Computational Design and Analysis of Binding Pockets at Protein-Protein Interaction Interfaces

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von

Susanne Eyrisch

Saarbrücken, 2009

Tag des Kolloquiums:	29. Oktober 2009
Dekan:	Prof. Dr. Stefan Diebels
Berichterstatter:	Prof. Dr. Volkhard Helms
	(vertreten durch PD Dr. Michael Hutter)
	Prof. Dr. Dr. Thomas Lengauer
Vorsitz	Prof. Dr. Rita Bernhardt
Akad. Mitarbeiter	Dr. Wei Gu

Acknowledgements

First and foremost, I would like to dedicate my thanks to my supervisor Prof. Dr. Volkhard Helms for offering me the opportunity to work in his group. During the last years, he supported me with fruitful discussions, valuable suggestions, and his encouragement. Especially, I am thankful for offering me the opportunity to attend many interesting conferences and workshops. In this context, I would also like to thank the Center for Bioinformatics, Saar for the initial funding.
I want to thank my co-workers, especially Tihamér Geyer for his technical support and assistance and the “Coffee Club” for very enjoyable scientific, as well as non-scientific discussions.
Furthermore, I thank Jan Fuhrmann and Dirk Neumann who kindly provided their BALLPass implementation to us and Anja Berwanger and Prof. Dr. Rita Bernhardt for giving me interesting insights into the “*in vitro*” world of Biochemistry.
Last, but now least I want to specially thank my friends Benjamin Eckert, Lars Steinbrück, Dorothea Emig, Peter Walter, and Andreas Schlicker for revising my thesis and my family for always believing in me.

Abstract

Protein-protein interactions play a pivotal role in most biological processes. Especially their function in controlling apoptosis makes them to important drug targets. But in contrast to enzymes, the applicability of existing *in silico* methods assisting the design of small-molecule inhibitors is abated by the intrinsic properties of protein-protein interaction interfaces. The central problem is that in the absence of inhibitors, accessible binding pockets are lacking in this region. In this thesis, we present computational approaches for designing and analyzing binding pockets located at protein-protein interaction interfaces. We observed that transient pockets not accessible in the unbound crystal structures of proteins involved in protein-protein interactions are frequently open in alternative protein conformations. At the native binding site, pockets suitable for accommodating known inhibitors were observed. Based on these findings, we studied how these pocket openings occur and developed different protocols for detecting and designing such ligand binding pockets. If no information about the binding site is available, the surface of the entire protein is sampled and all transient pockets opening on the protein surface are identified. If the binding site is approximately known, pockets of predefined properties are algorithmically designed at the desired location. After validating the protocols using three model systems, we show their application to two test systems.

Kurzfassung

Protein-Protein-Interaktionen sind wichtige Angriffspunkte für Wirkstoffe, da sie bei den meisten biologischen Prozessen eine entscheidende Rolle spielen. Im Gegensatz zu Enzymen ist jedoch die Anwendbarkeit existierender *in silico* Methoden zur Unterstützung der Entwicklung niedermolekularer Inhibitoren an Protein-Protein-Schnittstellen eingeschränkt. Das Kernproblem besteht hierbei darin, dass den Kristallstrukturen der ungebundenen Proteine häufig potentielle Bindungstaschen fehlen. In der vorliegenden Arbeit stellen wir computergestützte Ansätze zum Entwurf und zur Analyse von Bindungstaschen an Protein-Protein-Schnittstellen vor. Wir haben entsprechende Proteine untersucht und beobachtet, dass transiente Taschen, die in den ungebundenen Strukturen nicht zugänglich waren, häufig in alternativen Konformationen geöffnet sind und sich zudem als Bindungstaschen für bekannte Inhibitoren eignen. Des Weiteren haben wir untersucht, wie diese Taschenöffnungen zustande kommen und dieses Wissen in der Entwicklung neuer Vorgehensweisen zur Ermittlung solcher Ligandenbindungstaschen berücksichtigt. Ist keine Information über die Bindungsstelle verfügbar, wird die gesamte Proteinoberfläche nach transienten Taschen abgesucht. Ist die Bindungsstelle aber annähernd bekannt, können Bindungstaschen mit den gewünschten Eigenschaften algorithmisch entworfen werden. Nachdem diese Vorgehensweisen anhand dreier Modellsysteme validiert wurden, stellen wir deren Anwendung auf zwei Testsysteme vor.

Zusammenfassung

In der pharmazeutischen Forschung gewinnen computergestützte Methoden, die die Entwicklung neuer Wirkstoffe unterstützen, zunehmend an Bedeutung. So lässt sich die stetig steigende Anzahl neu entdeckter Enzym-Inhibitoren nicht nur auf die verbesserten experimentellen Screening-Techniken zurückführen, sondern auch auf den kontinuierlichen Fortschritt im strukturbasierten Wirkstoffdesign und die Tatsache, dass immer mehr hochauflösende Proteinstrukturen verfügbar werden. Ist die dreidimensionale Struktur des Zielproteins bekannt, kann man mittels computergestützter Methoden die zu blockierende Bindungsstelle ermitteln und deren chemische und geometrische Eigenschaften mit denen potentieller Liganden vergleichen. Obwohl strukturbasiertes Wirkstoffdesign sehr erfolgreich bei der Identifizierung von Inhibitoren eingesetzt wird, die auf die Wechselwirkung zwischen Proteinen (zumeist Enzyme) und kleinen Molekülen einwirken, lässt der Erfolg bei der Entdeckung von Liganden, die die Bildung von Protein-Protein-Komplexen modellieren, noch auf sich warten. Dabei liegt ein enormes therapeutisches Potential in der Hemmung von Protein-Protein Interaktionen, da diese eine entscheidende Rolle in fast allen wichtigen biologischen Prozessen spielen, wie zum Beispiel im Tumorwachstum oder der Immunantwort. Daher hat sich die Suche nach kleinen Molekülen, die eine entsprechende inhibierende Wirkung zeigen (auch SMPPIIs, “small-molecule protein-protein interaction inhibitors”, genannt), in den letzten Jahren zu einem sehr aktiven Forschungsfeld entwickelt. Jedoch sind bisher fast alle bekannten SMPPIIs mittels experimenteller Screening-Methoden entdeckt worden. Das strukturbasierte Wirkstoffdesign hat sich für diese Klasse von Proteinen als eine große Herausforderung erwiesen. Bei Enzymen befindet sich das aktive Zentrum für gewöhnlich in wohldefinierten, tiefen Bindungstaschen, in die potentielle Inhibitoren binden können. Im Gegensatz dazu befinden sich jedoch an den Schnittstellen der meisten ungebundenen Proteinstrukturen keine für die Ligandenbindung geeignete Vertiefungen. Daher ist es nahezu unmöglich Inhibitorbindungsstellen zu identifizieren, wenn diese nicht aus Ligand-gebundenen Kristallstrukturen bekannt sind. Selbst wenn der Bereich, in dem der Inhibitor bindet oder binden sollte, bekannt ist, verläuft die Anwendung computergestützter Methoden zur Suche nach vermeintlichen Treffern aus virtuellen Ligandenbibliotheken meist ohne Ergebnis, wenn keine potentielle Bindungstasche vorhanden ist, in die die Liganden platziert werden können. Das Ziel dieser Arbeit ist es daher, je nach verfügbarer Information über die Bindungsstelle geeignete Bindungstaschen zu ermitteln und zu analysieren oder so zu entwerfen, dass diese bestimmte Anforderungen erfüllen. Dass diese Vorgehensweise gerechtfertigt ist, zeigt unsere Eingangsstudie, in der wir drei Modellsysteme mittels Moleküldynamik-Simulationen in Wasser untersucht haben. Hierbei wurde in allen Fällen ein häufiges Auftreten von transienten Bindungstaschen, die nicht in der ungebundenen Startstruktur vorhanden waren, auf der Proteinoberfläche beobachtet. Da diese Modellsysteme so ausgewählt wurden, dass die Bindungsmoden eines Inhibitors aus einer Kristallstruktur bekannt ist, konnte dieses Wissen zur Validierung der Ansätze genutzt werden. Dadurch konnten wir zeigen, dass sich unter den beobachteten transienten Taschen auch die native Bindungstasche befindet und diese selbst in Abwesenheit ihres Liganden eine Form annimmt, in die der Inhibitor in einer der Kristallstruktur sehr ähnlichen Weise binden kann. Dieses Ergebnis weist darauf hin, dass die Benutzung transienter Bindungstaschen das strukturbasierte Wirkstoffdesign von Protein-Protein-Interaktionsinhibitoren erheblich erleichtern könnte.
In einer Folgestudie haben wir den Einfluss des Proteinrückgrates und des in Moleküldynamik-

Simulationen benutzten Lösungsmittels auf die Bildung von transienten Bindungstaschen untersucht und deren essentielle Bedeutung festgestellt. So wurden während einer Moleküldynamik-Simulation in Methanol mehr Taschenöffnungen als in der Vergleichssimulation in Wasser beobachtet. Des Weiteren waren diese Taschen größer und unpolarer, was darauf schließen lässt, dass das Öffnen solcher Taschen in Methanol energetisch günstiger ist als in Wasser. Darüber hinaus wurde der Einsatz von effizienteren Methoden zur Generierung von Proteinkonformationen geprüft, deren Ergebnisse jedoch denen der Moleküldynamik-Simulationen qualitativ unterlegen waren.
Aufgrund des hohen Zeitaufwandes dieser Simulationen haben wir eine weitere Vorgehensweise entwickelt, die angewendet werden sollte, wenn die Bindungsstelle der potentiellen Liganden annähernd bekannt ist. In solch einem Fall bietet es sich an, eine Bindungstasche mit den gewünschten Eigenschaften an einer bestimmten Stelle algorithmisch zu erzeugen. Im ersten Versuch der Umsetzung dieser Idee wird die Proteinoberfläche der potentiellen Bindungsregion nach energetisch günstigen Taschenpositionen abgesucht. Hierzu wird eine Kugel, die die Tasche repräsentiert, in die Proteinoberfläche gesetzt und das Protein energetisch minimiert, damit sich seine Konformation der Kugel anpasst und so eine Taschenvorstufe entsteht. Die so erzeugten Proteinkonformationen werden anschließend verfeinert, so dass die endgültigen Konformationen einen Kompromiss zwischen einer möglichst großen Tasche und einer möglichst geringen internen Proteinenergie darstellen. Da dieser Ansatz jedoch nur für zwei der drei Modellsysteme zufriedenstellende Ergebnisse lieferte, wurde eine verbesserte Vorgehensweise entwickelt, bei der die zu induzierende Tasche durch eine Anzahl kleiner Kugeln repräsentiert wird, deren Positionen in Abhängigkeit der Proteinkonformation gewählt werden. Die Grundidee ist hierbei, dass diese Kugeln anfangs stark mit den Proteinatomen überlappen, diese Überlappung jedoch mit zunehmender Anpassung der Proteinkonformation an die gewünschte Tasche reduziert wird. Diese Methode erlaubt es, neben der Position der zu erzeugenden Bindungstasche auch deren Volumen zu definieren. Die Anwendung dieses Ansatzes auf die drei Modellsysteme lieferte sehr vielversprechende Ergebnisse.
Die Erkenntnisse, die wir aus den hier beschriebenen Studien gewonnen haben, wurden abschliessend verwendet, um die Bindungsstellen und -moden experimentell bestimmter Liganden zweier Systeme vorherzusagen. Bei einem der Systeme war weder die Bindungsstelle der Liganden bekannt, noch auf welchem der an der Reaktion beteiligten Proteine sich diese befindet. Daher wurde in diesem Fall die gesamte Oberfläche aller in Frage kommender Proteine mit bekannter Kristallstruktur nach potentiellen Bindungstaschen abgesucht und getestet, ob die Liganden mit ausreichender Affinität darin binden können. Zusätzlich wurden alle mittels einer Moleküldynamik-Simulation des mutmaßlichen Zielproteins erzeugten transienten Taschen auf deren Eignung als Ligandenbindungstasche hin untersucht. Im zweiten Testsystem war die Bindungsregion der Liganden bekannt. Da das Protein an dieser Stelle jedoch eine außergewöhnlich hohe Flexibilität aufwies, haben wir uns auch hier für die Suche nach transienten Taschen mittels Moleküldynamik-Simulationen entschieden, in die die Liganden anschließend platziert wurden, um deren Eignung als Ligandenbindungstasche zu bewerten. In beiden Fällen konnten wir mittels der hier vorgestellten Methoden potentielle Bindungsstellen identifizieren und mögliche Bindungsmoden der Liganden vorschlagen.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AdR	Adrenodoxin Reductase
ASP	active site point
Adx	Adrenodoxin
BCL-X_L	basal cell lymphoma-extra large
DEE	dead-end elimination
DIZ	(2S)-2-(4-chlorophenyl)-2-[(3S)-3-(4-chlorophenyl)-7-iodo-2,5-dioxo-1,3-dihydro- 1,4-benzodiazepin-4-yl]acetic acid
FRH	5-[[2,3-dichloro-4-[5-[1-[2-[[(2R)-2-(diaminomethylideneamino)-4-methylpentanoyl] amino] acetyl] piperidin-4-yl]-1-methylpyrazol-3-yl]phenoxy]methyl]furan-2-carboxylic acid
GMEC	global minimum-energy conformation
GPS	generic pocket sphere
HTS	High Throughput Screening
IL-2	interleukin-2
IL-2Rα	interleukin-2 α-receptor
(L)GA	(Lamarckian) genetic algorithm
MDM2	mouse double minute 2
MD	molecular dynamics
N3B	4-(4-fluorophenyl)-N-[3-nitro-4-(2-phenylsulfanylethylamino)phenyl]sulfonylbenzamide
NMA	normal mode analysis
PDB	protein data bank
PID	pocket identifier
PLA	pocket lining atom
PPI	protein-protein interaction
Put	putrescine
(Q)SAR	(quantitative) structure activity relationship
RCS	relaxed complex scheme
RMSD	root mean square deviation
RMSF	root mean square fluctuation
SMPPII	small-molecule protein-protein interaction inhibitor
Spd	spermidine
Spn	spermine
XIAP	X chromosome-linked inhibitor of apoptosis protein

Chapter 1

Introduction

About 2,300 years ago, Aristotle stated: “The whole is more than the sum of its parts.” Along the same lines, a simple conglomeration of different molecules in a specified ratio in a biological cell does not result in a living organism. It is the strictly coordinated and regulated interaction of these parts that constitutes life. The first chapter provides an introduction into protein-protein interactions that are the drug design targets tackled in this thesis, and into the objective of this work.

