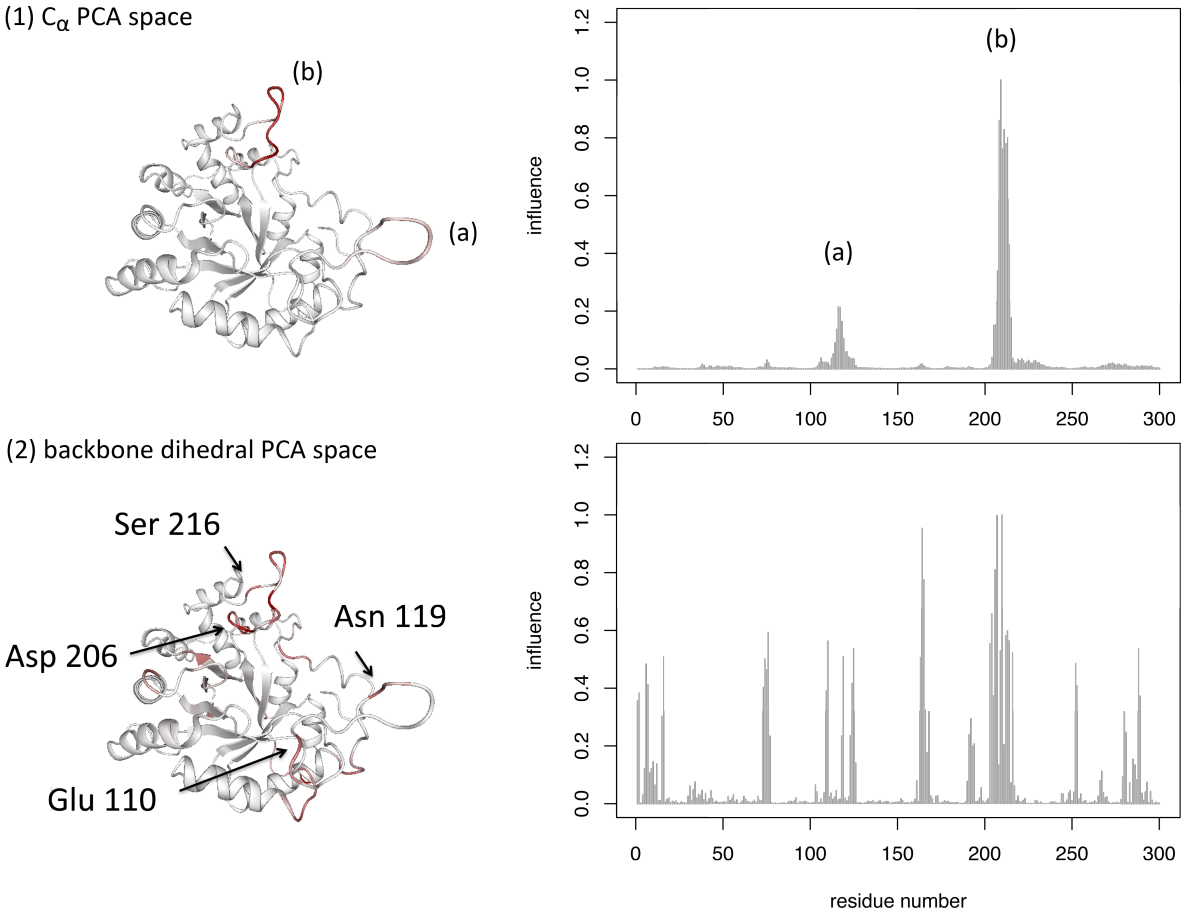


Figure D.3: Structural information of first principal component. AR is colored according to the influence of each residue to the first principal component determined in (1) C_α space and (2) dihedral space.

at both loop hinge regions which determine structural alterations. Glutamine 110 (Glu 120 in the original 2acr structure) and asparagine 119 relate to the motion of loop (a). For loop (b) we identified aspertic acid 206 and serine 216 as present in hinge regions causing loop flexibility. In this loop other residues between positions 206 and 216 show also dihedral backbone changes. This loop region is also known for its flexibility and the region $NADP^+$ binds to (residues 210 to 230 in the origianl 2acr).

Different protein conformations along MD simulation trajectories can be illustrated by projecting all PCA data points onto the first two principal components. Three-dimensional protein structures which are similar with respect to the first two principal components are placed in close proximity. In C_α space the points shown in Fig. D.4 occupy more or less one large region and feature tight and very sharp peaks. Fig. D.4 illustrates clearly that the cluster regions in backbone dihedral space are better separated and more distinct. Thus, in dihedral backbone space more than one high density region is visible. These regions might belong to certain, distinct protein conformations which are generated during the MD simulation.

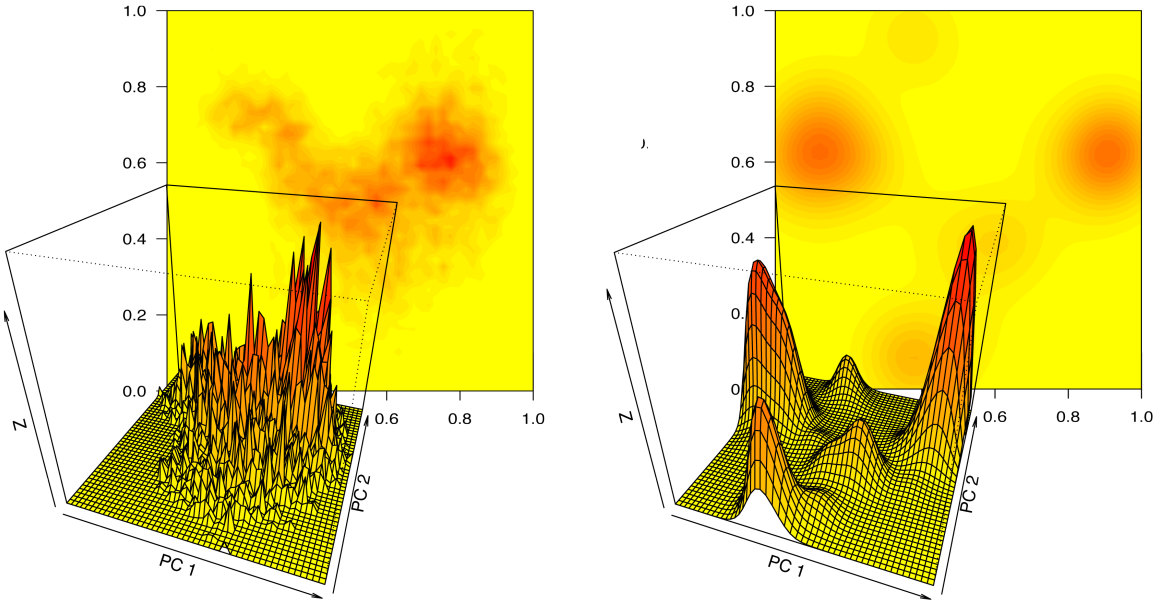


Figure D.4: Density plots of AR. Density plots of the first two principal components for AR are shown. 2acr was simulated at 300 K and PCA was performed in PCA C_α space (left) and in dihedral space (right).

Clustering diverse protein conformations

In order to identify distinct protein conformations we performed a clustering of the PCA results. The best clustering was determined by calculating volumes spanned by the resulting data points in three dimensions. These volumes were compared to the volume spanned by the 5,000 snapshots and intersections between the snapshot volume and cluster volumes were calculated. A value of 100% would indicate that the 150 cluster points span the same convex hull as all original 5,000 snapshot data points. For AR the best clustering results yields always a value above 75%, for all PCA spaces and different MD simulation setups. In general, this number is lower for both dihedral PCA spaces as compared to the equivalent Cartesian space. For example, the intersection of the volume in C_α space (MD simulation at 300 K) yield 83 % and the dihedral backbone space 75 % for the best cases. These two clustering results are illustrated in Fig. D.5. Comparing these plots it becomes evident that the clustering approach selects points present in higher density regions as cluster centers (colored points in Fig. D.5). It can also be observed that the whole space spanned by the 5000 snapshots is almost completely represented by the 150 cluster representatives. The 50 cluster centers alone do not cover that much space and are mostly present in interior regions. Except some snapshot points at the border regions and a few outliers are not captured by the 150 clustering points. This fact is valid for both illustrated PCA spaces: C_α and backbone dihedral space. Selecting the 100 cluster representatives additionally, represents the space spanned by all protein snapshots definitely better than just relying on the 50 k-means cluster centers alone. To be able to compare our results with other methods we produce a smaller subset of the 150 cluster representatives. It is also beneficial for our docking study to reduce the number of structures and eliminate very similar ones. In Fig. D.5 the darker grey dots and the colored ones are located in some cases in close proximity. These probably similar structure conformations are represented by only one in the following.