1.1 Molecular Interactions and their Modulation

Strictly speaking, all physiological processes, or biological processes in general (like reproduction, cell growth, signal transduction, cell recognition, and metabolism), involve interactions between molecules [1]. For example, the transfer of information from DNA to proteins as described by the central dogma of molecular biology is mediated by interactions between different kinds of molecules that interact with each other by forming *complexes* of variable stability. Complexes may contain two up to several thousands of molecules that bind to each other either covalently or non-covalently. In this context proteins are of particular interest. Besides representing the most abundant class of molecules by accounting for more than 50 % of the dry weight of cells [2], their

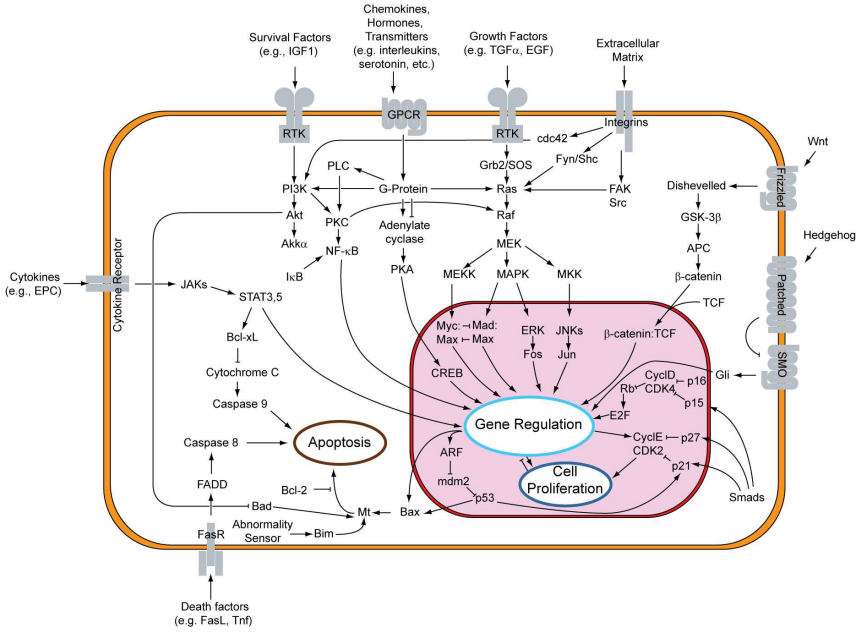


Figure 1.1: A part of the “whole” : Overview of signal transduction pathways (Figure taken from [3]).

Figure 1.2: A typical drug discovery pipeline.

molecular structures determine their biochemical functions and are, therefore, crucial for biological processes. The connection between the biochemical function of a protein and the output of a biological process is illustrated using the example of signal transduction. Such a process typically comprises a sequence of molecular interactions that are mediated by binding events. Upon binding, an effect is triggered (usually a chemical reaction or a structural change depending on the biochemical function of the involved protein) that is required for the activation of the next molecular interaction. This process is continued until the end of this *interaction pathway* is reached and the final result is obtained [4, 5]. Figure 1.1 provides an overview of signal transduction pathways in the cell and illustrates the complexity and variety of biomolecular interactions. Note that some of the proteins shown here are discussed later in this thesis. Although we focus on isolated interactions, one should always keep in mind that they are part of a large cellular network as depicted by this figure. Therefore, it is essential for the appropriate functioning of a biological process that all components interact in a proper way. Even a single aberrant molecular interaction may perturb the process and lead to an alleviated, abnormal, or even missing physiological effect and may be related to a disease [2].

In the ideal case, the medical treatment would consist of modulating these molecular interactions such that the appropriate physiological effect is recovered. Chemical substances (mostly small molecules) that exhibit such an impact on a living organism are called *drugs*. In contrast to the historical procedure that was mostly based on serendipity, the modern approach tackles the problem of discovering new drugs rationally [6]. An example for a drug discovery pipeline is shown in Figure 1.2. The first step in such a drug discovery project consists of identifying an eligible *target* that is involved in the aberrant interaction pathway, usually a protein, on which the potential drug should act. (Typical drug targets and their portion among the targets of all approved drugs are shown in Figure 1.3.) Subsequently, High Throughput Screening (HTS) libraries consisting of several thousand compounds are searched for so-called *hits* that modulate the activity of the target protein to the desired extent. Based on these hits, new analogs with improved pharmacological and biochemical properties (like improved potency and selectivity, reduced side-effects and toxicity) are then synthesized. These *lead* compounds serve as starting points for further refinement

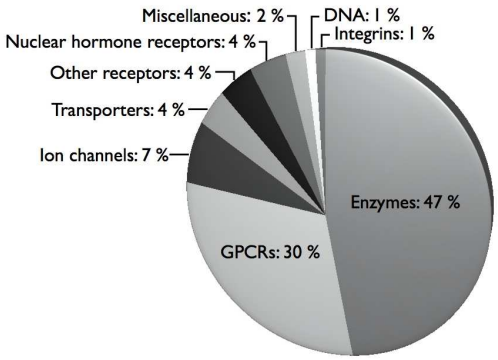
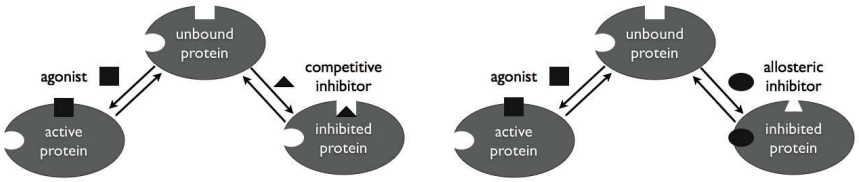


Figure 1.3: Marketed small-molecule drug targets by biochemical class (data taken from [9]). Most drug targets are proteins. Note that the large fraction of enzymes does not only indicate their importance, but also that the modulation of their activity is nowadays a quite successful enterprise.



(a) competitive drug binding (b) allosteric drug binding

Figure 1.4: The drug may either bind to the same site as the agonist ((a) into the quadratic pocket) and block this binding pocket, or to another site ((b) into the spherical pocket) and so induce conformational changes that encompass the binding site of the agonist. In the former case, the drug binding is called *competitive* and in the latter case it is called *allosteric*.

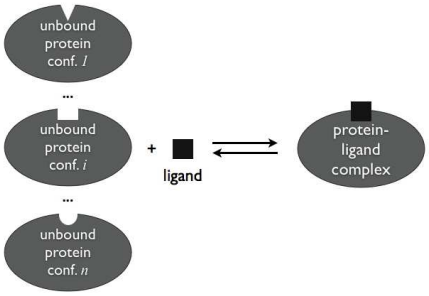
and testing. Afterwards, the most promising compounds, the *drug candidates* enter the preclinical and clinical test phases that are required for the final approval of the new drug. Nowadays most of these steps are assisted by computational methods [7, 8]. In this way, for example, the costs for the hit identification can be significantly reduced if only those compounds are tested *in vivo* that were predicted to be hits in a *virtual screening* campaign. Here, databases consisting of thousands of compounds are screened *in silico* to identify ligands that bind to the target protein with an appropriate computed affinity. In this work, we only focus on approaches that foster the *in silico* discovery of new hits. Further information about how computational methods assist the discovery and design of new drugs can be found in [7, 10–13].

Drugs, or ligands in general, commonly affect the behavior of their target proteins by non-covalently binding to them. The involved surface region of the protein where this interaction takes place is called *binding site*. A typical binding site for a small-molecule ligand is characterized by a cavity or depression on the protein surface, the so-called *binding pocket*, that accommodates the ligand in a protein-ligand complex. Depending on the ability of this complex to produce a functional response, one distinguishes between *agonists* and *antagonists*. An agonist alters the protein's activity (either positively or negatively) upon binding, whereas an antagonist (also called *inhibitor*) does not provoke a biological response itself. It solely functions by damping or blocking the binding of agonists. In the following, we focus on drugs that act as inhibitors because inhibition is the most commonly used strategy in modulating molecular interactions. The binding of the drug to the target protein may either be *competitive* or *allosteric*. In the former case, the drug binds to the same site on the protein as its natural ligand(s) but usually with higher affinity (Figure 1.4 (a)). In the latter case, the drug binds to a distinct site and triggers a conformational change that encompasses the binding site for the natural ligand(s) (Figure 1.4 (b)). In both cases, the physiological complex cannot be formed because the binding site is either occupied or distorted [6].

1.2 Structure-Based Drug Design

Structure-based drug design is an example of rational drug design where information about the three-dimensional structure of the studied molecules is used to assist the drug design process. Here, advantage is taken of the fact that in a complex, the protein and the ligand possess complementary geometric shapes and physicochemical properties. However, this observation only holds for proteins and ligands in their bound (*holo*) states. When it comes to proteins and ligands in their unbound (*apo*) states, the "lock-and-key" model (see Figure 1.5 (a)) as suggested by Emil Fischer in 1894 [14] that describes the binding site as rigid has been proven to be inaccurate. The currently most accepted model was published in 1958 by Daniel Koshland and is a modification to Fischer's "lock-and-key" model [15]. This so-called "induced-fit" model (see Figure 1.5 (b)) considers proteins and ligands as rather flexible structures that are able to reshape upon binding

(a) "lock-and-key" (b) "induced-fit"



(c) "conformational selection"

Figure 1.5: Different models for describing ligand binding: (a) "lock-and-key" [14], (b) "induced-fit" [15], and (c) "conformational selection" [16].

for maximizing their geometric and physicochemical complementarity. As the unbound receptor exists, like all molecules, in an ensemble of accessible conformations a further concept termed "conformational selection" [16] has been suggested to explain ligand binding. This model can be considered as an extension to the "induced-fit" model as it assumes that the ligand will "pick" the protein conformation into which it fits best *before* any conformational changes are induced to optimize the fit (see Figure 1.5 (c)).
In any case, structural information on the biomolecules involved in the targeted interactions facilitates the discovery and design of new drugs. If the three-dimensional structure of the target protein (or a homologous protein) is known, a *direct drug design* approach can be applied. Otherwise, available information about molecules binding to the same protein site can be employed in an *indirect drug design* approach. Note that both strategies do not exclude each other. Modern drug design projects commonly use a combination of both approaches (see [7] for examples).

1.2.1 Direct Drug Design

In an ideal case a high-quality three-dimensional atomic structure of the target protein has been determined by X-ray crystallography or NMR spectroscopy. Alternatively, the experimental structure can be substituted by a homology model derived from a (set of) protein(s) with similar sequence, and hence, structure (see [17] for a review). Having structural information at hand about the target protein then allows for the calculation of its physicochemical and geometrical properties

(a) deep binding pocket (b) flat surface pocket

Figure 1.6: The morphology of the binding site has a crucial influence on the degree of complementarity between the protein and its ligand: The deeper binding pocket in (a) tends to have a larger surface area and so more interactions between the protein and the ligand are possible than in (b). Therefore, the maximal degree of complementarity (illustrated here by the number of complementary charges) is also much higher.

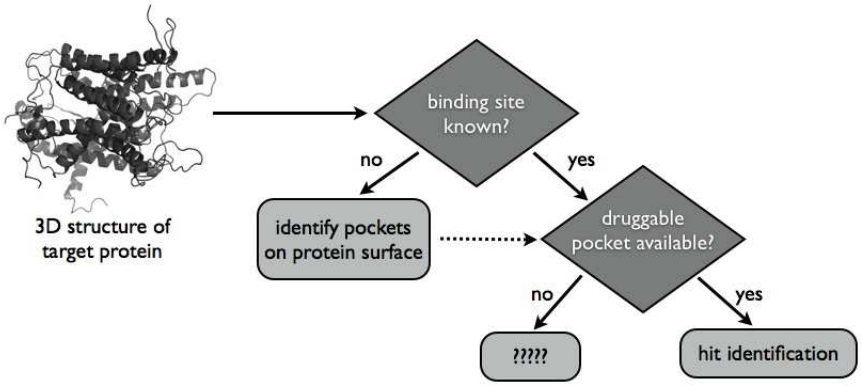


Figure 1.7: An example strategy using direct drug design for hit identification. Note that the existence of a (druggable) binding pocket is crucial for successful virtual screening.

as mentioned above. But equally important to having access to the protein structure is knowing the binding site of the potential drug as it represents the starting point for structure-based drug design [13]. Proteins often possess multiple pockets on their surface but not all of them are *druggable*, i.e. appropriate for ligand (or drug) binding. As designing competitive inhibitors is usually the most straightforward approach, one usually focusses on the pocket that accommodates the interaction to be prevented. An enzymatic reaction, for example, takes place in the active site that is usually located in the largest pocket on the protein surface [18].
A high-quality three-dimensional atomic structure of the apo protein may be more appropriate for drug design than a low-quality structure of the protein-ligand complex because detailed descriptions of putative ligand binding sites can be extracted from the apo structure when focussing on accessible pockets on the protein surface. An advantage of ligands binding into pockets is that the contact area between the protein and the ligand is much larger than when binding to a flat protein surface. The degree of complementarity and, thus, the specificity is much higher as well (compare Figure 1.6) [19]. Therefore, knowing the binding site is crucial for identifying potential drugs *in silico*. In addition, this site should be druggable. If these preconditions are fulfilled, huge compound libraries can be virtually screened for putative hits that fit into this binding site, e.g. by molecular docking. An example for a direct drug design strategy is illustrated in Figure 1.7. The steps covered in this thesis are explained in detail in Chapter 2.