We performed a hierarchical clustering to obtain only 20 new cluster representatives. In Fig D.6

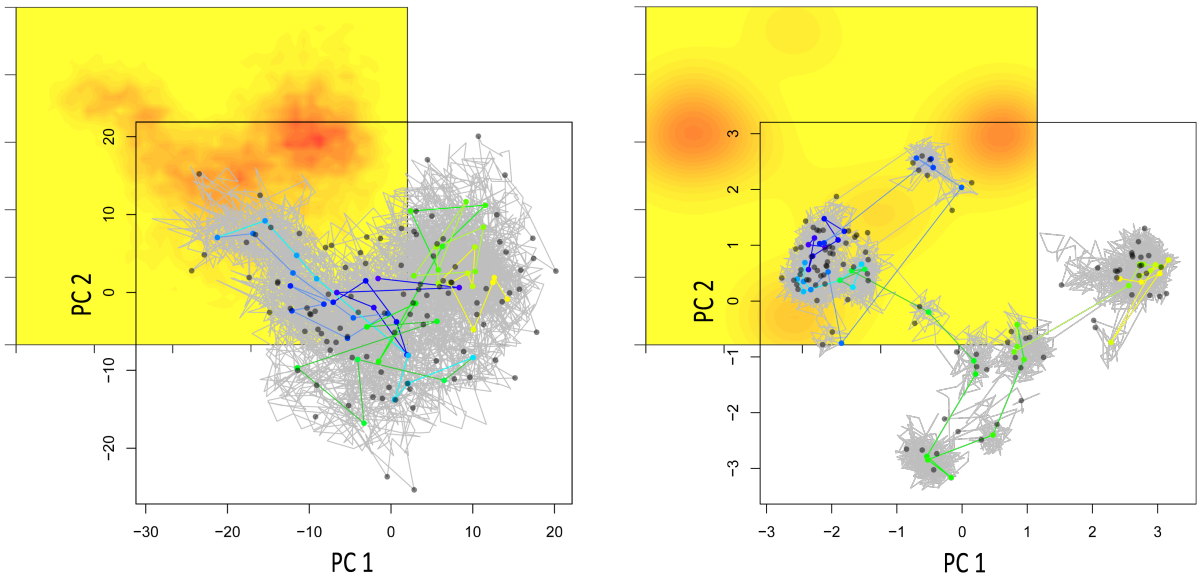


Figure D.5: Movement of MD trajectory structures. Gray lines indicate the movement of the snapshots along the MD simulation. From blue to yellow the 50 cluster representatives calculated with k-means are indicated. Black points represent the 150 cluster representatives.

cluster representatives are shown as black dots for the 20 clusters. The structures with the minimal RMSD to all other structures in the cluster are assigned as new cluster representatives. High density

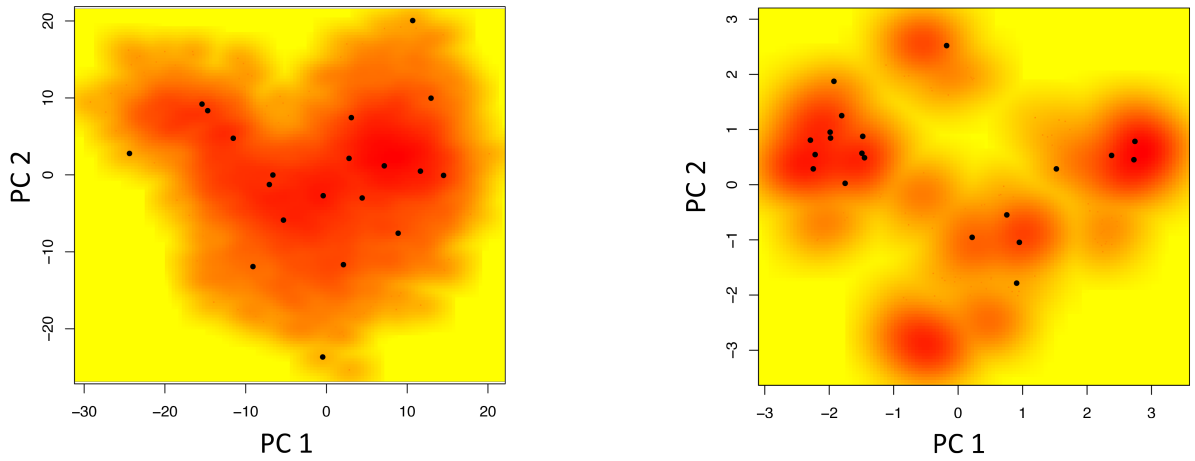


Figure D.6: Density plot with 20 cluster representatives. For AR simulated at 300 K 5000 snapshot data points colored according to their density (red – high density, yellow – low density). The black dots indicate 20 distinct new cluster representative structures in the space spanned by the first two principal components. The left plot shows data in C_α space and the right one in backbone dihedral space.

regions are colored red as opposed to yellow for lower density regions. In C_α space the new cluster representatives cover the space described by all 5000 snapshot data points relatively well. Compared to the 50 cluster centers specified by k-means clustering they even capture some outliers in border regions. Some higher density regions are not described by these 20 new cluster representatives in

backbone dihedral space. Unfortunately, the bottom part of the density plot is scarcely covered in this particular case.

These 20 cluster representatives, however, represent a set of structural conformations that are similar for both PCA spaces (see Fig. D.7). Loop regions (a) and (b) comprise various structures, whereby loop (b) (residues 210-230 in 2acr) shows larger differences. Some conformations differ slightly for PCA analyses in Cartesian and dihedral space, but we could show that both techniques identify similar structures. These conformations are even similar to relevant PDB conformations.

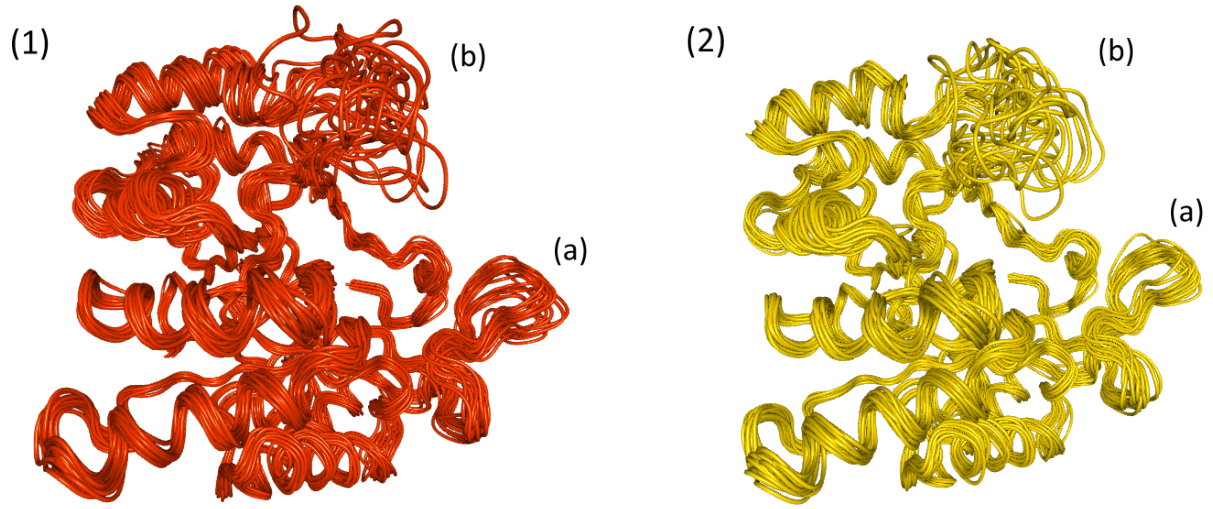


Figure D.7: Generated AR conformations. For AR simulated at 300 K 20 cluster representatives obtained in C_α (red structures) and backbone dihedral (yellow structures) space are superimposed. Loop regions (a) and (b) are flexible and different conformations are identified.

E Fourth Appendix

Table E.1: Rapid MD simulation protocol for re-scoring docking poses. Parameters for the minimization, relaxation, and production run steps in *AMBER* are described. The production run of this rapid implicit solvent MD simulation protocol is only 13 ps long.

Minimization

`imin=1, maxcyc=500, ncyc=500,`	*invoke minimization*
`ntpr=5,`	*print frequency*
`cut=12.0,`	*distance cutoff for non-bonded interactions*
`igb=7, gbsa=1,`	*GB model (igb=7) and SA calculation with LCPO*
`saltcon=0.2,`	*ionic strength*
`intdiel=1.0, extdiel=80.0,`	*interior and exterior dielectric constant*
`ntb=0, ntc=2, ntf=2,`	*no periodic boundary, SHAKE*

Relaxation

`imin=0, irest=0, ntx=1,`	*invoke MD simulation*
`nstlim=3000, dt=0.002,`	*run for 6 ps*
`ntpr=100, ntwx=100, ntwr=100,`	*output frequency*
`cut=12.0,`	*distance cutoff for non-bonded interactions*
`igb=7, gbsa=1,`	*GB model (igb=7) and SA calculation with LCPO*
`saltcon=0.2,`	*ionic strength*
`intdiel=1.0, extdiel=80.0,`	*interior and exterior dielectric constant*
`ntt=3, gamma_ln=2.5,`	*temperature control*
`temp0=0, tempi=300,`	*start/end temperature*
`ntb=0, ntc=2, ntf=2,`	*no periodic boundary, SHAKE*
`nrespa=1,`	*slowly varying forces are evaluated every step*

Production run

`imin=0, irest=1, ntx=5,`	*MD simulation, restart calculation*
`nstlim=6500, dt=0.002,`	*run for 13 ps*
`ntpr=100, ntwx=100, ntwr=100,`	*output frequency*
`cut=12.0,`	*distance cutoff for non-bonded interactions*
`igb=7, gbsa=1,`	*GB model (igb=7) and SA calculation with LCPO*
`saltcon=0.2,`	*ionic strength*
`intdiel=1.0, extdiel=80.0,`	*interior and exterior dielectric constant*
`ntt=3, gamma_ln=2.5,`	*temperature control*
`temp0=300.0, tempi=300,`	*start/end temperature*
`ntb=0, ntc=2, ntf=2,`	*no periodic boundary, SHAKE*
`nrespa=1,`	*slowly varying forces are evaluated every step*

Table E.2: Parameters for the PB section. We determined binding free energies of 18 urokinase-ligand complexes with the MM-PBSA approach provided by *AMBER* using the parameters described in this table.

PROC	2	*use Amber pbsa program*
REFE	0	*reference state for PB calculation*
INDI	1.0	*interior dielectric constant*
EXDI	80.0	*solvent dielectric constant*
SCALE	1.428571	*lattice spacing in number of grids per Å*
LINIT	1000	*number of interactions with the linear PB equation*
PRBRAD	0.0	*solvent probe radius in Å*
ISTRING	200.0	*ionic strength in mM for PB solver*
RADIOPT	0	*use radii from prmtop files for PB calculation*
NPOPT	1	*restart MD*
SURFTEN	0.0072	*surface tension*
SURFOFF	0.00	*offset to correct non-polar contribution*

Table E.3: Parameters for the GB section. We determined binding free energies of 18 urokinase-ligand complexes with the MM-PBSA approach provided by *AMBER* using the parameters described in this table.