1.2.2 Indirect Drug Design

If the structure of the protein target is unknown, new compounds will be designed on the basis of a hypothetical binding site that is derived from an analysis of the physicochemical and geometrical properties of known binders and non-binders. This approach is based on the principle of similarity that assumes that similar compounds produce similar effects [7]. This is again related to the complementarity between the protein and the ligand in their complexed state.
Examples for indirect drug design approaches are:

- Quantitative structure-activity relationship (QSAR) analysis
- Molecular shape analysis
- Pharmacophore generation and mapping

As indirect drug design is beyond the scope of this work, we refer to two reviews [7, 12] for further information.

1.3 Protein-Protein Complexes and their Modulation

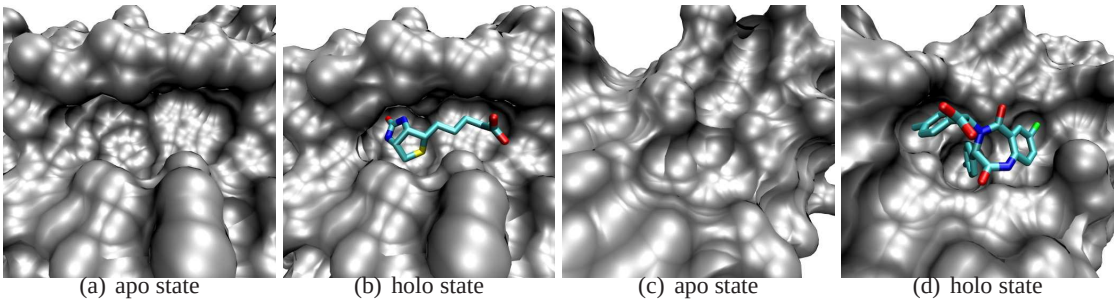
The interactions between proteins play a central role in most physiological processes. Interestingly, the protein-protein complexes that are formed during such interactions have different degrees of stability and duration [20]. Several possible classifications of these protein-protein complexes have been suggested (see [21] for a review). Here, we use the classification into *transient* and *permanent* complexes. In permanent complexes, proteins are only stable in oligomeric structures. Therefore this kind of interaction can be seen as a continuation of protein folding. Transient interactions, on the other hand, represent short-living complexes formed by proteins that are also stable in the apo form [20].
This work focuses on transient protein-protein interactions (PPIs) because they are involved in important biological processes like immune response and signal transduction (e.g. apoptosis or proliferation) and, thus, involve important drug targets [2, 8, 19, 20, 22–30].

1.3.1 Inhibiting Protein-Protein Interactions by Small Molecules

Designing small-molecule inhibitors that occupy enzyme active sites is nowadays a common and successful enterprise [29] as reflected by the small-molecule drug targets shown in Figure 1.3. But when it comes to small-molecule protein-protein interaction inhibitors (SMPPIIs), their design is widely regarded as a formidable challenge and so far, only a few SMPPIIs have been approved as drugs. The inherent difficulties mainly arise from the nature of protein-protein interaction interfaces [2, 23, 25–27, 29]. Table 1.1 compiles the differences to the design of inhibitors targeting interactions between enzymes and their small-molecule substrates. Figure 1.8 shows a particular example pointing out the contrast between binding sites in the apo and holo state for protein-protein interaction interfaces and small molecule - enzyme interactions. As protein-protein interaction interfaces mainly consist of hydrophobic residues they are quite featureless making it difficult to ensure that the small molecules bind with sufficient specificity. However, in many cases only a few residues contribute to high-affinity binding (so-called *hot spots* [31]) and, thus, a small-molecule inhibitor does not necessarily need to cover the entire protein-binding interface. The subset of the surface consisting of the hot spot residues is much smaller and, hence, suited to be masked by a small molecule [29]. This is underpinned by the fact that to date several SMPPIIs have been identified [2, 8, 19, 22, 23, 25, 27–29].

	PPI interface	**enzyme-substrate binding site**
morphology of binding site	relatively flat; often no deep binding pockets; often many small subpockets	well-defined deep binding pocket
similarity of binding site in the apo and holo state	major conformational changes upon ligand binding; binding pocket often not accessible in the apo state (Fig. 1.8 (c) + (d))	in most cases only minor conformational changes upon ligand binding (Fig. 1.8 (a) + (b))
surface area of binding site	1,500 - 3,000 Å^2 [25]	300 - 1,000 Å^2 [25]
spatial distribution of binding site	distributed	contiguous
natural ligands	proteins	small molecules
dominant interactions with binding partners	hydrophobic interactions	hydrogen bonds, salt bridges, and electrostatic forces

Table 1.1: Comparison between protein-protein interaction interfaces and deep enzyme pockets as drug binding sites. It should be mentioned here that enzymes may also contain binding sites for other proteins. But within this context, we focus on the active site or other binding pockets for small-molecule ligands.



(a) apo state (b) holo state (c) apo state (d) holo state

Figure 1.8: Ligand-induced conformational changes in enzyme active sites and protein-protein interaction interfaces: the binding sites of Biotin on the enzyme Streptavidin (a, b) and of the SMPPII DIZ on the protein MDM2 (c, d). (a) and (b) possess an almost identical binding site, whereas the pockets in (c) and (d) show noticeable differences.

Examples for targeted protein-protein interactions are:

- BCL-X_L - Bak
- MDM2 - p53
- IL-2 - IL-2Rα
- XIAP-BIR3 - Caspase-9

This emphasizes the fact that although finding SMPPIIs is indeed a great challenge, but it is not an impossible one.

1.3.2 Experimental Approaches for Targeting Protein-Protein Interactions by Small Molecules

Protein-protein interactions may be modulated by different classes of molecules. For example, if the interface of at least one protein consists of a short continuous amino acid sequence that contributes significantly to the overall binding affinity this binding patch can be mimicked by a peptide. Although peptides are generally inappropriate as oral drugs because of their poor metabolic stability and low bioavailability, they may serve as a lead compound that is subsequently optimized by chemical modifications like the inclusion of non-natural amino acids [27]. Another strategy is the identification of small-molecule binders by HTS of chemical libraries. The advantages of this approach are that it is also applicable if the binding site cannot be mimicked by a peptide and that the hit or lead compounds are already small molecules, so that there is no need to mimick a peptide by a small molecule as in the first case. A further advantage is that no prior knowledge about the location or the constitution of the interaction interface is required. However, the hits identified in an initial functional screening have to be further tested to rule out assay-specific artifacts. Moreover, as the assays usually contain both (or even several different) proteins that form the targeted complex, it has to be clarified to which protein the compound binds. The success of a HTS crucially depends on the size and diversity of the compound library. In fact, it may happen that the conversion of hits identified by HTS to lead compounds is unsuccessful. A possible explanation is that most screening libraries are designed for “traditional” drug targets like enzymes or G-protein-coupled receptors (compare Figure 1.3) and are, due to the different physicochemical properties of protein-protein interaction interfaces, unsuitable for binding to them [25].
An alternative approach that overcomes this problem is *fragment-based screening* [23]. The idea behind this strategy is that initially a library of small organic fragments (with masses typically less than 200 Da) are screened for active representatives. These are linked or otherwise optimized to generate a small set of drug-sized molecules that are tested for an improved binding affinity.

The advantage of this approach is that a huge chemical space can be probed while only a minimal number of compounds has to be synthesized. The most commonly used methods for fragment-based screening are *SAR by NMR* and *Tethering*. SAR (Structure-activity relationship) by NMR is an NMR-based method that identifies fragments that bind to proximal subsites of the protein. These fragments are then linked or merged by using a combination of structure-based design and SAR [32]. In Tethering, a protein residue located near the binding site is mutated to cysteine and the protein reacts with a library of disulfide-containing fragments. At equilibrium, the protein-fragment complexes with highest affinity will predominate [33]. The drawback of this technique is that the location of the binding site of the fragments has to be known in advance and that the mutation may influence the affinity towards the screened fragments.

1.3.3 Computational Approaches for Targeting Protein-Protein Interactions by Small Molecules

Although it is nowadays possible to test up to 100,000 compounds a day in HTS, potential hits could be identified even more efficiently if the size of the compound library is decreased. To this end *in silico* approaches exist that design compound libraries either structure- or diversity-based. Diversity-based library design, for example, generates focussed libraries by similarity clustering while maintaining the diversity of the complete library. If structural information about the target protein is available, it can be used to select or design compounds that are to be tested experimentally. Several published studies reported the identification of SMPPIIs by using a combination of *in silico* and *in vitro* screening. As for most studied protein-protein complexes no small-molecule binders had been identified before, direct drug design approaches were applied. In all examples reported so far, the location of the binding site was already known from experimental studies revealing hot spot residues or predicted from high-quality (complex) crystal structures that contained accessible binding pockets. (Case studies are described, for example, in [8, 22, 27, 29].) Analogously to the standard direct drug design approach applied to "traditional" drug targets, the procedure for protein-protein interactions illustrated in Figure 1.9 comprises two key levels: the prediction of potential binding sites from hot spot analysis, pocket detection, and/or detection of

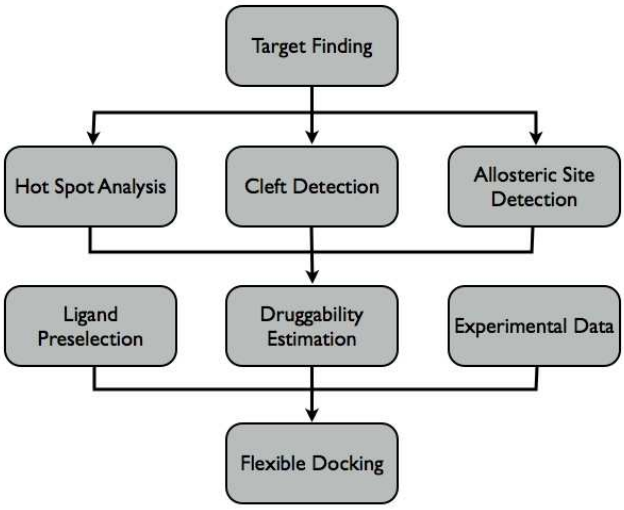


Figure 1.9: Flowchart as suggested in [26] describing an *in silico* approach for the discovery of SMPPIIs. This procedure comprises two key levels: identification of an appropriate binding site and virtual screening by docking ligands into the selected receptor region. Note that this approach relies on the identification of a druggable binding site from the target structure.

allosteric sites and the actual virtual screening step in which ligands are docked into the receptor site [26]. So far, only a few studies [34–37] considered the flexibility of the binding region. In fact, docking into flat cavities located at protein-protein interaction interfaces is more challenging than docking into deep enzyme pockets. Najmanovich *et al.* analyzed the conformational changes of side-chains upon ligand binding [38]. By comparing their findings to a similar study about the side-chain rearrangements upon protein-protein association [39], they concluded that side-chains in ligand binding pockets are more rigid than those in protein-protein interaction interfaces. Thus, it appears that other studies that did not incorporate receptor flexibility were only successful because their target structure already contained a well-defined pocket that was appropriate for ligand binding. Note that this is not always the case, especially if only a crystal structure of the apo protein is available. In such a case the procedure described before is not applicable and the identification of potential hit compounds is completely reliant on the success of *in vivo* or *in vitro* screening approaches.

1.4 Goal of this Work

The present work introduces computational approaches that assist in the design of small-molecule inhibitors at protein-protein interaction interfaces. A particular focus is placed on the identification of binding sites for potential hits. This step is crucial for the design and discovery of SMPPIIs. As outlined above, detecting binding sites in enzymes is not very challenging, even if solely structural information about the apo conformation is available. In contrast to this, for proteins involved in protein-protein interactions, missing knowledge about the binding site may impede the whole drug discovery pipeline. If only the apo structure of the target protein is on-hand, the whole protein surface has to be considered when searching for ligand binding sites. Moreover, as illustrated in Figure 1.8, detecting binding sites in apo proteins may fail because putative conformational changes that result from ligand binding are not represented by a single conformation. To this end, we developed different approaches for identifying or designing putative ligand binding pockets using apo protein structures as outlined in Figure 1.10. After giving an overview of the underlying theory

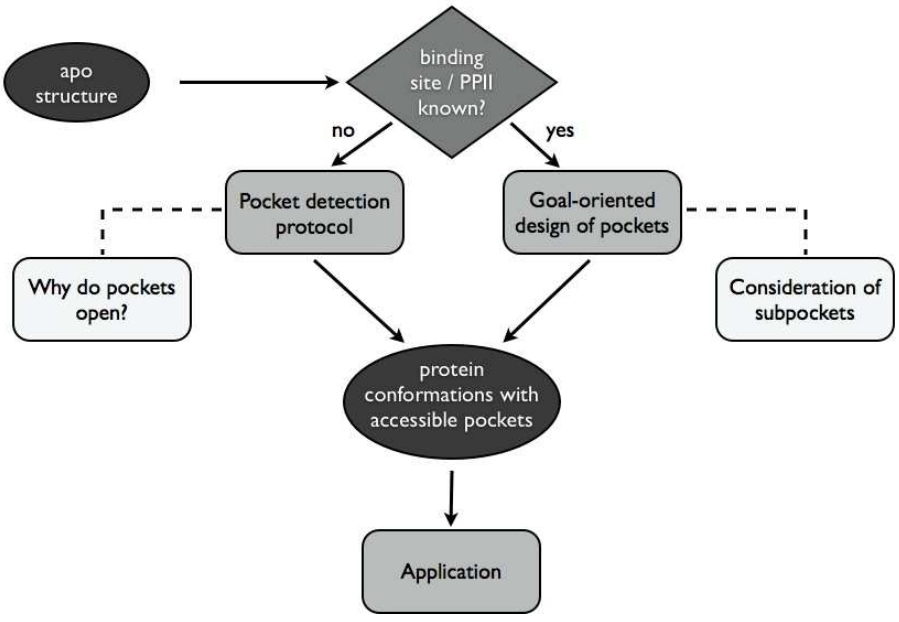


Figure 1.10: Flowchart illustrating the goal of this work. Every rectangular box is covered by a chapter (Chapters 3 - 7).

and (alternative) methods in Chapter 2, our initial pocket detection protocol based on molecular dynamics simulations will be introduced in Chapter 3. As this approach led to the surprising result that many pockets not accessible in the starting structure are open in other conformations, the underlying mechanisms of these pocket openings were studied using modified versions of this protocol and the findings will be presented in Chapter 4. If the protein-protein interaction interface is known (e.g. from the complex structure), the search for accessible binding pockets can be limited to this area. Besides, the reduction of the search space makes the goal-oriented design of binding pockets computationally feasible as will be shown in Chapter 5. However, binding pockets for SMPPIIs often consist of several subpockets. Considering this fact, an extension was developed that will be described in Chapter 6. Finally, in Chapter 7, the methodology will be applied to two test systems: Adrenodoxin and the BIR-2 domain of XIAP.

Chapter 2

Background

Since the 1980s, *in silico* methods have become more and more important in drug design projects [12]. The underlying concepts and algorithms most relevant for the present work will be introduced in this chapter.

2.1 Statistical Thermodynamics of Binding Reactions

Molecules exist in a crowded world. Besides many other species of molecules that are present in large numbers in (or outside) the cell, they are also surrounded by many molecules of their kind. But not all of them are exactly in the same state. Some molecules, for example, may be part of a complex while others are in the unbound state. As a consequence, experiments rather study the behavior of an ensemble of molecules than that of a single molecule. The behavior of such a macroscopic system is characterized by its volume, temperature, pressure, number of particles, total energy, and a variety of other macroscopic parameters and is described by *thermodynamics*. *Energy*, for example, is one of the most important concepts in chemistry. It determines which molecules exist (and how they look like), which reactions occur (and in which direction they are executed), and how a system behaves. Thus, the calculation of energy is essential in computational structural biology. The term energy is commonly defined as a quantity describing the amount of work that can be performed by a force. It exists in different forms (e.g. potential and kinetic energy) that can be transformed into each other while the total energy of a closed system (its *internal energy* U) is always conserved (first law of thermodynamics). The heat transfer during a reaction taking place in a closed system is calculated by the *enthalpy* H

$$H = U + pV \tag{2.1}$$

where pV is the work required to allocate space of volume V against a constant pressure p. Together with the *entropy* S, interpreted as a measure how close the system is to equilibrium, and the absolute temperature T, the enthalpy is used to calculate *Gibb's free energy* that defines how much process-initiating work can be obtained from an isothermal and isobaric closed system:

$$G = H - T \cdot S \tag{2.2}$$

The change in free energy,

$$\Delta G = \Delta H - T \cdot \Delta S \tag{2.3}$$

indicates which reactions take place under the given conditions:

- $\Delta G < 0$: reaction occurs spontaneously
- $\Delta G = 0$: no reaction occurs (system is in equilibrium)
- $\Delta G > 0$: reaction occurs non-spontaneously

In Chapter 1, the term *affinity* was used to describe the binding stability. As the ligand L, the target protein P, and the complex PL occur more than once and in more than one state, their concentrations are measured after the equilibrium state

$$P + L \rightleftharpoons PL \tag{2.4}$$

is reached. The ratio of their concentrations is called *dissociation constant* and is defined as

$$K_d = \frac{[P] \cdot [L]}{[PL]} \tag{2.5}$$

Note that this relationship describes the binding affinity of a ligand towards its target protein. For example, in virtual screening one tries to identify hits that are predicted to bind with K_d rates in the micromolar range [22]. This affinity is influenced by non-covalent interactions between the two binding partners like electrostatic interactions, hydrogen bonds, and van der Waals interactions and is, thus, also represented by the ratio between the speed of the dissociation k_{off} and the association k_{on} of the complexes. Under standard conditions, the direction of this binding reaction at steady state depends on the energy free difference of the two states

$$\Delta G = RT \cdot \ln K_d \tag{2.6}$$

where R is the gas constant [40].
In contrast to experiments that measure the behavior of ensembles of molecules, computational approaches usually calculate the behavior of an individual molecule. How can these microscopic results be used to explain the macroscopic results of the experiments? This link is provided by the *partition function* q, the key quantity of statistical mechanics. It can be used to calculate all macroscopic functions [41]. For a single molecule in the canonical NVT ensemble (see Section 2.3.1), it is defined as the sum over all possible energy states E_i accessible at temperature T

$$q = \sum_{i=states}^{\infty} e^{\frac{-E_i}{k_B T}} \tag{2.7}$$

where k_B is the Boltzmann constant. When considering macroscopic systems in thermal equilibrium, one is usually interested in the average microstate with energy E_i, i.e. the state i for which the probability P_i is maximal. In the canonical ensemble, the probabilities follow the well-known Boltzmann distribution,

$$P_i = \frac{1}{q} \cdot e^{\frac{-E_i}{k_B T}} \tag{2.8}$$

The partition function is used as a normalization factor to ensure that the probabilities sum up to 1.

2.2 Energy Evaluation by Force Fields

Force field methods are an efficient way to calculate the potential energy for a given conformation of even very large atomic systems. In contrast to the very time-consuming and computationally intensive quantum mechanical methods, they do not consider electrons as individual particles but approximate the electronic energy by a parametric function of the nuclear coordinates. This means that the atomic movements are treated by Newtonian mechanics, the so-called *molecular mechanics* [42]. In molecular mechanics, atomic nuclei and electrons are merged into point-like force centers and covalent bonds are represented by springs of different stiffness. Thus, the bonded interactions, i.e. the stretching, bending, and improper torsion of the bonds (shown in Fig. 2.1) can

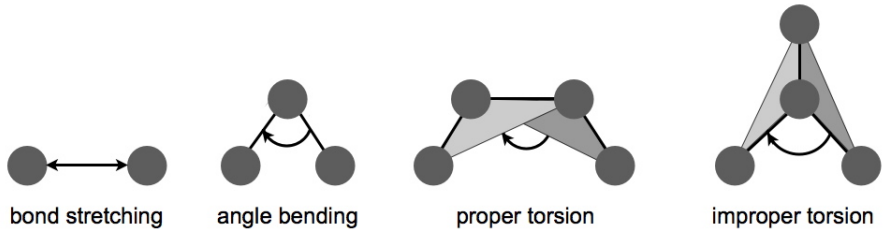


Figure 2.1: In molecular mechanics, atoms are represented by balls and bonds by springs. This allows describing the bonded interactions (except for proper torsions) by Hooke's law.

be described by Hooke's law. It is assumed that the values of these interactions fluctuate around equilibrium values and the magnitude of these fluctuations is characterized by the corresponding spring force constants. As an exception, proper torsions are modeled as sums of suitable cosine functions. Interactions between pairs of non-bonded atoms i and j with a distance of r_{ij} are described by a Lennard-Jones potential and an electrostatic potential following Coulomb's law. The Lennard-Jones potential defining the van der Waals interaction is defined as

$$U_{vdW}(ij) = 4\epsilon_{ij} \cdot \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) \tag{2.9}$$

where σ_{ij} denotes the separation of the two atoms for which the potential is zero and ϵ_{ij} denotes the energy minimum (well depth) of this potential. The electrostatic interaction between two atoms with charges q_i and q_j is given by

$$U_{ES}(ij) = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r} \tag{2.10}$$

with ϵ_0 being the electric constant and ϵ_r being the relative dielectric constant. In this context, a *force field* is a set of parameters and functions derived from quantum mechanical calculations and experimental data that is used to describe the potential energy of a molecular system. All atoms are assigned a type that defines their partial charge and van der Waals radius, as well as the equilibrium values (b_0 for bond lengths, ϕ_0 for angles, and ξ_0 for improper dihedrals) and force constants k for bonded interactions with other atom types. The potential energy is then approximated by an empirical function U of the three-dimensional coordinates of the system s [42]:

$$\begin{aligned} U(s) &= \sum_{bonds(ij)} \frac{k^{(ij)}}{2} \left(r_{ij} - b_0^{(ij)} \right)^2 + \sum_{angles(ijk)} \frac{k^{(ijk)}}{2} \left(\phi_{ijk} - \phi_0^{(ijk)} \right)^2 \\ &+ \sum_{improper\ torsions(ijkl)} k^{(ijkl)} \left(\xi_{ijkl} - \xi_0^{(ijkl)} \right)^2 \\ &+ \sum_{proper\ torsions(ijkl)} \sum_{n=0}^{N} \frac{k^{(ijkl)}}{2} \left(1 + \cos(n^{(ijkl)} \tau_{ijkl} - \tau_0^{(ijkl)}) \right)^2 \\ &+ \sum_{pairs(ij)} \left(U_{vdW}(ij) + U_{ES}(ij) \right) \end{aligned} \tag{2.11}$$

As above, r_{ij} denotes the distance between two atoms i and j, ϕ_{ijk} denotes the angle between three atoms i, j, and k, ξ_{ijkl} the improper, and τ_{ijkl} the proper dihedral angle between four atoms i, j, k, and l. In the term defining the proper torsions, n is the multiplicity, a value that that gives the number of minima in the function, and $\tau_0^{(ijkl)}$ is the phase factor which determines the equilibrium values of the dihedral angles.

Many different force fields have been developed for different purposes like the application to proteins, DNA or RNA, and small organic molecules. The most commonly used force fields for proteins are AMBER [43], CHARMM [44], GROMOS [45], and OPLS-AA [46]. However, one should keep in mind that the potential energy calculated by a force field is just an approximation and has no physical meaning. Its correctness highly depends on the parameterization and the assignment of the correct atom types. Furthermore, as the electrons are not considered explicitly, the effects of delocalized π-electron systems and of polarizability are neglected and the application is limited to systems in the electronic ground state. A long-range interactions cutoff is usually used to speed-up the calculations of the non-bonded interactions. In the molecular dynamics simulations of this work, long-range electrostatic interactions were considered by the Particle-Mesh-Ewald method [47].
Force field energies can be considered as steric energies because they measure the excess energy relative to a hypothetical molecule (where all bond lengths, angles, and torsions are at their equilibrium values) without non-bonded interactions. The potential energies calculated for chemically different molecules use different terms (due to different atom types, bonds, etc.) and, thus, their zero points differ from each other. In other words, their energies cannot be compared [41]. However, when considering conformers of the same biomolecular system, the use of force fields is highly recommended. They can even be used to approximate the *potential energy surface* of the system, a $3N - 6$ dimensional hypersurface that is defined by the potential energy of all possible conformers of a system with N atoms [42]. Therefore, force fields are widely used to minimize the energy of a protein, to search for multiple energetically favorable conformations, the so-called *conformational sampling*, and to score docking complexes.

2.3 Conformational Sampling

When dealing with experimental protein structures, one should always keep in mind that proteins are flexible molecules and that their dynamics cannot be described by a single conformation. Especially proteins in solution exist as an ensemble of energetically accessible conformations and so their flexible structure is best described when capturing as many different representative conformations as possible. Besides, X-ray or NMR structures represent time-averaged coordinates

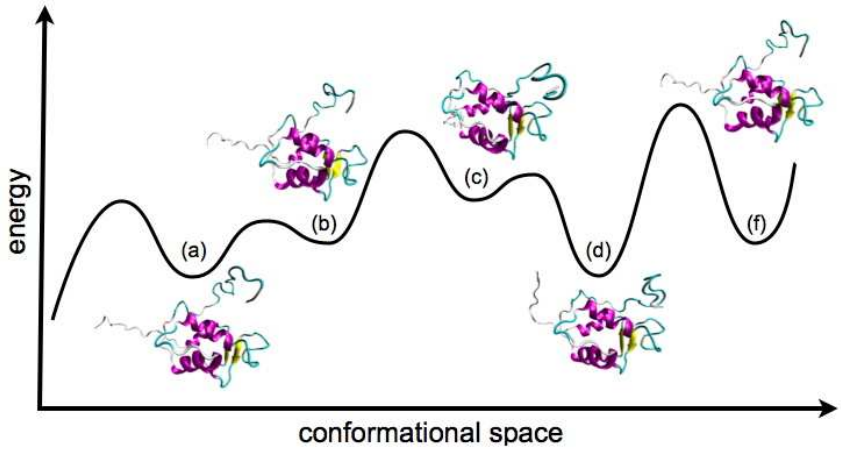


Figure 2.2: Proteins exist in many different conformations arising from e.g. displacements of secondary structure elements. According to the Boltzmann distribution, their population is dependent on the free energy of the various conformations. Here, (a) and (d) represent the most favorable conformations and (c) the least favorable one.

and are often derived under non-physiological conditions (e.g. in crystals instead of solution, at low temperature, at too low or too high pH values) [48]. They represent only one out of many different possible conformations the protein may adopt. However, not all conformations have the same energy as depicted in Figure 2.2 and, thus, not the same probability of occurring as defined in equation 2.8. Therefore, conformations that correspond to low-energy states of the protein are more frequently observed than conformations of high energy. This finding reduces the search space when looking for dominant protein conformations. Conformational sampling can thus be interpreted as searching for low-energy protein conformations [49]. A large number of conformational sampling methods for proteins have been developed. Systematic search methods scanning the complete or a significant fraction of the conformational space can only be applied to small molecules having a few degrees of freedom. When sampling the conformational space of proteins, heuristic search methods have to be applied that consider only a tiny fraction of the conformational space but aim at generating a conformational ensemble that is as representative (in the Boltzmann weighted sense) as possible [49, 50]. Such methods can be roughly divided into the following types:

- Non-step methods generating conformations that are independent from each other (e.g. (t)CONCOORD, NMA, methods that sample side-chain rotamers)
- Step methods generating a new conformation from the previous one (e.g. Monte Carlo methods, molecular dynamics simulations)

In the following, the conformational sampling methods that are of importance for this thesis are shortly introduced.