IGB	7	*use GB model (igb=7)*
GBSA	1	*perform gbsa calculation*
INTDIEL	1.0	*interior dielectric constant*
EXTDIEL	80.0	*solvent dielectric constant*
SALTCON	0.2	*ionic strength in M for GB*
RADIOPT	0	*use radii from prmtop files for GB calculation*
SURFTEN	0.0072	*surface tension*
SURFOFF	0.00	*offset to correct non-polar contribution*

Table E.4: Binding free energies without flexibility. Binding free energies for all 18 urokinase inhibitors determined experimentally, using the docking program *Glide*, and four different force field-based approaches are presented. The force field-based methods (MM-PBSA and MM-GBSA) were tested with two different force fields: ff99SB and ff03. Pearson and Spearman correlations between experimental and calculated binding free energies are calculated.

				ff99SB				ff03			
		Glide		PBSA		GBSA		PBSA		GBSA	
Ligand	-pK$_i$	scores	rank	ΔG	rank	ΔG	rank	ΔG	rank	ΔG	rank
1	-7.3	-8.20	8	-62.77	1	-30.83	2	-60.68	2	-30.42	3
2	-7.2	-7.62	11	-62.45	2	-23.38	9	-62.36	1	-24.14	9
3	-6.5	-9.01	2	-49.39	5	-22.23	12	-49.04	5	-21.94	12
4	-6.2	-7.73	10	-46.08	10	-26.13	8	-46.08	10	-26.01	8
5	-6.0	-7.98	9	-48.44	8	-27.74	6	-48.45	7	-28.27	6
6	-5.8	-8.60	6	-40.70	13	-22.64	10	-40.57	13	-23.04	11
7	-5.7	-8.46	7	-48.16	9	-22.26	11	-47.76	8	-23.09	10
8	-5.6	-8.76	4	-51.99	4	-30.12	4	-52.13	4	-30.84	2
9	-5.5	-9.73	1	-57.20	3	-32.01	1	-56.43	3	-32.30	1
10	-5.2	-7.52	12	-48.88	7	-29.55	5	-49.03	6	-29.51	4
11	-5.1	-9.01	3	-41.03	12	-16.83	14	-40.59	12	-16.29	14
12	-5.0	-5.63	14	-36.13	15	-8.36	17	-36.24	15	-9.95	17
13	-4.9	-8.73	5	-45.58	11	-26.62	7	-45.28	11	-26.73	7
14	-4.7	-7.12	13	-49.27	6	-30.63	3	-47.53	9	-29.24	5
15	-4.5	-5.01	15	-37.15	14	-14.59	15	-37.04	14	-15.68	15
16	-4.3	-4.36	18	-27.54	17	-14.59	16	-27.66	17	-15.50	16
17	-4.3	-4.46	16	-30.10	16	-17.01	13	-30.66	16	-18.29	13
18	-4.0	-4.37	17	-25.16	18	-3.70	18	-24.68	18	-3.69	18
R =		0.63	0.58	0.83	0.75	0.54	0.49	0.83	0.79	0.55	0.50

Table E.5: Binding free energies with flexibility. Binding free energies for all 18 urokinase inhibitors determined experimentally, using the docking program *Glide*, and four different force field-based approaches are presented. The force field-based methods (MM-PBSA and MM-GBSA) were tested with two different force fields: ff99SB and ff03. Pearson and Spearman correlations between experimental and calculated binding free energies are calculated.

				ff99SB				ff03			
		Glide		PBSA		GBSA		PBSA		GBSA	
Ligand	-pK$_i$	scores	rank	ΔG	rank	ΔG	rank	ΔG	rank	ΔG	rank
1	-7.3	-8.20	8	-81.66	1	-64.66	1	-78.53	1	-62.56	1
2	-7.2	-7.62	11	-65.20	2	-42.46	3	-66.40	2	-49.97	2
3	-6.5	-9.01	2	-57.27	4	-42.99	2	-60.25	3	-48.68	3
4	-6.2	-7.73	10	-54.55	5	-41.19	5	-50.48	10	-40.87	8
5	-6.0	-7.98	9	-53.08	7	-39.04	7	-51.64	9	-35.47	13
6	-5.8	-8.60	6	-47.58	10	-36.26	9	-46.66	13	-37.19	12
7	-5.7	-8.46	7	-54.44	6	-42.14	4	-51.82	8	-40.90	7
8	-5.6	-8.76	4	-49.21	8	-37.70	8	-54.06	5	-43.91	5
9	-5.5	-9.73	1	-58.97	3	-40.49	6	-59.26	4	-45.14	4
10	-5.2	-7.52	12	-46.27	11	-34.08	11	-53.74	6	-43.76	6
11	-5.1	-9.01	3	-48.22	9	-32.20	12	-51.83	7	-40.59	9
12	-5.0	-5.63	14	-42.10	14	-26.34	15	-42.50	14	-27.08	15
13	-4.9	-8.73	5	-42.11	13	-30.56	13	-47.29	11	-38.53	10
14	-4.7	-7.12	13	-44.98	12	-35.87	10	-46.8	12	-37.36	11
15	-4.5	-5.01	15	-41.91	15	-30.37	14	-41.85	15	-28.06	14
16	-4.3	-4.36	18	-28.71	18	-19.64	18	-27.33	18	-18.35	17
17	-4.3	-4.46	16	-33.38	16	-23.09	16	-27.56	17	-17.23	18
18	-4.0	-4.37	17	-30.91	17	-19.88	17	-30.64	16	-19.59	16
R =		0.63	0.58	0.92	0.92	0.88	0.93	0.89	0.81	0.86	0.80

F Contributions and Publications

Contributions

Chapter 2 - Biological Background

This chapter is part of a manuscript [82] that has been accepted at *Bentham eBooks*. Nina Monika Fischer (NMF) conceived the manuscript. In addition to myself (NMF) Oliver Kohlbacher (OK) contributed to this manuscript.

Chapter 4 - Studying protein-DNA complexes at the atomic level

This project is part of a manuscript [30] that has been accepted at *Nucleic Acids Research*. It is joint work with Luise H. Brand (LHB), Dierk Wanke (DW), Klaus Harter (KH), and Oliver Kohlbacher (OK). DW, LHB, and NMF conceived and designed the project. The contributions to this project of LHB, DW, and KH have not been incorporated in detail in this chapter. LHB and DW conducted *in vitro* experiments presented in [30]. LHB provided the DNA sequence logos of Fig. 4.10 created with Weblogo version 3.0 [58]. NMF performed the computational experiments, NMF and OK analyzed the computational results.

Chapter 5 - Studying flexibility of protein-ligand complexes

Section 5.2 - Representing major movements in protein-ligand docking

In addition to myself (NMF), Mirco Michel (MM), Christopher Mohr (CM), and Oliver Kohlbacher (OK) contributed to this project. NMF and OK conceived and designed the project. MM implemented the PCA analysis tool. CM tested and compared different clustering algorithms. NMF performed the computational experiments presented in this chapter. NMF and OK analyzed the data.

Section 5.3 - Rapid molecular dynamics simulation protocol for re-scoring docked ligand poses

In addition to myself (NMF), Wolfgang M. Schneider (WMS), Andreas Kämper (AK), and Oliver Kohlbacher (OK) contributed to this project. NMF and OK conceived and designed this project. AK assisted with obtaining ligand protonation states. NMF developed the high-throughput MD simulation protocol. WMS performed the docking experiment, conducted MD simulations, and tested different MM-PBSA parameter settings. NMF, WMS, and OK contributed to the discussion. This project was presented as poster at the German Conference on Chemoinformatics in 2010 and is included in the conference proceedings [83].

Figures

All figures illustrating protein, DNA, or ligand structure representations are visualized using BALLView [202, 203].

Publications

Accepted and published manuscripts and abstracts

- Brand, L.H.*, **Fischer, N.M.***, Harter, K., Wanke, D., Kohlbacher, O.
  Elucidating the evolutionary conserved DNA-binding specificities of WRKY transcription factors by molecular dynamics and in vitro binding assays. (accepted at Nucleic Acids Research) 2013.
  *shared corresponding authors
    - Text and figures from this manuscript appear in the Abstract and in Chapter 2 and 4 of this thesis.

- **Fischer, N.M.**, Kohlbacher, O.
  Structural insights into physical interactions of transcription factor-DNA complexes at an atomic level. (accepted to eBook "The Analysis of Cis-elements: Current developments, knowledge and applications to gene networking", edited by K. Berendzen, Bentham Science Publisher) 2013.
    - Text and figures from this manuscript appear in Sections 2.1 and 2.3 and in Chapter 4 of this thesis.

- Schärfe, C., Taeger, J., Reuter, P., **Fischer, N.M.**, Krüger, J., Thiel, P., Wissinger, B., Kohlbacher, O.
  Development of a pharmacophore model for pharmacological chaperones targeting mutant trafficking deficient CNG channels. J Cheminform. 2013; 5(Suppl 1): O18. Published online 2013.

- Wanke, D., Brand, L.H., **Fischer, N.M.**, Peschke, F., Kilian, J., and Berendzen K.W.
  Implications of DNA-nanostructures by Hoogsteen-dinucleotides on transcription factor binding. In QP-PQ: Quantum Probability and White Noise Analysis - Vol. 30, pp. 351-362. "Quantum Bio-Informatics V: From Quantum Information to Bio-Informatics". World Scientific (ISBN: 978-981-4460-01-9) 2013.

- Schumann, M., Röttig, M., **Fischer, N.M.**, Kohlbacher, O.
  A framework and workflow system for virtual screening and molecular docking. J Cheminform. 2011; 3(Suppl 1):P15. Published online 2011.

- **Fischer, N.M.**, Schneider, W.M, Kohlbacher, O.
  Rescoring of docking poses using force field-based methods. J Cheminform. 2010; 2(Suppl 1):P52. Published online 2010.
    - The concept and results of this project appear in Section 5.3 of this thesis.

Bibliography

1. Adler, K., Beyreuther, K., Fanning, E., Geisler, N., Gronenborn, B., Klemm, A., Müller-Hill, B., Pfahl, M., and Schmitz, A. How lac repressor binds to DNA. *Nature*, 237(5354):322–327, 1972.

2. Alamanova, D., Stegmaier, P., and Kel, A. Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC Bioinformatics*, 11:225, 2010.