2.3.1 Molecular Dynamics Simulations

Classical molecular dynamics (MD) simulations calculate the time-dependent behavior of a molecular system, so-called *trajectories* [42]. It is a very powerful and complex technique and conformational sampling is only one possible application. The system may be simulated in different thermodynamic ensembles:

- the *micro-canonical (NVE) ensemble* (constant number of atoms N, volume V, and energy E)
- the *canonical (NVT) ensemble* (constant number of atoms N, volume V, and temperature T)
- the *isothermal-isobaric (NpT) ensemble* (constant number of atoms N, pressure p, and temperature T)

The new configuration of the system consisting of N interacting atoms i with coordinates s_i and mass m_i at time step t is calculated from the previous configuration by integrating Newton's law of motion

$$m_i \frac{\delta^2 s_i}{\delta t^2} = F_i \tag{2.12}$$

where the forces F_i are the negative derivatives of the potential function $U(s_1, \ldots, s_N)$:

$$F_i = -\frac{\delta U}{\delta s_i} \tag{2.13}$$

These equations have to be solved for each atom in small time steps (usually 1-2 fs). This is normally done using *Verlet methods* like the *leap-frog algorithm*. The forces $F(t)$ computed from

the coordinates s at time t are used to update the velocities of the atoms v at time $t + \Delta t$. In order to obtain more accurate values, mid-step velocities at time $t + \frac{\Delta t}{2}$ are calculated by

$$v(t + \frac{\Delta t}{2}) = v(t - \frac{\Delta t}{2}) + \frac{F(t)}{m}\Delta t \tag{2.14}$$

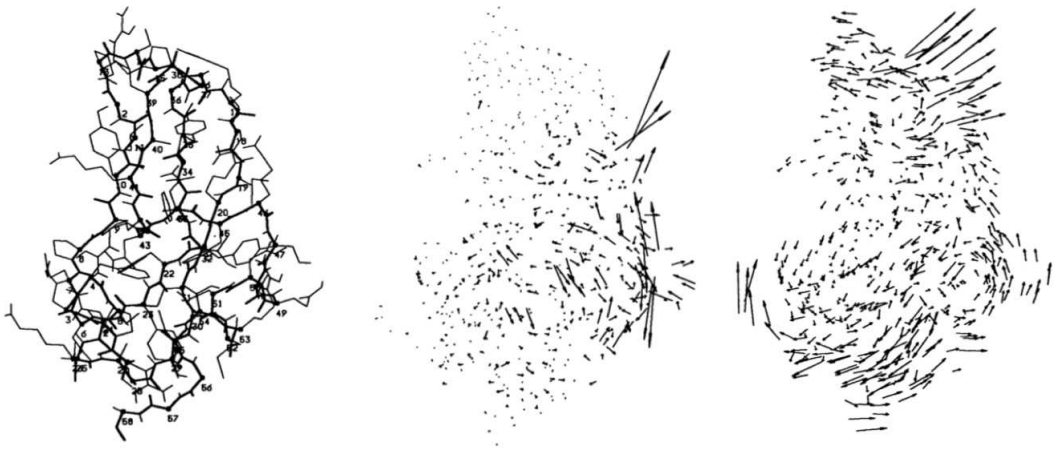
The new velocities are then used to update the coordinates s:

$$s(t + \Delta t) = s(t) + v(t + \frac{\Delta t}{2})\Delta t \tag{2.15}$$

In the classical setup the molecule(s) are placed in a solvent box. As the number of atoms may be very large, especially when using explicit solvent molecules, several approximations in addition to those related to the use of force fields are needed. Treating bonds as constraints instead of oscillators in the equation of motion allows the use of larger time steps [51]. Furthermore, the behavior of the system at the boundary of the simulation box may be unnatural. Thus, periodic boundary conditions are used to simulate a bulk system without real phase boundaries.
Although MD simulations of biomolecular systems are computationally very expensive, they may nowadays be extended to multiple microseconds of simulation time, depending on the system size [52]. Note that only those states may be sampled that occur on time scales comparable to the simulation length.

2.3.2 Normal Mode Analysis

Another very-well established method for studying conformational changes in biomolecules is *normal mode analysis* (NMA) [53, 54]. Like in MD simulations, a force field is used to calculate the potential energy of particular conformations. The virtue of this technique is that it can identify the inherent collective modes along which overall protein dynamics takes place, whereas other methods like MD simulations sample the protein dynamics only along coupled modes. These so-called *normal modes* are linear independent concerted motions of atoms that oscillate with the same frequency around a local energy minimum (see Fig. 2.3 for an example). For calculating



(a) (b) (c)

Figure 2.3: Normal mode analysis of the bovine pancreatic trypsin inhibitor protein (a) identified displacement vectors of atoms (shown as arrows) oscillating with the same frequency. Here, the normal modes with a frequency of 118.8 cm^{-1} (b) and 6.9 cm^{-1} (c) are illustrated (Figure taken from [53]).

the normal modes, it is assumed that the potential energy function around a local minimum is harmonic and that the normal modes represent harmonic vibrations around these energy valleys. This strong assumption about the molecule's harmonic behavior requires a preceding exhaustive energy minimization. Given an energy minimized system of N atoms, the $3N$ x $3N$ mass-weighted Hessian matrix H is calculated by

$$H_{ij} = \frac{\delta^2 U}{\delta s_i \delta s_j \sqrt{m_i m_j}} \tag{2.16}$$

where m denotes the mass, U the potential energy function, and s_i and s_j the atomic x, y, or z coordinates. The calculation of the normal modes can then be reduced to an eigenvalue problem

$$H\Psi = \Psi\Omega \tag{2.17}$$

where Ψ is a matrix containing the eigenvectors of H as columns and Ω is the diagonal eigenvalue matrix. The eigenvectors are the normal modes σ_i that contain the amplitude and direction of motion for each atom and the eigenvalues are the corresponding squared frequencies of oscillation ω_i^2. The root mean-square fluctuation (RMSF) of a normal mode is then given by

$$\sigma_i = \sqrt{\frac{k_B T}{\omega_i^2}} \tag{2.18}$$

Normal modes with lowest frequencies result in delocalized motions involving more distant parts of a protein, i.e. oscillations of larger amplitude where a large number of atoms is involved. In contrast, localized motions like bond stretching result from normal modes with higher frequencies [53–55]. Hence, conformational changes (e.g. induced-fit effects) can be described by a linear combination of displacements along the eigenvectors. Likewise, new conformations can be sampled by random displacements along the eigenvectors.
However, one should always keep in mind that NMA is based on the strong assumption that all vibrations are harmonic. If this was to be true, any molecular motion could be exactly expressed as a linear combination of normal modes. But at 300 K, many vibrations are anharmonic and, thus, not all protein dynamics observable during MD simulations or in experiments can be described by a superposition of normal modes. Yet, it has been shown for several proteins that functionally important conformational variations, like conformational changes upon ligand binding, can be described by a single or multiple low-frequency normal modes [56, 57]. But note that as NMA is very memory-intensive, the energy calculations are performed in vacuum and, thus, solvent effects are neglected in the analysis of the protein dynamics.

2.3.3 CONCOORD and tCONCOORD

MD simulations and NMA suffer from the fact that high-energy barriers are hard to overcome and the sampling is restricted to a local energy basin of the potential energy surface. CONCOORD (CONstraints to COORDinates) [58] and its extension named tCONCOORD [59] are efficient methods avoiding this problem by generating random protein conformations that fulfill previously determined distance bounds. The method consists of two steps illustrated in Figure 2.4. At first, all pairwise interatomic distances in the starting structure (usually an energy minimized experimental structure) are measured and classified by the program *dist*. For each atom pair the distance range is then set to the measured value plus or minus a tolerance value that reflects the strength of this interaction. For example, the allowed distance ranges for covalently bonded pairs are quite small, whereas for atom pairs with hydrophobic interactions larger deviations from the distance observed in the starting structures are accepted. In the second step, the program *disco* tries to

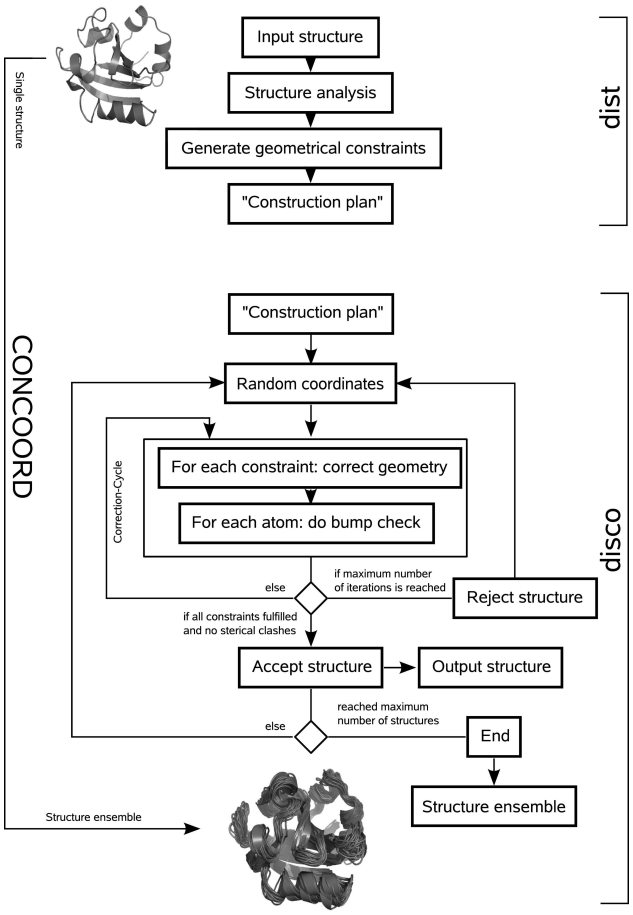


Figure 2.4: The (t)CONCOORD algorithm (Figure taken from `http://www.mpibpc.mpg.de/groups/de_groot/dseelig/tconcoord.html`).

find conformations that fulfill all these distance constraints. The procedure starts from random coordinates and iteratively applies corrections to the positions of all atoms that violate the distance constraints until all violations are removed. The authors showed that the generated conformations are similar to those obtained from MD simulations [58].
Whereas the goal of CONCOORD is the generation of a conformational ensemble around a given starting structure, the extension tCONCOORD tries to predict conformational transitions and, thus, is able to sample the conformational space more exhaustively. The main difference to the original implementation is the estimation of the stability of hydrogen bonds by analyzing the environment with respect to hydrophobic protection. Only stable hydrogen bonds that are not likely to be broken by water molecules are translated into distance constraints. Interestingly, when starting from the apo protein structure, this extension is able to generate conformations that are very similar to experimentally determined ligand-bound structures [59].

2.3.4 Sampling Side-Chain Rotamers

When considering the conformational space of the protein side-chains only (e.g. homology modeling with fixed backbone conformation) one can discretize the search space into so-called side-chain *rotamers*. Analyzing available protein structures revealed that the dihedral angles $\chi_1, \ldots, \chi_4$ of the side-chains tend to cluster around particular values representing low-energy side-chain conformations (see Figure 2.5) [60]. This observation led to the compilation of *rotamer libraries*

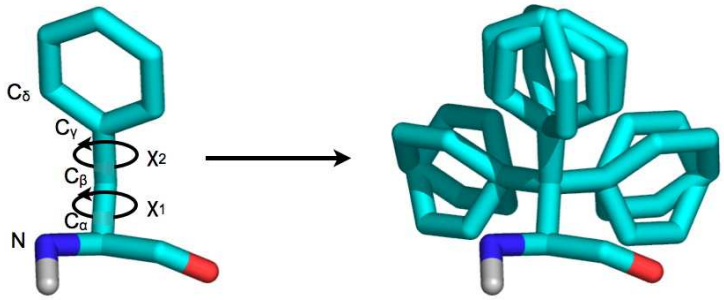


Figure 2.5: Phenylalanine has two χ angles: χ_1 is the torsion N - C_α - C_β - C_γ and χ_2 is the torsion C_α - C_β - C_γ - C_δ. The rotamers, the energetically favorable side-chain conformations, can be represented by a combination of different χ_1 and χ_2 angles.

listing the observed rotamers (defined by the χ-angles) for each amino acid together with the probabilities of their occurrence [61, 62]. As these *backbone-independent* rotamer libraries do not take the possible relation between a χ-angle and the local backbone conformation (i.e. the secondary structure) into account, a second type of rotamer libraries was introduced, the *backbone-dependent* rotamer libraries. Here, rotamers are defined as a function of the backbone dihedral angles ψ and ϕ [63]. An example for the usage of rotamer sets is the side-chain prediction problem where the protein backbone is fixed and the correct rotamer has to be assigned to each residue such that the total energy of the resulting protein conformation is minimal. This is an important task in homology modeling, ab initio protein structure prediction, and structure-based drug design [64]. The sampling of side-chain rotamers is a combinatorial problem and even for small proteins with 100 residues having an average number of 5 rotamers per residue, 5^{100} different conformations are possible. However, many combinations of rotamers cannot exist in the same low-energy conformation. When calculating the global minimum-energy conformation (GMEC), the Dead-End Elimination (DEE) theorem [65] can be used to reduce the search space by removing all rotamers that cannot occur in the GMEC. The program *SCWRL* uses DEE to restrict the conformational space given by a backbone-dependent rotamer library such that the remaining space can be searched exhaustively for low-energy side-chain combinations [66]. The drawback of *SCWRL* and other methods using DEE is that the GMEC highly depends on the rotamer library and the potential energy function used. By keeping the best rotamer per residue only, many other low-energy conformations are ignored. However, in many applications it is beneficial to consider an ensemble of low-energy conformations instead of just a single one. The program *IRECS*, for example, handles side-chain flexibility by calculating several energetically favorable rotamer combinations [67].