3. Alibés, A., Nadra, A. D., Masi, F. D., Bulyk, M. L., Serrano, L., and Stricher, F. Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: the Pax6 example. *Nucleic Acids Res*, 38(21):7422–7431, 2010.

4. AlQuraishi, M. and McAdams, H. H. Direct inference of protein-DNA interactions using compressed sensing methods. *Proc Natl Acad Sci USA*, 108(36):14819–14824, 2011.

5. Amaro, R. E., Baron, R., and McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J Comput Aided Mol Des*, 22(9):693–705, 2008.

6. Anandakrishnan, R. and Onufriev, A. Analysis of basic clustering algorithms for numerical estimation of statistical averages in biomolecules. *J Comput Biol*, 15(2):165–184, 2008.

7. Anderson, A. C., O'Neil, R. H., Surti, T. S., and Stroud, R. M. Approaches to solving the rigid receptor problem by identifying a minimal set of flexible residues during ligand docking. *Chem Biol*, 8(5):445–457, 2001.

8. Angarica, V. E., Pérez, A. G., Vasconcelos, A. T., Collado-Vides, J., and Contreras-Moreira, B. Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, 9:436, 2008.

9. Antosiewicz, J., McCammon, J. A., and Gilson, M. K. Prediction of pH-dependent properties of proteins. *J Mol Biol*, 238(3):415–436, 1994.

10. Antosiewicz, J., McCammon, J. A., and Gilson, M. K. The determinants of pKas in proteins. *Biochemistry*, 35 (24):7819–7833, 1996.

11. Babu, M. M., Iyer, L. M., Balaji, S., and Aravind, L. The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. *Nucleic Acids Res*, 34(22):6505–6520, 2006.

12. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci USA*, 98(18):10037–10041, 2001.

13. Bank, R. E. and Holst, M. A New Paradigm for Parallel Adaptive Meshing Algorithms. *SIAM Review*, 54:291–232, 2006.

14. Beamer, L. J. and Pabo, C. O. Refined 1.8 Å crystal structure of the lambda repressor-operator complex. *J Mol Biol*, 227(1):177–196, 1992.

15. Beierlein, F. R., Kneale, G. G., and Clark, T. Predicting the effects of basepair mutations in DNA-protein complexes by thermodynamic integration. *Biophys J*, 101(5):1130–1138, 2011.

16. Benos, P. V., Bulyk, M. L., and Stormo, G. D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, 30(20):4442–4451, 2002.

17. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. Molecular dynamics with coupling to an external bath. *J Chem Phys*, 81:3684–3690, 1984.

18. Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24 (11):1429–1435, 2006.

19. Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R., and Schneider, B. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J*, 63(3):751–759, 1992.

20. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res*, 28:235–242, 2000.

21. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., and Zardecki, C. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 6 No 1):899–907, 2002.

22. Berman, H. M., Westbrook, J., Feng, Z., Iype, L., Schneider, B., and Zardecki, C. The Nucleic Acid Database. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 6 No 1):889–898, 2002.

23. Betts, M. J. and Sternberg, M. J. An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng*, 12(4):271–283, 1999.

24. Bezdek, J. C. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers Norwell, MA, USA, 1981.

25. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 31(1):365–370, 2003.

26. Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des*, 8(3):243–256, 1994.

27. Böhm, H. J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des*, 12(4):309–323, 1998.

28. Bondi, A. van der Waals Volumes and Radii. *J Chem Phys*, 64:441, 1964.

29. Bonnet, P. and Bryce, R. A. Molecular dynamics and free energy analysis of neuraminidase-ligand interactions. *Protein Sci*, 13(4):946–957, 2004.

30. Brand, L., Fischer, N., Harter, K., Wanke, D., and Kohlbacher, O. Elucidating the evolutionary conserved DNA-binding specificities of WRKY transcription factors by molecular dynamics and in vitro binding assays. Accepted at Nucleic Acids Research, 2013.

31. Brand, L. H., Kirchler, T., Hummel, S., Chaban, C., and Wanke, D. DPI-ELISA: a fast and versatile method to specify the binding of plant transcription factors to DNA in vitro. *Plant Methods*, 6:25, 2010.

32. Broughton, H. B. A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening. *J Mol Graph Model*, 18(3):247–57, 302–4, 2000.

33. Brown, S. P. and Muchmore, S. W. High-throughput calculation of protein-ligand binding affinities: modification and adaptation of the MM-PBSA protocol to enterprise grid computing. *J Chem Inf Model*, 46(3):999–1005, 2006.

34. Brown, S. P. and Muchmore, S. W. Rapid estimation of relative protein-ligand binding affinities using a high-throughput version of MM-PBSA. *J Chem Inf Model*, 47(4):1493–1503, 2007.

35. Bulyk, M. L., Huang, X., Choo, Y., and Church, G. M. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci USA*, 98(13):7158–7163, 2001.

36. Caflisch, A. and Karplus, M. Acid and thermal denaturation of barnase investigated by molecular dynamics simulations. *J Mol Biol*, 252(5):672–708, 1995.

37. Carlson, H. A., Masukawa, K. M., Rubins, K., Bushman, F. D., Jorgensen, W. L., Lins, R. D., Briggs, J. M., and McCammon, J. A. Developing a dynamic pharmacophore model for HIV-1 integrase. *J Med Chem*, 43(11): 2100–2114, 2000.

38. Case, D. A., Darden, T. A., Cheatham, T. E., III, C. L. S., J. Wang, R. E. D., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Roberts, B., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossvai, I., Wong, K. F., Paesani, F., Vanicek, J., Liu, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.-J., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P. A. *AMBER 10*. University of California, San Francisco, 2008.

39. Case, D. A., Darden, T. A., Cheatham, T. E., III, C. L. S., J. Wang, R. E. D., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Roberts, B., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossvai, I., Wong, K. F., Paesani, F., Vanicek, J., Liu, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.-J., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P. A. *AMBER 11*. University of California, San Francisco, 2010.

40. Cavasotto, C. N. and Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J Mol Biol*, 337(1):209–225, 2004.

41. Cavasotto, C. N., Kovacs, J. A., and Abagyan, R. A. Representing receptor flexibility in ligand docking through relevant normal modes. *J Am Chem Soc*, 127(26):9632–9640, 2005.

42. Cheatham, T. E. and Kollman, P. A. Molecular dynamics simulations highlight the structural differences among DNA:DNA, RNA:RNA, and DNA:RNA hybrid duplexes. *J Am Chem Soc*, 119:4805–4825, 1997.

43. Chen, C. and Chen, Z. Isolation and characterization of two pathogen- and salicylic acid-induced genes encoding WRKY DNA-binding proteins from tobacco. *Plant Mol Biol*, 42(2):387–396, 2000.

44. Chen, Q. K., Hertz, G. Z., and Stormo, G. D. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci*, 11(5):563–566, 1995.

45. Chen, S., Gunasekera, A., Zhang, X., Kunkel, T. A., Ebright, R. H., and Berman, H. M. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. *J Mol Biol*, 314(1):75–82, 2001.

46. Chen, Y., Dey, R., and Chen, L. Crystal structure of the p53 core domain bound to a full consensus site as a self-assembled tetramer. *Structure*, 18(2):246–256, 2010.

47. Cherrak, I., Mauffret, O., Santamaria, F., Hocquet, A., Ghomi, M., Rayner, B., and Fermandjian, S. L-nucleotides and 8-methylguanine of d(C1m8G2C3G4C5LG6LC7G8C9G10)2 act cooperatively to promote a left-handed helix under physiological salt conditions. *Nucleic Acids Res*, 31(23):6986–6995, 2003.

48. Chervenak, M. C. and Toone, E. J. A direct measure of the contribution of solvent reorganization to the enthalpy of ligand binding. *J Am Chem Soc*, 116:10533–10539, 1994.

49. Chiu, S. W., Clark, M., Subramaniam, S., and Jakobsson, E. J. Collective motion artifacts arising in long-duration molecular dynamics simulations. *J Comput Chem*, 21:121–131, 2000.

50. Ciolkowski, I., Wanke, D., Birkenbihl, R. P., and Somssich, I. E. Studies on DNA-binding selectivity of WRKY transcription factors lend structural clues into WRKY-domain function. *Plant Mol Biol*, 68(1-2):81–92, 2008.

51. Claussen, H., Buning, C., Rarey, M., and Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol*, 308(2):377–395, 2001.

52. Cody, V., Luft, J. R., Pangborn, W., and Gangjee, A. Analysis of three crystal structure determinations of a 5-methyl-6-N-methylanilino pyridopyrimidine antifolate complex with human dihydrofolate reductase. *Acta Crystallogr D Biol Crystallogr*, 59(Pt 9):1603–1609, 2003.

53. Cohen, S. X., Moulin, M., Hashemolhosseini, S., Kilian, K., Wegner, M., and Müller, C. W. Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. *EMBO J*, 22(8):1835–1845, 2003.

54. Coifman, R. R. and Maggioni, M. Diffusion wavelets. *Appl Comput Harmon Anal*, 21:53–94, 2006.

55. Connolly, M. L. Analytical molecular surface calculation. *J Appl Cryst*, 16(5):548–558, 1983.

56. Cormack, R. S., Eulgem, T., Rushton, P. J., Köchner, P., Hahlbrock, K., and Somssich, I. E. Leucine zipper-containing WRKY proteins widen the spectrum of immediate early elicitor-induced WRKY transcription factors in parsley. *Biochim Biophys Acta*, 1576(1-2):92–100, 2002.

57. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, J. D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollmann, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc*, 117:5179–5197, 1995.

58. Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190, 2004.

59. Dahlquist, F. W. Remembering Daniel E. Koshland Jr. (1920âĂŞ2007). *Protein Sci*, 12:2583âĂŞ–2584, 2007.

60. Damm, K. L. and Carlson, H. A. Exploring experimental sources of multiple protein conformations in structure-based drug design. *J Am Chem Soc*, 129(26):8225–8235, 2007.

61. Darden, T., Pearlman, D., and Pedersen, L. G. Ionic charging free energies: Spherical versus periodic boundary conditions. *J Chem Phys*, 79:926–935, 1983.

62. Das, P., Moll, M., Stamati, H., Kavraki, L. E., and Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc Natl Acad Sci USA*, 103(26):9885–9890, 2006.

63. Davis, I. W. and Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J Mol Biol*, 385(2): 381–392, 2009.

64. de Pater, S., Greco, V., Pham, K., Memelink, J., and Kijne, J. Characterization of a zinc-dependent transcriptional activator from Arabidopsis. *Nucleic Acids Res*, 24(23):4624–4631, 1996.

65. de Vries, S. J., van Dijk, A. D. J., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A. M. J. J. HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins*, 69(4):726–733, 2007.

66. Deng, Y., Glimm, J., Wang, Y., Korobka, A., Eisenberg, M., and Grollman, A. P. Prediction of protein binding to DNA in the presence of water-mediated hydrogen bonds. *J Molecular Modeling*, 5(7):125–133, 1999.

67. Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125(7):1731–1737, 2003.

68. Donald, J. E., Chen, W. W., and Shakhnovich, E. I. Energetics of protein-DNA interactions. *Nucleic Acids Res*, 35(4):1039–1047, 2007.

69. Dong, F., Vijayakumar, M., and Zhou, H.-X. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. *Biophys J*, 85(1):49–60, 2003.

70. Dror, R. O., Jensen, M. Ø., Borhani, D. W., and Shaw, D. E. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *J Gen Physiol*, 135(6):555–562, 2010.

71. Duan, M.-R., Nan, J., Liang, Y.-H., Mao, P., Lu, L., Li, L., Wei, C., Lai, L., Li, Y., and Su, X.-D. DNA binding mechanism revealed by high resolution crystal structure of Arabidopsis thaliana WRKY1 protein. *Nucleic Acids Res*, 35(4):1145–1154, 2007.

72. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem*, 24(16):1999–2012, 2003.

73. Dunitz, J. D. Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions. *Chem Biol*, 2(11):709–712, 1995.

74. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des*, 11(5):425–445, 1997.

75. Elrod-Erickson, M., Rould, M. A., Nekludova, L., and Pabo, C. O. Zif268 protein-DNA complex refined at 1.6 Å: a model system for understanding zinc finger-DNA interactions. *Structure*, 4(10):1171–1180, 1996.

76. Endres, R. G., Schulthess, T. C., and Wingreen, N. S. Toward an atomistic model for predicting transcription-factor binding sites. *Proteins*, 57(2):262–268, 2004.

77. Eulgem, T. and Somssich, I. E. Networks of WRKY transcription factors in defense signaling. *Curr Opin Plant Biol*, 10(4):366–371, 2007.

78. Eulgem, T., Rushton, P. J., Robatzek, S., and Somssich, I. E. The WRKY superfamily of plant transcription factors. *Trends Plant Sci*, 5(5):199–206, 2000.

79. Ferguson, A. L., Panagiotopoulos, A. Z., Kevrekidis, I. G., and Debenedetti, P. G. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chem Phys Lett*, 509:1–11, 2011.

80. Fermi, G., Perutz, M. F., Shaanan, B., and Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol*, 175(2):159–174, 1984.

81. Fischer, E. Einfluss der Konfiguration auf die Wirkung der Enzyme. *Berichte der deutschen chemische Gesellschaft*, 27:2985–2993, 1894.

82. Fischer, N. M. and Kohlbacher, O. Structural insights into physical interactions of transcription factor-DNA complexes at an atomic level. Accepted to eBook "The Analysis of Cis-elements: Current developments, knowledge and applications to gene networking" edited by Kenneth Berendzen, Bentham Science Publisher, 2013.

83. Fischer, N. M., Schneider, W. M., and Kohlbacher, O. Rescoring of docking poses using force field-based methods. *J. Cheminform*, 2(Suppl 1):P52, 2010.

84. Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E. Z., Thorpe, M., and Gerstein, M. The Database of Macromolecular Motions: new features added at the decade mark. *Nucleic Acids Res*, 34(Database issue):D296–D301, 2006.

85. Fogolari, F., Esposito, G., Viglino, P., and Molinari, H. Molecular mechanics and dynamics of biomolecules using a solvent continuum model. *J Comput Chem*, 22(15):1830–1842, 2001.

86. Fogolari, F., Brigo, A., and Molinari, H. Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys J*, 85(1):159–166, 2003.

87. Forgy, E. Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, 21: 768–769, 1965.

88. Frech, K., Herrmann, G., and Werner, T. Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res*, 21(7):1655–1664, 1993.

89. Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*, 47(7):1739–1749, 2004.

90. Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C., and Mainz, D. T. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*, 49(21):6177–6196, 2006.

91. Frith, M. C., Hansen, U., and Weng, Z. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878–889, 2001.

92. Fry, C. J. and Farnham, P. J. Context-dependent transcriptional regulation. *J Biol Chem*, 274(42):29583–29586, 1999.

93. García, A. E. Large-Amplitude Non-Linear Motions in Proteins. *Phys. Rev. Lett.*, 68:2696–2700, 1992.

94. Ge, W., Schneider, B., and Olson, W. K. Knowledge-based elastic potentials for docking drugs or proteins with nucleic acids. *Biophys J*, 88(2):1166–1190, 2005.

95. Gerstein, M. and Krebs, W. A database of macromolecular motions. *Nucleic Acids Res*, 26(18):4280–4290, 1998.

96. Gilbert, W. and Maxam, A. The nucleotide sequence of the lac operator. *Proc Natl Acad Sci USA*, 70(12): 3581–3584, 1973.

97. Gilbert, W. and Müller-Hill, B. The lac operator is DNA. *Proc Natl Acad Sci USA*, 58(6):2415–2421, 1967.

98. Gohlke, H. and Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J Comput Chem*, 25(2):238–250, 2004.

99. Gohlke, H. and Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem Int Ed Engl*, 41(15):2644–2676, 2002.

100. Gohlke, H., Kiel, C., and Case, D. A. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J Mol Biol*, 330(4):891–913, 2003.

101. Gordon, J. C., Myers, J. B., Folta, T., Shoja, V., Heath, L. S., and Onufriev, A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res*, 33(Web Server issue):W368–W371, 2005.

102. Graves, A. P., Shivakumar, D. M., Boyce, S. E., Jacobson, M. P., Case, D. A., and Shoichet, B. K. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *J Mol Biol*, 377(3):914–934, 2008.

103. Guo, J. X. and Gmeiner, W. H. Molecular dynamics simulation of the human U2B" protein complex with U2 snRNA hairpin IV in aqueous solution. *Biophys J*, 81(2):630–642, 2001.

104. Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., and Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem*, 47(7):1750–1759, 2004.

105. Harrison, D. H., Bohren, K. M., Ringe, D., Petsko, G. A., and Gabbay, K. H. An anion binding site in human aldose reductase: mechanistic implications for the binding of citrate, cacodylate, and glucose 6-phosphate. *Biochemistry*, 33(8):2011–2020, 1994.

106. Hartigan, J. A. and Wong, M. A. A k-means clustering algorithm. *Applied Statistics*, 28:100 – 108, 1979.

107. Harvey, S. C., Tan, R. K. Z., and Cheatham, T. E. The flying ice cube: velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J Comput Chem*, 18:726–740, 1998.

108. Havranek, J. J., Duarte, C. M., and Baker, D. A simple physical model for the prediction and design of protein-DNA interactions. *J Mol Biol*, 344(1):59–70, 2004.

109. Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chem Phys Lett*, 246:122–129, 1995.

110. Hawkins, G. D., Cramer, C. J., and Truhlar, D. G. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J Phys Chem*, 100:19824–19839, 1995.

111. Hegde, R. S. The papillomavirus E2 proteins: structure, function, and biology. *Annu Rev Biophys Biomol Struct*, 31:343–360, 2002.

112. Hert, S. and Schirra, S. 3D Convex Hulls. In CGAL User and Reference Manual. CGAL Editorial Board, 4.2 Edition, 2013.

113. Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput*, 4:435–447, 2008.

114. Hildebrandt, A., Dehof, A. K., Rurainski, A., Bertsch, A., Schumann, M., Toussaint, N. C., Moll, A., Stöckel, D., Nickels, S., Mueller, S. C., Lenhof, H.-P., and Kohlbacher, O. BALL–biochemical algorithms library 1.3. *BMC Bioinformatics*, 11:531, 2010.

115. Hinderhofer, K. and Zentgraf, U. Identification of a transcription factor specifically expressed at the onset of leaf senescence. *Planta*, 213(3):469–473, 2001.

116. Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D., and Shakked, Z. DNA bending by an adenine–thymine tract and its role in gene regulation. *Proc Natl Acad Sci USA*, 98(15):8490–8495, 2001.

117. Hockney, R. W. Potential calculation and some applications. *Methods Comput Phys*, 1970.

118. Höglund, A. and Kohlbacher, O. From sequence to structure and back again: approaches for predicting protein-DNA binding. *Proteome Sci*, 2(1):3, 2004.

119. Holm, L. and Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res*, 38(Web Server issue): W545–W549, 2010.

120. Honig, B. and Rohs, R. Biophysics: Flipping Watson and Crick. *Nature*, 470(7335):472–473, 2011.

121. Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. Errors in protein structures. *Nature*, 381(6580):272, 1996.

122. Hornak, V. and Simmerling, C. Targeting structural flexibility in HIV-1 protease inhibitor binding. *Drug Discov Today*, 12(3-4):132–138, 2007.

123. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–725, 2006.

124. Hornak, V., Okur, A., Rizzo, R. C., and Simmerling, C. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc Natl Acad Sci USA*, 103(4):915–920, 2006.

125. Howard, E. I., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Zandt, M. V., Sibley, E., Bon, C., Moras, D., Schneider, T. R., Joachimiak, A., and Podjarny, A. Ultrahigh resolution drug design I: details of interactions in human aldose reductase-inhibitor complex at 0.66 Å. *Proteins*, 55(4):792–804, 2004.

126. Huo, S., Wang, J., Cieplak, P., Kollman, P. A., and Kuntz, I. D. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. *J Med Chem*, 45(7):1412–1419, 2002.