2.4 Detection of Binding Sites on Protein Surfaces

Characterizing the surface of the studied protein is crucial for understanding and predicting its function. As the function of most proteins is closely related to binding specific partners, this is also reflected in the properties of their molecular surfaces. Only if the interface possesses the requisite complementarity, binding will occur with the required affinity [68]. This complementarity is provided by physicochemical and sterical features. Therefore, the ubiquitous question in structure-based drug design, "Does protein A bind molecule B with sufficient affinity?" can be broken down into two smaller questions: "Which surface region of protein A is complementary to molecule B?" and "How does the complex AB look like?" This section introduces computational techniques for answering the first question, the second question is addressed in the next section.

If B is a small molecule, its binding site on protein A will most likely contain a binding pocket. Many algorithms have been presented that aim at identifying pockets on protein surfaces [18, 69–73]. Due to their focus on concave regions, these methods are usually *geometry-based*. More general methods use *energy-based* approaches for predicting surface regions that are endowed with physicochemical features that may account for high affinity ligand binding [74–76]. Here, only a digest of the established methods is presented.

2.4.1 Geometry-based Detection of Binding Sites

Algorithms of this category detect cavities on protein surfaces. The advantage of this purely geometrical definition of a binding site is the independence from the ligand's physicochemical properties. One typically assumes that the ligand binds into one of the largest pockets available on the protein surface. However, several studies showed that the ligand indeed binds into the largest pocket in 72 % to 84 % of the complexes in the used data sets [18, 77] suggesting that not only the largest cavity, but also the smaller ones are of interest. Pocket detection methods using geometric criteria can be further subdivided into *grid-based* approaches where the protein is mapped onto a

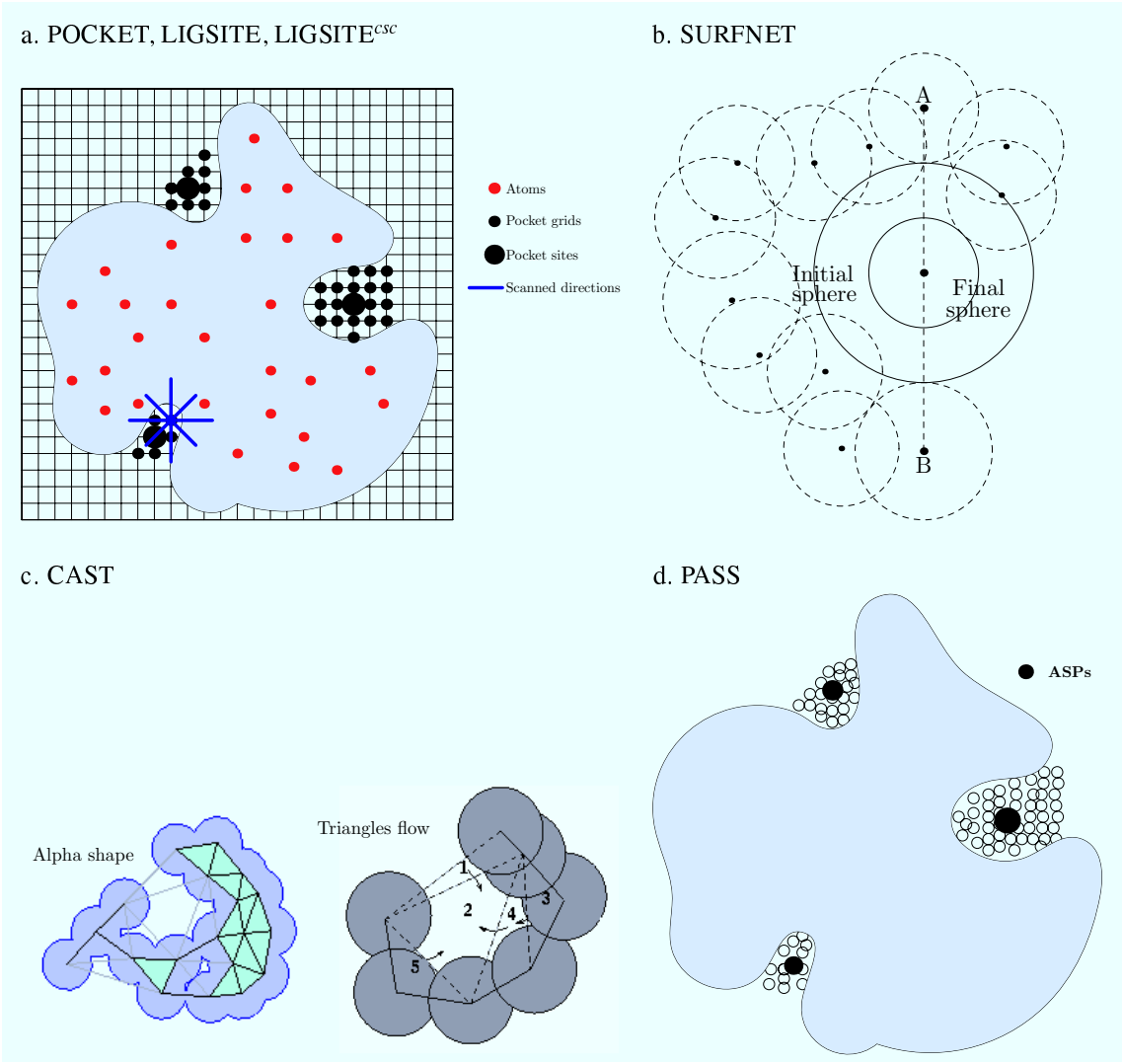


Figure 2.6: Geometry-based algorithms for the detection of pockets (Figure taken from [69]).

3D grid (Fig. 2.6 (a)) and *non grid-based* approaches (Fig. 2.6 (b) - (d)).
Prominent examples for grid-based approaches are:

- POCKET [72] scans the x, y, and z axes for a sequence of grid points that starts and ends with a point inside the protein and has a period of solvent grid points in between.
- LIGSITE [73] extends the POCKET algorithm by additionally scanning the four cubic diagonals.
- LIGSITE^{CSC} [69] refines LIGSITE by scanning for surface-solvent-surface instead of protein-solvent-protein events and re-ranks the pockets by the degree of conservation of the involved surface residues.
- Pocket-Finder [74] extends the LIGSITE algorithm by setting a threshold for the minimal number of protein-solvent-protein events.

While these method are all based on the same idea, the methodologies of the non grid-based approaches differ significantly from each other. The most important ones are:

- PASS [70] uses probe spheres to incrementally fill pockets.
- SURFNET [71] generates a set of interpenetrating spheres that are placed between two atoms and do not contain any other atoms.
- CAST [18] uses alpha shapes and discrete flow theory to compute pockets.

In this thesis the PASS algorithm is used. In the following, it is introduced in more detail.

PASS

PASS (Putative Active Sites with Spheres) [70] uses probe spheres for characterizing regions of buried volume on protein surfaces. Based on the size and burial extent (i.e. number of protein atoms within a given radius) of these volumes, the approach identifies positions likely to represent

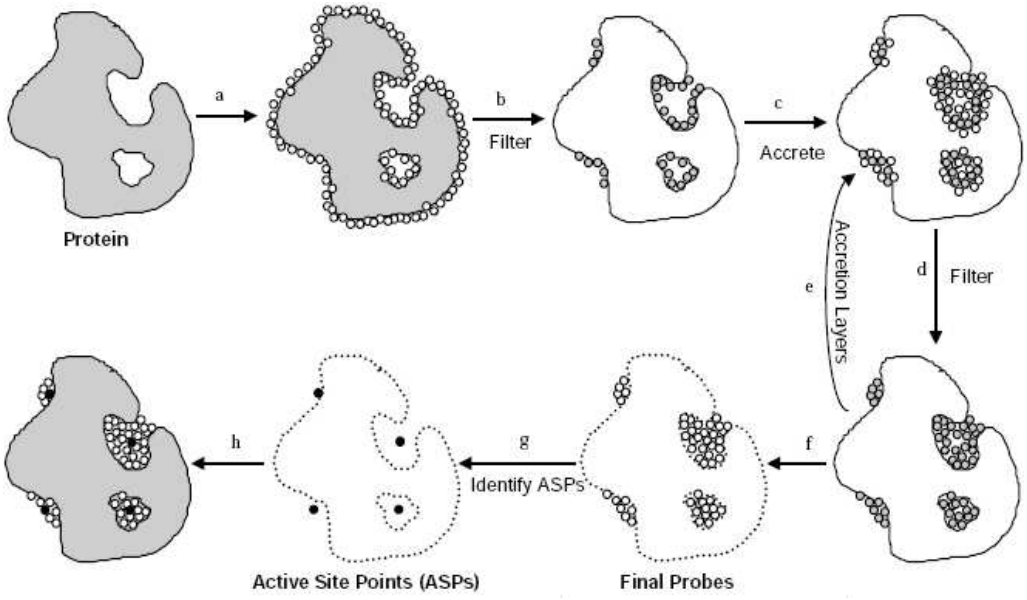


Figure 2.7: The PASS algorithm (Figure taken from [70]).

binding sites. An outline of the algorithm is shown in Figure 2.7. Given a protein in PDB format, PASS assigns the elemental atomic radii. The actual pocket screening process starts with coating the protein surface with an initial layer of spherical probes (radius: 1.8 Å) according to a three-point geometry (a). Subsequently these probes are filtered (b) to remove all probes that

- clash with protein atoms,
- are not sufficiently buried (burial count below a given threshold), or
- lie within 1 Å of a more buried probe

Afterwards accretion layers of smaller probes (radius: 0.7 Å) are added onto the previously identified probes (c) and are filtered (d) as described in step (b). These two steps are repeated (e) until a layer is encountered in which no newly-found probes survive the filters (f). For each probe in this final set of probes a weight is calculated that reflects the number of probes in the vicinity and their burial extent. The active site points (ASPs) are determined by cycling through all probes in descending order of their probe weight, considering only those probes with a weight above a given threshold that are separated by a minimum distance of 8.0 Å from the previously identified ASPs (g). The output of the algorithm is the placed probes and the identified ASPs representing the potential binding sites (h).
Initially, we have used the binary executable file made available by the developers of PASS at `http://www.ccl.net/cca/software/UNIX/pass/overview.shtml`. At a later stage, we shifted to the BALLPass implementation by Jan Fuhrmann and Dirk Neumann (CBI, Saarbrücken).

2.4.2 Energy-based Detection of Binding Sites

Alternatively to the geometry-based algorithms, binding sites can also be detected by energetic criteria. Here, it is assumed that the ligand binds to the energetically most favorable site on the protein surface [78]. In principle, the common idea of these approaches is to calculate the interaction energy between small probes and the protein. The variety of used probes ranges from single methyl probes to sets of many different organic probes and ions. Note that the more different probe types are used the more detailed is the characterization of the putative binding sites (e.g. information about potential hydrogen bond donors and acceptors, electrostatic, hydrophobic, or aromatic interactions).
In this class of algorithms, binding sites are not (solely) predicted by cavity detection, the physicochemical properties of these surface regions that may account for high-affinity ligand binding (i.e. ligand complementarity) are also considered. Many algorithms are grid-based, so that the interaction energy between the molecular probe and the protein is calculated for each grid point. Other approaches use the grid just for the initial placement of the probes and optimize their positions by energy minimizations. Examples are:

- GRID [79] calculates the interaction energy between a chosen probe and the protein at each grid point of a cubic grid placed onto a region of the protein or onto the entire protein.
- Q-SiteFinder [74] determines the van der Waals energy between a methyl probe and the protein for each grid point and detects pockets by clustering the most favorable ones.
- CS-Map [75] uses a grid for the initial placement of small organic functional groups, moves them around to minimize the interaction energy with the protein atoms, and finally clusters and ranks the positions to predict hot spots for binding of drug-like molecules.
- MCSS [76] requires a prior definition of the binding site. The strength of this algorithm is that protein flexibility is taken into account while favorable positions and orientations for small organic functional groups are predicted.

2.5 Molecular Docking

Knowing the potential site of interaction between protein and ligand does not directly allow for inferring the binding affinity and the conformation of the complex. The binding affinity depends on the *binding mode* of the ligand, i.e. its bound conformation and its orientation relative to the protein. The problem of predicting a protein-ligand complex is tackled by *molecular docking*. Docking can be described as a combination of two components: searching for favorable configurations and scoring them. The output is a list of predicted protein-ligand complexes, the *docking poses*, ranked by their *docking score* that evaluates the affinity of the complexes. Both components are crucial, though error-prone. The configuration space that has to be sampled is huge. Even if the ligand as well as the protein (or *receptor* as they are often called within this context) are rigid, six degrees of freedom have to be considered. But ligands tend to possess several rotatable bonds and so not only the ligand's orientation relative to the protein has to be sampled, but also the internal degrees of freedom of the ligand. Moreover, when treating the protein as a rigid body, induced-fit effects are neglected. Yet the size of the search space impedes a complete sampling and so the native binding mode may be missed. But even with a complete list of possible configurations, finding the native complex is not guaranteed. Scoring functions predict which docking poses occur in nature. They may be, for example, force-field based, knowledge based, or empirical and usually include weights for individual terms that have been fitted using a training set of protein-ligand complexes. However, one cannot expect them to work perfectly for each kind of protein-ligand interaction [80, 81].
Molecular docking is nowadays a crucial component of many drug discovery projects. Note that docking approaches always represent a compromise between exactness and computational feasibility. Especially when applied early in a drug discovery project, i.e. when docking is used in virtual screening to identify potential hits among several thousands of compounds, speed is an important issue although the predicted affinities should still be reliable. On the other hand, when applied in the lead optimization phase, where the number of putative ligands has significantly decreased and the objective is a reliable differentiation of their predicted binding affinities, exactness is more important. One possibility to evaluate the reliability of a docking program is *re-docking*, where a ligand is docked back into the protein conformation taken from its experimentally known complex structure. In a perfect scenario, a docking pose that corresponds to the native complex would be scored best and, thus, ranked number 1. The similarity of a docking pose to the native complex is measured by the root mean square deviation (RMSD) of the involved atoms (usually the atoms of the ligand). A docking pose with an RMSD ≤ 2 Å is considered native-like. Note that if only docking poses with high RMSD values were calculated, this may indicate an insufficient sampling. If docking poses with low RMSD values are exclusively predicted to be unfavorable, this may hint at an unsuitable scoring function [10, 11, 80, 81].
While early docking algorithms like DOCK [82] considered the protein and the ligand as rigid, more and more flexibility has been incorporated in recent years. Nowadays, treating all rotatable bonds of ligands as flexible is standard in modern docking algorithms and even receptor flexibility is handled more and more successfully [48]. In the following, a few popular docking software packages are introduced.

2.5.1 Docking Flexible Ligands into Rigid Receptors

In general, ligands change their conformations upon binding to a receptor. Although this results in a loss of degrees of freedom and so in a free energy penalty in the order of 0.4 kcal/mol per torsion [83], this energy increase is compensated by interactions between protein and ligand and solvent reordering so that the total free energy of binding is favorable. The magnitude of the conformational changes varies between complexes. As about 70% of all drug-like molecules possess

2-8 rotatable, non-terminal bonds [84], handling at least ligand flexibility is pivotal for a successful prediction of their binding modes. Two common approaches exist for sampling binding poses that incorporate flexible ligands: fragmentation of the ligand molecule followed by an incremental reconstruction and global optimization of the ligand coordinates. The best known docking tools are:

- The package DOCK [85] considers ligand flexibility since version 4.0. The algorithm starts with filling the binding site with overlapping spheres that represent clusters of pseudo-atoms. The ligand is divided into fragments and a rigid portion, the "anchor" is superposed onto these pseudo-atoms by geometric matching. The anchor positions are then evaluated and energetically minimized using a force-field based scoring function which incorporates intermolecular energy terms that were precomputed on a grid. The best initial docking results are selected and the remaining ligand fragments are incrementally added in different orientations, optimized by a short energy minimization, and pruned such that only a predefined number of partial binding configurations has to be considered in the next step.

- FlexX [86] divides the ligand into fragments and chooses a base fragment that is then placed at several promising positions in the binding pocket, independently of the rest of the ligand. Subsequently, the ligand is incrementally reconstructed using a greedy strategy. Interactions are classified by the strength of their geometric constraints and described by spherical surfaces whose parameters depend on the type of interaction. The empirical scoring function estimates the free binding energy of the protein-ligand complexes by penalizing deviations from the ideal geometry for hydrogen bonds, ionic, aromatic, and lipophilic interactions.

- Glide [87] makes use of a "docking funnel" for progressively narrowing down the search space and so allowing for more accurate scoring functions. In a preprocessing step, sets of fields that represent the shape and properties of the receptor on a grid are computed. After an initial sampling of ligand conformations, promising poses are selected and the ligand is energetically minimized in the field of the receptor using a force-field based energy function. The conformations of the very best candidate poses are refined using a Monte Carlo sampling. Finally, the docking poses are re-scored and ranked using a more accurate scoring function combining force-field and empirical based terms.

- AutoDock [88] uses a genetic algorithm (GA) to optimize the ligand coordinates with respect to the protein and evaluates the docking poses by an empirical scoring function. As this is the docking method used in this thesis, it is introduced in detail in the following paragraph.

AutoDock3

Before the actual docking step of AutoDock3 [88] is launched, a grid of user-defined size and spacing is placed at the binding site. At each grid point, the electrostatic potential, as well as the interaction energy between the protein and the different atom types available in the ligand are precomputed. The docking step itself then tries to optimize the interaction energy between the protein and the ligand. To this end, several search procedures are provided. The most prevalent is the Lamarckian genetic algorithm (LGA) in which the translation, rotation, and conformation of the ligand with respect to the protein are coded by *state variables*. In the context of genetic algorithms, these state variables correspond to genes that define the ligand coordinates. The *fitness* of a ligand is calculated by the AutoDock3 scoring function that is based on force field energies and tries to estimate the free binding energy of the complex in solution by implementing the

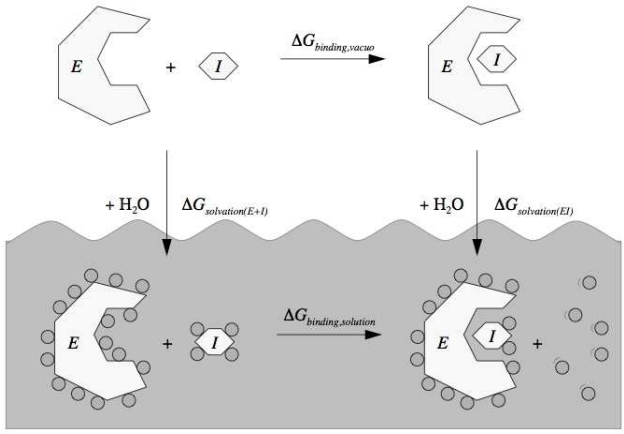


Figure 2.8: The AutoDock3 scoring function estimates the change of free energy upon binding in solvent. As the change in free energy is independent from the path, the thermodynamic cycle can be used to derive $\Delta G_{binding,solution}$ (Figure taken from the AutoDock3 manual).

thermodynamic cycle shown in Figure 2.8:

$$\Delta G_{binding,solution} = \Delta G_{binding,vacuo} + \Delta G_{solvation(EI)} - \Delta G_{solvation(E+I)} \tag{2.19}$$

The scoring function is given by

$$\begin{aligned} \Delta G &= \Delta G_{vdW} \sum_{ij} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\ &+ \Delta G_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ &+ \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \\ &+ \Delta G_{tor} N_{tor} \\ &+ \Delta G_{sol} \sum_{i,j} (S_i V_j + S_j V_i)\, e^{-r_{ij}^2/2\sigma^2} \end{aligned} \tag{2.20}$$

The ΔG terms are coefficients fitted by a linear regression analysis of a set of 30 protein-ligand complexes of known binding affinities. The summations run over all pairs of ligand atoms i and protein atoms j, as well as over all ligand atom pairs that are separated by at least three bonds. $E(t)$ denotes a directional weight that is based on the angle, t, between the probe and the target atom and a Coulombic electrostatic potential. Upon binding, the ligand looses several conformational degrees of freedom resulting in a loss of entropy. In this scoring function, the entropy cost is assumed to be proportional to its number of sp^3 bonds, N_{tor}. The last term estimates the *desolvation energy*, the energy gain or loss arising from the removal of solvent molecules from the binding interface of the protein and the ligand. Here, the desolvation energy is estimated by a variant of the method of Stouten [89] that evaluates the percentage of volume around the ligand occupied by protein atoms, where V_i denotes the atomic fragmental volume and S_i the solvation term for atom i, and σ a gaussian distance constant.

After assessing the fitness of each *individual*, the best ones pass their genes on to the next generation, where they are recombined, randomly mutated, or left unchanged. In the LGA, this global search implemented in the genetic algorithm is combined with a local search that performs an energy minimization on the atomic coordinates and is applied to a user-defined fraction of individuals of each generation. After a predefined maximum number of energy evaluations or generations, the

best individuals corresponding to docking solutions are reported. Usually several independent docking runs are performed and the different solutions are clustered by their RMSD.

2.5.2 Docking Flexible Ligands into Flexible Receptors

Not only ligands, but also proteins undergo conformational changes upon binding. Here, the trade-off between efficiency and accuracy is even more an issue as considering conformational changes of the receptor in addition to the ligand significantly increases the number of degrees of freedom that have to be sampled. These induced-fit effects may comprise subtle rearrangements of a few side-chains located at the binding site, local adaptions of the backbone, or even movements of whole protein domains [90]. Not surprisingly, docking programs handling only ligand flexibility are quite successful for many molecular systems, while for others, they completely fail in predicting the protein-ligand complex. For example when re-docking ligands into their receptors in the bound conformation, the docking program GOLD that handles ligand and receptor flexibility was able to find a docking pose within 2 Å of the native conformation for 91% of the test set. But when another conformation of these receptors is used, the docking performance significantly dropped to 72% [91]. This observation highlights the need for docking protocols that incorporate receptor flexibility. Actually, there are two possibilities to fulfill this requirement. The receptor flexibility may either be represented by a conformational ensemble generated by an external program or by considering different crystal structures, or by sampling conformational changes directly during the search step in the docking program itself. For completeness, we also mention a third possibility for handling receptor flexibility implicitly, the so-called *soft docking* approaches. They represent the simplest way for tackling sterical clashes arising from missing induced-fit effects when a ligand is docked into a rigid receptor. Instead of changing the receptor conformation, the docking pose is scored optimistically by tolerating an overlap of the ligand with the receptor surface. To this end, the repulsive contributions to the energy function are reduced or the van der Waals radii of the receptor atoms are scaled down. Yet, this approach will only yield accurate docking results if subtle side-chain rearrangements in the binding site are sufficient for accommodating the ligand [50].
Most docking approaches introduced previously have actually been extended to model receptor flexibility. The underlying changes will be discussed in the following overview of docking protocols that incorporate receptor flexibility.

Representing Receptor Flexibility by Conformational Ensembles

The maybe most trivial solution is to dock the ligands against every receptor conformation taken from a conformational ensemble. In this case, each individual conformation can be treated as rigid during the actual docking step. The conformations may be derived from different experimental structures of the protein, extracted from MD simulations, or from any other method for conformational sampling. The advantage of this kind of protocol is its modularity, i.e. sampling receptor flexibility and docking poses are independent from each other. Any conformational sampling approach can be combined with any docking program and the conformations to be considered can be selected beforehand. Moreover, the degree of receptor flexibility is unlimited. It is noteworthy that these protocols are rather based on the theory of conformational selection than on the theory of pure induced-fit effects [48]. The drawback of this kind of handling receptor flexibility is that the quality of the docking results crucially depends on the used conformations. If relevant conformations are lacking, e.g. if no appropriate binding pocket is available, then the best docking program will not be able to predict a reasonable binding pose. Moreover, the computational demand may be quite high when using a large conformational ensemble. On the other hand, it was shown that the ligand may bind to receptor conformations which are not highly populated [93, 94], suggesting

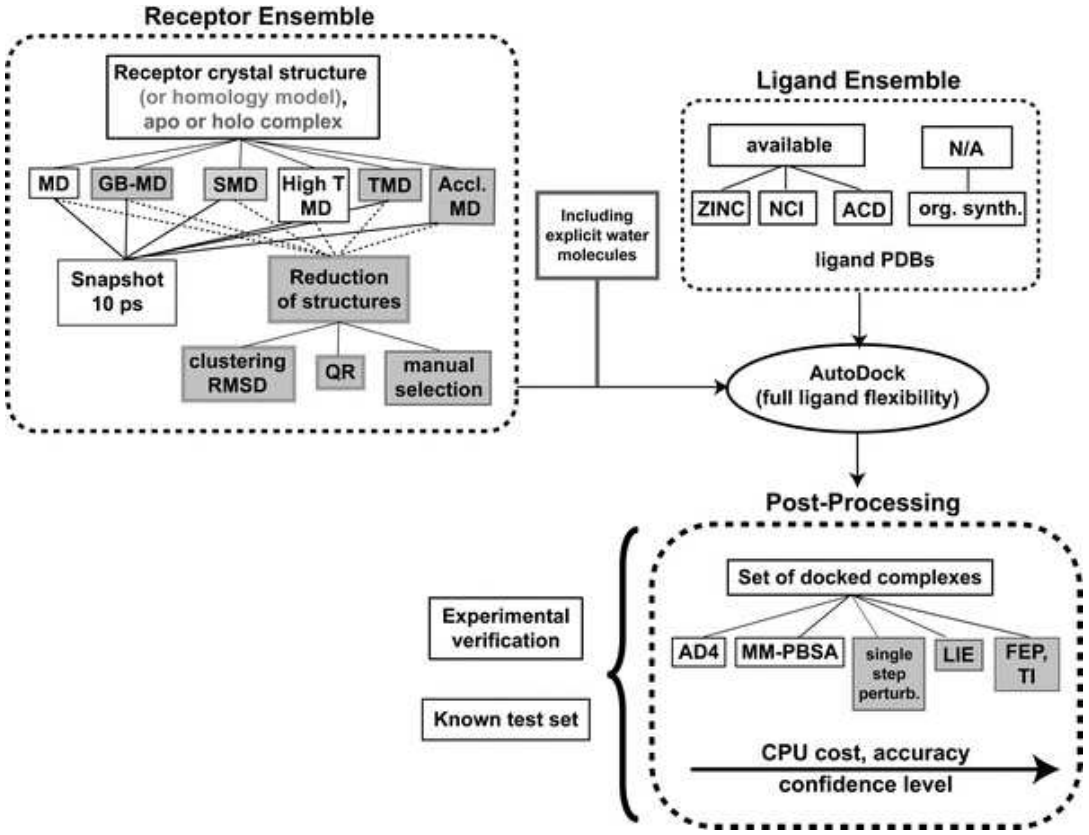


Figure 2.9: An overview of the improved relaxed complex scheme. The improvements to the original RCS are shown in gray background. The receptor ensembles can be generated by classical MD simulations or by simulation techniques that enhance the conformational sampling like Generalized-Born MD (GB-MD), steered MD (SMD), high temperature MD (High T MD), targeted MD (TMD), or accelerated MD (Accl. MD). The ligands are taken from existing or newly assembled databases and docked with AutoDock into the receptor ensemble. The docking poses are then re-scored or re-evaluated in the Post-Processing stage using the AutoDock version 4.0 scoring function (AD4), molecular-mechanics Poisson-Boltzmann surface area (MM-PBSA), single step perturbation, LIE, FEP or TI technique (Figure taken from [92]).

the importance of preselecting conformations that may contain eligible binding sites. For example, the original *relaxed complex scheme* combines MD simulations of the receptor in explicit water with a subsequent rapid docking of the ligands into the extracted MD snapshots, and an accurate re-scoring of the docking poses [93, 94]. In the *improved relaxed complex scheme* shown in Figure 2.9, docking is performed on a reduced set of MD snapshots containing only non-redundant receptor conformations [92].