127. Hwang, M. J., Stockfisch, T. P., and Hagler, A. T. Derivation of Class II Force Fields. 2. Derivation and Characterisation of a Class II Force Field, CFF93, for the Alkyl Functional Group and Alkane Molecules. *J Am Chem Soc*, 116:2515–2525, 1994.

128. Ichiye, T. and Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, 11(3):205–217, 1991.

129. Jacobson, M. P., Friesner, R. A., Xiang, Z., and Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol*, 320(3):597–608, 2002.

130. Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., and Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins*, 55(2):351–367, 2004.

131. Jakalian, A., Jack, D. B., and Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J Comput Chem*, 23(16):1623–1641, 2002.

132. Janin, J. Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure*, 7(12):R277–R279, 1999.

133. Jayaram, B. and Jain, T. The role of water in protein-DNA recognition. *Annu Rev Biophys Biomol Struct*, 33: 343–361, 2004.

134. Jiang, F. and Kim, S. H. "Soft docking": matching of molecular surface cubes. *J Mol Biol*, 219(1):79–102, 1991.

135. Johnson, C. S., Kolevski, B., and Smyth, D. R. TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. *Plant Cell*, 14(6):1359–1375, 2002.

136. Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267(3):727–748, 1997.

137. Jones, S., van Heyningen, P., Berman, H. M., and Thornton, J. M. Protein-DNA interactions: A structural analysis. *J Mol Biol*, 287(5):877–896, 1999.

138. Jorgensen, W. L., Chandrasekhar, J., Madura, J., and Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys*, 79:926–935, 1983.

139. Joshi, R., Passner, J. M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M. A., Jacob, V., Aggarwal, A. K., Honig, B., and Mann, R. S. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, 131(3):530–543, 2007.

140. Journot-Catalino, N., Somssich, I. E., Roby, D., and Kroj, T. The transcription factors WRKY11 and WRKY17 act as negative regulators of basal resistance in Arabidopsis thaliana. *Plant Cell*, 18(11):3289–3302, 2006.

141. Kaufmann, L. and Rousseeuw, P. J. Clustering by means of medoids. *Statistical Data Analysis based on the L Norm*, 405 – 445, 1987.

142. Kendrew, J. C. The three-dimensional structure of a protein molecule. *Sci Am*, 205:96–110, 1961.

143. Kendrew, J. C. The structure of globular proteins. *Comp Biochem Physiol*, 4:249–252, 1962.

144. Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillps, D. C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

145. Kendrew, J. C., Watson, H. C., Strandberg, B. E., Dickerson, R. E., Philipps, D. C., and Shore, V. C. The amino-acid sequence x-ray methods, and its correlation with chemical data. *Nature*, 190:666–670, 1961.

146. Kerzmann, A., Neumann, D., and Kohlbacher, O. SLICK–scoring and energy functions for protein-carbohydrate interactions. *J Chem Inf Model*, 46(4):1635–1642, 2006.

147. Keserû, G. M. and Kolossváry, I. Fully flexible low-mode docking: application to induced fit in HIV integrase. *J Am Chem Soc*, 123(50):12708–12709, 2001.

148. Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*, 26(12):1351–1359, 2008.

149. Kinoshita, T., Miyake, H., Fujii, T., Takakura, S., and Goto, T. The structure of human recombinant aldose reductase complexed with the potent inhibitor zenarestat. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 4):622–626, 2002.

150. Klapper, I., Hagstrom, R., Fine, R., Sharp, K., and Honig, B. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. *Proteins*, 1(1):47–59, 1986.

151. Klebe, G. and Böhm, H. J. Energetic and entropic factors determining binding affinity in protein-ligand complexes. *J Recept Signal Transduct Res*, 17(1-3):459–473, 1997.

152. Knegtel, R. M., Kuntz, I. D., and Oshiro, C. M. Molecular docking to ensembles of protein structures. *J Mol Biol*, 266(2):424–440, 1997.

153. Kollman, P. A., Weiner, S., Seibel, G., Lybrand, T., Singh, U. C., Caldwell, J., and Rao, S. N. Modeling complex molecular interactions involving proteins and DNA. *Ann N Y Acad Sci*, 482:234–244, 1986.

154. Kono, H. and Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, 35(1): 114–131, 1999.

155. Koo, H. S., Wu, H. M., and Crothers, D. M. DNA bending at adenine . thymine tracts. *Nature*, 320(6062): 501–506, 1986.

156. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci USA*, 44(2):98–104, 1958.

157. Koudelka, G. B. and Carlson, P. DNA twisting and the effects of non-contacted bases on affinity of 434 operator for 434 repressor. *Nature*, 355(6355):89–91, 1992.

158. Koudelka, G. B., Harrison, S. C., and Ptashne, M. Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature*, 326(6116):886–888, 1987.

159. Kuhn, B. and Kollman, P. A. Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J Med Chem*, 43(20):3786–3791, 2000.

160. Kuhn, B., Gerber, P., Schulz-Gasch, T., and Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *J Med Chem*, 48(12):4040–4048, 2005.

161. Lam, P. Y., Jadhav, P. K., Eyermann, C. J., Hodge, C. N., Ru, Y., Bacheler, L. T., Meek, J. L., Otto, M. J., Rayner, M. M., and Wong, Y. N. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science*, 263(5145):380–384, 1994.

162. Lamoureux, J. S., Stuart, D., Tsang, R., Wu, C., and Glover, J. N. M. Structure of the sporulation-specific transcription factor Ndt80 bound to DNA. *EMBO J*, 21(21):5721–5732, 2002.

163. Lamoureux, J. S., Maynes, J. T., and Glover, J. N. M. Recognition of 5'-YpG-3' sequences by coupled stacking/hydrogen bonding interactions with amino acid residues. *J Mol Biol*, 335(2):399–408, 2004.

164. Lawson, C. L., Swigon, D., Murakami, K. S., Darst, S. A., Berman, H. M., and Ebright, R. H. Catabolite activator protein: DNA binding and transcription activation. *Curr Opin Struct Biol*, 14(1):10–20, 2004.

165. Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol*, 235(1):345–356, 1994.

166. Leach, A. R. *Molecular Modeling: Principles and Application*. Pearson Education Limited, 2001.

167. Lee, M. C. and Duan, Y. Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins*, 55(3):620–634, 2004.

168. Lee, M.-L. T., Bulyk, M. L., Whitmore, G. A., and Church, G. M. A statistical model for investigating binding probabilities of DNA nucleotide sequences using microarrays. *Biometrics*, 58(4):981–988, 2002.

169. Li, J., Brader, G., Kariola, T., and Palva, E. T. WRKY70 modulates the selection of signaling pathways in plant defense. *Plant J*, 46(3):477–491, 2006.

170. Lin, J.-H., Perryman, A. L., Schames, J. R., and McCammon, J. A. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J Am Chem Soc*, 124(20):5632–5633, 2002.

171. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78(8):1950–1958, 2010.

172. Linse, S. S. Scientific Background on the Nobel Prize in Chemistry 2012. Studies of G-protein-coupled receptors. The Royal Swedish Academy of Sciences, 2012.

173. Liu, L. A. and Bader, J. S. Ab initio prediction of transcription factor binding sites. *Pac Symp Biocomput*, pages 484–495, 2007.

174. Liu, L. A. and Bradley, P. Atomistic modeling of protein-DNA interaction specificity: progress and applications. *Curr Opin Struct Biol*, 22(4):397–405, 2012.

175. Lloyd, S. P. Least squares quantization in pcm. *IEEE Transations on Information Theory*, 28:128 – 137, 1982.

176. Lu, B. Z., Chen, W. Z., Wang, C. X., and Xu, X.-J. Protein molecular dynamics with electrostatic force entirely determined by a single Poisson-Boltzmann calculation. *Proteins*, 48(3):497–504, 2002.

177. Lu, Q. and Luo, R. A Poisson–Boltzmann dynamics method with nonperiodic boundary condition. *J. Chem. Phys.*, 119:11035–11047, 2003 2003.

178. Luo, M., Dennis, E. S., Berger, F., Peacock, W. J., and Chaudhury, A. MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in Arabidopsis. *Proc Natl Acad Sci USA*, 102(48):17531–17536, 2005.

179. Luo, R., David, L., and Gilson, M. K. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J Comput Chem*, 23(13):1244–1253, 2002.

180. Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biol*, 1(1):REVIEWS001, 2000.

181. Luscombe, N. M., Laskowski, R. A., and Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*, 29(13):2860–2874, 2001.

182. Lyne, P. D., Lamb, M. L., and Saeh, J. C. Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J Med Chem*, 49(16):4805–4808, 2006.

183. Ma, P. C., Rould, M. A., Weintraub, H., and Pabo, C. O. Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. *Cell*, 77(3):451–459, 1994.

184. Maerkl, S. J. and Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.

185. Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J Mol Biol*, 253(2):370–382, 1995.

186. Mandel-Gutfreund, Y., Baron, A., and Margalit, H. A structure-based approach for prediction of protein binding sites in gene upstream regions. *Pac Symp Biocomput*, pages 139–150, 2001.

187. Mangelsen, E., Kilian, J., Berendzen, K. W., Kolukisaoglu, U. H., Harter, K., Jansson, C., and Wanke, D. Phylogenetic and comparative gene expression analysis of barley (Hordeum vulgare) WRKY transcription factor family reveals putatively retained functions between monocots and dicots. *BMC Genomics*, 9:194, 2008.

188. Mangoni, M., Roccatano, D., and Nola, A. D. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins*, 35(2):153–162, 1999.

189. Matthews, B. W. Protein-DNA interaction. No code for recognition. *Nature*, 335(6188):294–295, 1988.

190. Maxwell, J. C. Molecules. *Nature*, 1873.

191. May, A. and Zacharias, M. Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. *J Med Chem*, 51(12):3499–3506, 2008.

192. McQueen, J. Some methods for classifciation and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281 – 297, 1967.

193. Mehler, E. L. and Eichele, G. Electrostatic effects in water-accessible regions of proteins. *Biochemistry*, 23(17): 3887–3891, 1984.

194. Meiler, J. and Baker, D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins*, 65(3):538–548, 2006.

195. Melo, F. and Feytmans, E. Novel knowledge-based mean force potential at atomic level. *J Mol Biol*, 267(1): 207–222, 1997.

196. Melo, F. and Feytmans, E. Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol*, 277(5):1141–1152, 1998.

197. Melo, F., Devos, D., Depiereux, E., and Feytmans, E. ANOLEA: a www server to assess protein structures. *Proc Int Conf Intell Syst Mol Biol*, 5:187–190, 1997.

198. M.Holst. Adaptive Numerical Treatment of Elliptic Systems on Manifolds. *Advances in Computational Mathematics*, 15:139–191, 2001.

199. Misra, V. K., Hecht, J. L., Sharp, K. A., Friedman, R. A., and Honig, B. Salt effects on protein-DNA interactions. The lambda cI repressor and EcoRI endonuclease. *J Mol Biol*, 238(2):264–280, 1994.

200. Miyamoto, S. and Kollman, P. A. SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem*, 13:952–962, 1992.

201. Mohr, C. Flexibility in Proteins – Clustering Molecular Dynamics Trajectories. Master's thesis, University of Tübingen, 2012.

202. Moll, A., Hildebrandt, A., Lenhof, H.-P., and Kohlbacher, O. BALLView: an object-oriented molecular visualization and modeling framework. *J Comput Aided Mol Des*, 19(11):791–800, 2005.

203. Moll, A., Hildebrandt, A., Lenhof, H.-P., and Kohlbacher, O. BALLView: a tool for research and education in molecular modeling. *Bioinformatics*, 22(3):365–366, 2006.

204. Mongan, J., Simmerling, C., McCammon, J. A., Case, D. A., and Onufriev, A. Generalized Born model with a simple, robust molecular volume correction. *J Chem Theory Comput*, 3(1):156–169, 2007.

205. Moroni, E., Caselle, M., and Fogolari, F. Identification of DNA-binding protein target sequences by physical effective energy functions: free energy analysis of lambda repressor-DNA complexes. *BMC Struct Biol*, 7:61, 2007.

206. Morozov, A. V., Havranek, J. J., Baker, D., and Siggia, E. D. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res*, 33(18):5781–5798, 2005.

207. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*, 30(16):2785–2791, 2009.

208. Muybridge, E. The Horse in motion. "Sally Gardner," owned by Leland Stanford; running at a 1:40 gait over the Palo Alto track. *Library of Congress Prints and Photographs Division Washington, D.C.*, 1878.

209. Muybridge, E. *Animal Locomotion: an Electro-Photographic Investigation of Connective Phases of Animal Movements*. J.B. Lippincott Co., 1887.

210. Muybridge, E. *âĂIJThe GallopâĂİ, Animals in Motion*. New York: Dover Publications, 1957.

211. Nelson, H. C., Finch, J. T., Luisi, B. F., and Klug, A. The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature*, 330(6145):221–226, 1987.

212. Ng, A. Y., Jordan, M. I., and Weiss, Y. *Advances in Neural Information Processing Systems - On spectral clustering: Analysis and an algorithm*. MIT Press, 2001.

213. Nicholls, A. and Honig, B. A Rapid Finite Difference Algorithm, Utilizing Successive Over-Relaxation to Solve the Poisson-Boltzmann Equation. *J Comput Chem*, 12:435–445, 1991.

214. Nikolova, E. N., Kim, E., Wise, A. A., O'Brien, P. J., Andricioaei, I., and Al-Hashimi, H. M. Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*, 470(7335):498–502, 2011.

215. Nikolova, E. N., Gottardo, F. L., and Al-Hashimi, H. M. Probing transient Hoogsteen hydrogen bonds in canonical duplex DNA using NMR relaxation dispersion and single-atom substitution. *J Am Chem Soc*, 134(8):3667–3670, 2012.

216. Noskov, S. Y. and Lim, C. Free energy decomposition of protein-protein interactions. *Biophys J*, 81(2):737–750, 2001.

217. O'Boyle, N. M., Liebeschuetz, J. W., and Cole, J. C. Testing assumptions and hypotheses for rescoring success in protein-ligand docking. *J Chem Inf Model*, 49(8):1871–1878, 2009.

218. Ohishi, H., Tsukamoto, K., Hiyama, Y., Maezaki, N., Tanaka, T., and Ishida, T. Amine free crystal structure: the crystal structure of d(CGCGCG)2 and methylamine complex crystal. *Biochem Biophys Res Commun*, 348(3): 794–798, 2006.

219. Okimoto, N., Futatsugi, N., Fuji, H., Suenaga, A., Morimoto, G., Yanai, R., Ohno, Y., Narumi, T., and Taiji, M. High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. *PLoS Comput Biol*, 5(10):e1000528, 2009.

220. Onufriev, A., Bashford, D., and Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J Phys Chem B*, 104:3712–3720, 2000.

221. Onufriev, A., Case, D. A., and Bashford, D. Effective Born radii in the generalized Born approximation: the importance of being perfect. *J Comput Chem*, 23(14):1297–1304, 2002.

222. Onufriev, A., Bashford, D., and Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*, 55(2):383–394, 2004.

223. Osterberg, F., Morris, G. M., Sanner, M. F., Olson, A. J., and Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins*, 46 (1):34–40, 2002.

224. Pabo, C. O. and Nekludova, L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J Mol Biol*, 301(3):597–624, 2000.

225. Pabo, C. O. and Sauer, R. T. Protein-DNA recognition. *Annu Rev Biochem*, 53:293–321, 1984.

226. Pabo, C. O. and Sauer, R. T. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem*, 61:1053–1095, 1992.

227. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, 2009.

228. Pearson, W. R., Wood, T., Zhang, Z., and Miller, W. Comparison of DNA Sequences with Protein Sequences. *Genomics*, 46(1):24–36, 1997.

229. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A., and Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys J*, 92(11): 3817–3829, 2007.

230. Perryman, A. L., Lin, J.-H., and McCammon, J. A. Optimization and computational evaluation of a series of potential active site inhibitors of the V82F/I84V drug-resistant mutant of HIV-1 protease: an application of the relaxed complex method of structure-based drug design. *Chem Biol Drug Des*, 67(5):336–345, 2006.

231. Perutz, M. Early days of protein crystallography. *Methods Enzymol*, 114:3–18, 1985.

232. M. F.Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. *Nature*, 185 (4711):416–422, 1960.

233. Pietrucci, F., Marinelli, F., Carloni, P., and Laio, A. Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *J Am Chem Soc*, 131(33):11811–11818, 2009.

234. Pruitt, K. D. and Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*, 29 (1):137–140, 2001.

235. Ptashne, M. Specific binding of the lambda phage repressor to lambda DNA. *Nature*, 214(85):232–234, 1967.

236. Raha, K. and Merz, K. M. Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J Med Chem*, 48(14): 4558–4575, 2005.

237. Rahi, S. J., Virnau, P., Mirny, L. A., and Kardar, M. Predicting transcription factor specificity with all-atom models. *Nucleic Acids Res*, 36(19):6209–6217, 2008.

238. Rarey, M., Kramer, B., and Lengauer, T. Time-efficient docking of flexible ligands into active sites of proteins. *Proc Int Conf Intell Syst Mol Biol*, 3:300–308, 1995.

239. Rashin, A. A. Hydration Phenomena, Classical Electrostatics, and the Boundary Element Method. *J Phys Chem*, 94:1725–1733, 1990.

240. Raub, S., Steffen, A., Kämper, A., and Maria, C. M. AIScore chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes. *J Chem Inf Model*, 48(7):1492–1510, 2008.

241. Reyes, C. M. and Kollman, P. A. Structure and thermodynamics of RNA-protein binding: using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change. *J Mol Biol*, 297(5):1145–1158, 2000.

242. Rizzo, R. C., Toba, S., and Kuntz, I. D. A molecular basis for the selectivity of thiadiazole urea inhibitors with stromelysin-1 and gelatinase-A from generalized born molecular dynamics simulations. *J Med Chem*, 47(12): 3065–3074, 2004.

243. Rocchia, W., Alexov, E., and Honig, B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *J Phys Chem B*, 105:6507–6514, 2001.

244. Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., and Honig, B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem*, 23(1):128–137, 2002.

245. Rognan, D., Lauemoller, S. L., Holm, A., Buus, S., and Tschinke, V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem*, 42(22):4650–4658, 1999.

246. Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248–1253, 2009.

247. Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*, 79:233–269, 2010.

248. Rushton, P. J., Torres, J. T., Parniske, M., Wernert, P., Hahlbrock, K., and Somssich, I. E. Interaction of elicitor-induced DNA-binding proteins with elicitor response elements in the promoters of parsley PR1 genes. *EMBO J*, 15(20):5690–5700, 1996.

249. Rushton, P. J., Somssich, I. E., Ringler, P., and Shen, Q. J. WRKY transcription factors. *Trends Plant Sci*, 15 (5):247–258, 2010.

250. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J Comput Physics*, 23:327–341, 1977.

251. Ryde, U. Molecular dynamics simulations of alcohol dehydrogenase with a four- or five-coordinate catalytic zinc ion. *Proteins*, 21(1):40–56, 1995.

252. Sadiq, S. K. and De Fabritiis, G. Explicit solvent dynamics and energetics of HIV-1 protease flap opening and closing. *Proteins*, 78(14):2873–2885, 2010.

253. Schneider, B. and Berman, H. M. Hydration of the DNA bases is local. *Biophys J*, 69(6):2661–2669, 1995.

254. Schneider, B., Cohen, D. M., Schleifer, L., Srinivasan, A. R., Olson, W. K., and Berman, H. M. A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. *Biophys J*, 65(6): 2291–2303, 1993.

255. Schneider, B., Patel, K., and Berman, H. M. Hydration of the phosphate group in double-helical DNA. *Biophys J*, 75(5):2422–2434, 1998.

256. Schneider, W. M. Rescoring of Docking Poses. Master's thesis, University of Tübingen, 2009.

257. Schultz, S. C., Shields, G. C., and Steitz, T. A. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, 253(5023):1001–1007, 1991.

258. Schutz, C. N. and Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins*, 44(4):400–417, 2001.

259. Schwabe, J. W. The role of water in protein-DNA interactions. *Curr Opin Struct Biol*, 7(1):126–134, 1997.

260. Seeliger, D., Buelens, F. P., Goette, M., de Groot, B. L., and Grubmüller, H. Towards computational specificity screening of DNA-binding proteins. *Nucleic Acids Res*, 39(19):8281–8290, 2011.

261. Seeman, N. C., Rosenberg, J. M., and Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci USA*, 73(3):804–808, 1976.

262. Sham, Y. Y., Muegge, I., and Warshel, A. The effect of protein relaxation on charge-charge interactions and dielectric constants of proteins. *Biophys J*, 74(4):1744–1753, 1998.

263. Shao, J., Tanner, S. W., Thompson, N., and Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J Chem Theory Comput*, 3(6):2312–2334, 2007.

264. Shen, Q.-H., Saijo, Y., Mauch, S., Biskup, C., Bieri, S., Keller, B., Seki, H., Ulker, B., Somssich, I. E., and Schulze-Lefert, P. Nuclear activity of MLA immune receptors links isolate-specific and basal disease-resistance responses. *Science*, 315(5815):1098–1103, 2007.

265. Shrivastava, T. and Tahirov, T. H. Three-dimensional structures of DNA-bound transcriptional regulators. *Methods Mol Biol*, 674:43–55, 2010.

266. Siggers, T. W. and Honig, B. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res*, 35(4):1085–1097, 2007.

267. Singh, N. and Warshel, A. Toward accurate microscopic calculation of solvation entropies: extending the restraint release approach to studies of solvation effects. *J Phys Chem B*, 113(20):7372–7382, 2009.

268. Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17(4):355–362, 1993.

269. Sotriffer, C. A., Krämer, O., and Klebe, G. Probing flexibility and "induced-fit" phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations. *Proteins*, 56(1):52–66, 2004.

270. Spinelli, S., Liu, Q. Z., Alzari, P. M., Hirel, P. H., and Poljak, R. J. The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie*, 73(11):1391–1396, 1991.

271. Stamati, H., Clementi, C., and Kavraki, L. E. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins*, 78(2):223–235, 2010.

272. Steinbrecher, T. and Labahn, A. Towards accurate free energy calculations in ligand protein-binding studies. *Curr Med Chem*, 17(8):767–785, 2010.

273. Steuber, H., Zentgraf, M., Motta, C. L., Sartini, S., Heine, A., and Klebe, G. Evidence for a novel binding site conformer of aldose reductase in ligand-bound state. *J Mol Biol*, 369(1):186–197, 2007.

274. Steuber, H., Heine, A., Podjarny, A., and Klebe, G. Merging the binding sites of aldose and aldehyde reductase for detection of inhibitor selectivity-determining features. *J Mol Biol*, 379(5):991–1016, 2008.

275. Stillinger, F. H. and Rahman, A. Improved Simulation of Liquid Water by Molecular Dynamics. *Journal of Chemical Physics*, 60:1545–1557, 1974.

276. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

277. Stormo, G. D. and Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet*, 11(11): 751–760, 2010.

278. Temiz, N. A. and Camacho, C. J. Experimentally based contact energies decode interactions responsible for protein-DNA affinity and the role of molecular waters at the binding interface. *Nucleic Acids Res*, 37(12):4076–4088, 2009.

279. Thompson, D. C., Humblet, C., and Joseph-McCarthy, D. Investigation of MM-PBSA rescoring of docking poses. *J Chem Inf Model*, 48(5):1081–1091, 2008.

280. Totrov, M. and Abagyan, R. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins*, Suppl 1:215–220, 1997.

281. Totrov, M. and Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol*, 18(2):178–184, 2008.

282. Tsui, V. and Case, D. A. Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *J Am Chem Soc*, 122:2489–2498, 2000.

283. Tsui, V. and Case, D. A. Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulation. *Biopolymers*, 56:275–291, 2001.

284. Ulker, B. and Somssich, I. E. WRKY transcription factors: from DNA binding towards biological function. *Curr Opin Plant Biol*, 7(5):491–498, 2004.

285. van Dijk, M. and Bonvin, A. M. J. J. A protein-DNA docking benchmark. *Nucleic Acids Res*, 36(14):e88, 2008.

286. van Dijk, M. and Bonvin, A. M. J. J. Pushing the limits of what is achievable in protein-DNA docking: benchmarking HADDOCK's performance. *Nucleic Acids Res*, 38(17):5634–5647, 2010.

287. von Hippel, P. H. and Berg, O. G. On the specificity of DNA-protein interactions. *Proc Natl Acad Sci USA*, 83 (6):1608–1612, 1986.

288. Walker, R. *The Development of a QM/MM Based Linear Response Method and its Application to Proteins*. PhD thesis, Imperial College London, 2003.

289. Wang, J., Cieplak, P., and Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem*, 21(12): 1049–1074, 2001.

290. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–1174, 2004.

291. Wang, J., Wang, W., Kollman, P. A., and Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model*, 25(2):247–260, 2006.

292. Wang, R., Lu, Y., Fang, X., and Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J Chem Inf Comput Sci*, 44(6):2114–2125, 2004.

293. Wang, Z., Yang, P., Fan, B., and Chen, Z. An oligo selection procedure for identification of sequence-specific DNA-binding activities associated with the plant defence response. *Plant J*, 16(4):515–522, 1998.

294. Wanke, D. and Harter, K. Analysis of plant regulatory DNA sequences by the yeast-one-hybrid assay. *Methods Mol Biol*, 479:291–309, 2009.

295. Warshel, A. and Russell, S. T. Calculations of electrostatic interactions in biological systems and in solutions. *Q Rev Biophys*, 17(3):283–422, 1984.

296. Warshel, A., Russell, S. T., and Churg, A. K. Macroscopic models for studies of electrostatic interactions in proteins: limitations and applicability. *Proc Natl Acad Sci USA*, 81(15):4785–4789, 1984.

297. Warwicker, J. and Watson, H. C. Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. *J Mol Biol*, 157(4):671–679, 1982.

298. Watson, J. D. and Crick, F. H. C. A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, 1953.

299. Wei, K.-F., Chen, J., Chen, Y.-F., Wu, L.-J., and Xie, D.-X. Molecular phylogenetic and expression analysis of the complete WRKY transcription factor family in maize. *DNA Res*, 19(2):153–164, 2012.

300. Weis, A., Katebzadeh, K., Söderhjelm, P., Nilsson, I., and Ryde, U. Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. *J Med Chem*, 49(22):6596–6606, 2006.

301. Weiser, J., Shenkin, P. S., and Still, W. C. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *J Comput Chem*, 20:217–230, 1999.

302. Wendt, M. D., Rockway, T. W., Geyer, A., McClellan, W., Weitzberg, M., Zhao, X., Mantei, R., Nienaber, V. L., Stewart, K., Klinghofer, V., and Giranda, V. L. Identification of novel binding interactions in the development of potent, selective 2-naphthamidine inhibitors of urokinase. Synthesis, structural analysis, and SAR of N-phenyl amide 6-substitution. *J Med Chem*, 47(2):303–324, 2004.

303. Wiederstein, M. and Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*, 35(Web Server issue):W407–W410, 2007.

304. Wilson, D. K., Bohren, K. M., Gabbay, K. H., and Quiocho, F. A. An unlikely sugar substrate site in the 1.65 Å structure of the human aldose reductase holoenzyme implicated in diabetic complications. *Science*, 257(5066): 81–84, 1992.

305. Woda, J., Schneider, B., Patel, K., Mistry, K., and Berman, H. M. An analysis of the relationship between hydration and protein-DNA interactions. *Biophys J*, 75(5):2170–2177, 1998.

306. Wong, C. F., Kua, J., Zhang, Y., Straatsma, T. P., and McCammon, J. A. Molecular docking of balanol to dynamics snapshots of protein kinase A. *Proteins*, 61(4):850–858, 2005.

307. Xu, X., Chen, C., Fan, B., and Chen, Z. Physical and functional interactions between pathogen-induced Arabidopsis WRKY18, WRKY40, and WRKY60 transcription factors. *Plant Cell*, 18(5):1310–1326, 2006.

308. Yamasaki, K., Kigawa, T., Inoue, M., Tateno, M., Yamasaki, T., Yabuki, T., Aoki, M., Seki, E., Matsuda, T., Nunokawa, E., Ishizuka, Y., Terada, T., Shirouzu, M., Osanai, T., Tanaka, A., Seki, M., Shinozaki, K., and Yokoyama, S. Solution structure of an Arabidopsis WRKY DNA binding domain. *Plant Cell*, 17(3):944–956, 2005.

309. Yamasaki, K., Kigawa, T., Watanabe, S., Inoue, M., Yamasaki, T., Seki, M., Shinozaki, K., and Yokoyama, S. Structural basis for sequence-specific DNA recognition by an Arabidopsis WRKY transcription factor. *J Biol Chem*, 287(10):7683–7691, 2012.

310. Yang, A. Y.-C., Källblad, P., and Mancera, R. L. Molecular modelling prediction of ligand binding site flexibility. *J Comput Aided Mol Des*, 18(4):235–250, 2004.

311. Yanover, C. and Bradley, P. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Res*, 39(11):4564–4576, 2011.

312. Yoshida, T., Nishimura, T., Aida, M., Pichierri, F., Gromiha, M. M., and Sarai, A. Evaluation of free energy landscape for base-amino acid interactions using ab initio force field and extensive sampling. *Biopolymers*, 61(1): 84–95, 2001.

313. Zakrzewska, K., Bouvier, B., Michon, A., Blanchet, C., and Lavery, R. Protein-DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies. *Phys Chem Chem Phys*, 11 (45):10712–10721, 2009.

314. Zhao, Y., Granas, D., and Stormo, G. D. Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12):e1000590, 2009.

315. Zheng, Z., Qamar, S. A., Chen, Z., and Mengiste, T. Arabidopsis WRKY33 transcription factor is required for resistance to necrotrophic fungal pathogens. *Plant J*, 48(4):592–605, 2006.

316. Zou, X., Sun, Y., and Kuntz, I. D. Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model. *J Am Chem Soc*, 121:8088–8043, 1999.