Alternatively, the complete conformational ensemble is supplied to the docking program and several conformations are combined into one receptor representation during the search step [48]. An example for such a docking approach is FlexE where varying parts are considered as discrete alternative conformations and are combinatorially joined for yielding new receptor representations for docking [95]. The virtue of this more advanced approach is that local receptor conformations providing a favorable contribution to binding can be combined into a single structure suitable for accommodating the ligand. Nevertheless, the tolerated structural differences among the conformational ensemble are limited. Very diverse conformations are difficult to combine and too similar ones are inappropriate for modeling receptor flexibility. In addition, this composite structure may not be a physiological accessible conformation and so the predicted binding modes and energies are artificial [48].

Intrinsic Receptor Flexibility

Including receptor flexibility in sampling putative docking poses can be regarded as a direct implementation of the induced-fit theory. Many docking programs take advantage of the finding that for many proteins only subtle changes in amino acid side-chains are sufficient to achieve a collision-free ligand binding, also known as the "minimal rotation hypothesis" [96]. Hence, a common and time-efficient approach is considering multiple side-chain conformations of a few selected residues located at the binding site by sampling torsion angles or using predefined rotamers taken from a rotamer library [48, 50]. In 1994, Leach presented one of the first docking approaches that handled side-chain flexibility [97]. The residue conformations were taken from a rotamer library and the docking poses were incrementally built up using the A* algorithm (a graph-based method for finding optimal paths from a given initial state to a goal state) in combination with DEE. Another pioneering docking program taking into account receptor flexibility at least partially was GOLD [98, 99] that defines "fitting points" on hydrogen bonding and hydrophobic groups of the ligand and the receptor to place the ligand into the binding site. Similar to AutoDock, GOLD uses a GA for calculating the docking pose. Besides the dihedrals of the rotatable bonds of the ligand, its ring geometries, and the mapping of the fitting points, the conformations of hydrogen bonding terminal bonds of some side-chains are optimized as well. As stated on the GOLD webpage [1], the latest version also features side-chain and backbone flexibility for a few user-defined residues. Likewise, in AutoDock4 [100], the user can define rotatable bonds for a few residues and the side-chain conformations are encoded as genes analogously to the flexible ligand. In contrast, ROSETTALIGAND [101] allows for full side-chain flexibility by optimizing the side-chains and the ligand position simultaneously. In some cases incorporating side-chain flexibility only is not sufficient to model ligand-induced conformational changes. Although backbone movements are often limited to single loops, these local adaptions are not accounted for by side-chain mobility. But even if only the ψ and ϕ angles are varied, considering backbone flexibility will result in a huge search space and an adequate sampling is impossible. Therefore, docking approaches that handle backbone flexibility explicitly have to narrow down the search space further [50]. This can be done efficiently by using harmonic modes (e.g. derived from NMA or MD simulations) to model deformations of the binding pockets [102, 103]. The use of "flexibility trees" [104] is an alternative way for focussing on molecular motions that modify the binding site by encoding the conformational subspace using a small number of variables. They are used by the docking program FLIPDock [105] where a flexible ligand is docked into fully flexible receptors. When sampling the docking poses, the variables representing the receptor flexibility are searched concurrently with the conformation and placement of the ligand. Further approaches comprise several steps, starting with a rough placement of the ligand into the binding site and a subsequent optimization of the docking pose by MD simulations [106] or energy minimizations followed by Monte Carlo minimizations [107]. Other docking protocols combine well-established approaches for sampling, docking, minimization, and scoring. For example, Fleksy [108] is a combination of rotamer sampling, soft-docking using FlexE, refinement of the docking poses using energy minimization, and re-scoring.

[1] `http://www.ccdc.cam.ac.uk/products/life_sciences/gold/`

Chapter 3

Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction

In this chapter our protocol for identifying and analyzing transient pockets will be described. Using this protocol, we could show that the native binding pockets of three protein systems open spontaneously during MD simulations of the apo protein structures in water. This study was published in the *Journal of Medicinal Chemistry* in 2007 [109]. Besides, the three model systems utilized to validate all approaches presented in this thesis will be introduced.

3.1 Introduction

Targeting protein-protein interactions by small molecules is full of challenges (see Table 1.1). The first hurdle is the identification of a favorable binding site. A solution to this problem is straightforward if small-molecule binders have been discovered experimentally and/or the (protein-protein or protein-ligand) complex structure is available. In such a case, the holo protein conformation possessing a binding pocket can be used for virtual screening of ligand libraries. Bowman et al, for example, identified new inhibitors for the MDM2-p53 interaction by docking ligands into a dynamic receptor-based pharmacophore model [36]. However, they used the p53-bound X-ray structure of MDM2 that already possessed a well defined binding pocket at the interface. But how shall we handle proteins for which only apo crystal structures are available? Such structures often lack deep cavities or clearly shaped binding pockets that could be used for identifying binding sites for putative ligands, e.g. by docking. Although docking methods that account for (partial) receptor flexibility have proven to be quite promising even if rigid docking to the apo protein structure failed, they depend on a definition of the binding site [92, 101–103, 105–108] because sampling the entire protein surface would be computationally very costly. Furthermore, the extent of receptor flexibility that can be modeled by these docking protocols may not be sufficient to handle conformational changes at protein-protein interfaces. For example, the side-chains located at these interfaces are on average more flexible than those located in binding pockets. Thus, the structure-based design of small molecules inhibiting protein-protein interactions is generally considered to be challenging. Nonetheless, once a putative binding pocket is available, it can be targeted like those pockets found in enzymes [26].

When dealing with crystal structures, it is important to keep in mind that they typically represent only one out of many possible protein conformations. With respect to pockets, this means that it is impossible to deduce from a single structure whether and where cavities are available. In other words, a protein may possess pockets that are only accessible in conformations different from the crystal structure and may serve as more favorable binding sites that would be missed when its conformational dynamics is not properly considered. For example, Frembgen-Kesner and Elcock successfully identified in MD simulations an alternative binding site of the p38 MAP kinase which was not accessible in the crystal structure [110]. This observation can be explained with the high

mobility of residues on protein surfaces [111] and led us to the assumption that transient pockets that are large and deep enough to bind small-molecule inhibitors may open from time to time on the protein surface.
In this chapter, we will show that transient pockets may provide a starting point for the *in silico* drug design for cases in which standard screening methods would fail, for example, when no potential binding pocket could be identified. For three model systems (MDM2-p53, Interleukin-2-Interleukin-2 α-receptor, and BCL-X_L-Bak) the PASS program clearly identified the native ligand binding pocket in the inhibitor-bound structures of each system, whereas the binding pockets were not or only partly accessible in the apo structures. Thus, these systems are perfectly suited for validating our pocket detection protocol.

3.2 Model Systems

Although SMPPIIs are known for several protein-protein interactions, we required that for ideal model systems the crystal structures of

- the protein-protein complex (the *complex structure*),
- the apo protein (the *apo structure*),
- and of the protein with a small-molecule inhibitor bound in the interface region (the *holo structure*)

should be available in the Protein Data Bank (PDB) [112]. The three selected model systems fulfilling these criteria will be introduced in the following.

3.2.1 BCL-X_L - Bak

The basal cell lymphoma-extra large (BCL-X_L) protein belongs to the BCL-2 family which mediates apoptosis and functions primarily by forming protein-protein complexes with other members of the BCL-2 family (see Figure 1.1). It is an anti-apoptotic protein that acts by binding the pro-apoptotic Bak protein and so inhibits its function [113]. Like other anti-apoptotic proteins, it is overexpressed in some forms of cancer, resulting in an increased cell proliferation [114]. Hence, the interaction between BCL-X_L and Bak is a promising drug target [115].
The structure of BCL-X_L consists of two central hydrophobic α-helices surrounded by five amphipathic helices [116] as shown in Figure 3.1. Bak is also an α-helical protein that is mainly defined by a single α-helix, the BH3 region (see Fig. 3.2 (a)) [117]. It was shown that the BH3 region of Bak is sufficient to bind to BCL-X_L and, thus, to promote cell death, suggesting that a

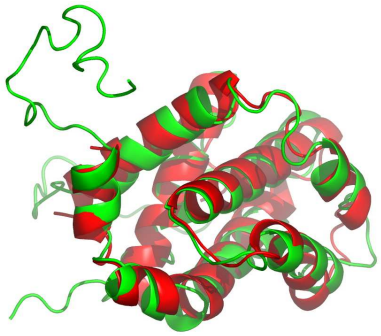
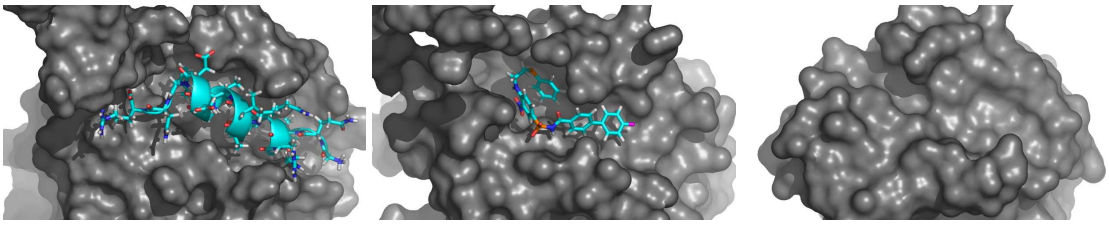


Figure 3.1: Cartoon representation of BCL-X_L in its apo (shown in red) and holo (shown in green) conformations. The backbone RMSD between the two structures is 1.7 Å. Note that the C-terminal region is missing in the apo structure.



(a) (b) (c)

Figure 3.2: Surface representation of the binding interface of BCL-X_L complexed with Bak (a), the SMPPII N3B (b), or in the apo state (c). All complexes are shown in the same orientation. Note that (a) and (b) show extra portions of the protein that could not be resolved in the apo structure shown in (c).

small molecule mimicking the BH3 region of Bak may reestablish pro-apoptotic activity [117]. The following crystal structures were selected:

- apo BCL-X_L (PDB code 1R2D, X-ray structure with a resolution of 1.95 Å) [116]
- complex between a Bak derived peptide and BCL-X_L (PDB code 1BXL, minimized average NMR structure) [118]
- complex between the SMPPII N3B (4-(4-fluorophenyl)-N- [3-nitro-4-(2-phenylsulfanyl-ethylamino) phenyl] sulfonylbenzamide) and BCL-X_L (PDB code 1YSI, NMR structure) [115]

The crystal structures of the complexes reveal that the BH3 region of Bak binds to a hydrophobic cleft formed by the BH1, BH2, and BH3 regions of BCL-X_L as shown in Figure 3.2 (a). The complex between BCL-X_L and Bak is mainly stabilized by intermolecular electrostatic interactions (Arg^{76}, Asp^{83}, Asp^{84} from the Bak peptide and Glu^{129}, Arg^{139}, Arg^{100} from BCL-X_L) and by hydrophobic interactions between the NH_2-terminal residues of the Bak peptide and the BH1 region of BCL-X_L (Val^{126}, Leu^{130}, Phe^{146}). Mutant studies identified Tyr^{101}, Leu^{130}, Gly^{138}, and Arg^{139} as key interacting residues of BCL-X_L [118]. The NMR structure of the inhibitor bound complex revealed that the inhibitor binds into two distinct but proximal subsites in the BH3-binding groove (see Figures 3.2 (b) and A.2 (a), Appendix). One moiety of N3B binds near Arg^{139}, whereas the other one occupies a hydrophobic subpocket formed by Tyr^{101}, Leu^{108}, Val^{126}, and Phe^{146}. Hence, some of the most relevant interactions between BCL-X_L and Bak are mimicked [115]. The binding site is also present in the apo form of BCL-X_L but the groove is more narrow as Figure 3.2 (c) reveals.
BCL-X_L is a widely studied system. For example, Brown and Hajduk used their method for calculating the druggability of a binding site [119] to show that the BH3 binding groove of apo BCL-X_L becomes more druggable during MD simulations [120]. Novak et al. used MD simulations and free energy calculations to study the influence of ligand-induced conformational changes on the activity of known inhibitors [121]. They showed that the improvement of the binding affinity is directly related to a reduced local flexibility of specific regions in the binding groove.

3.2.2 Interleukin-2 - Interleukin-2 α-receptor

Interleukin-2 (IL-2) is an immunoregulatory cytokine and a member of the four-helix bundle cytokine superfamily that is a central part of the immune response [122]. IL-2 binds sequentially to the α- (IL-2Rα), β- (IL-2Rβ), and common γ- chain (γ_c) receptor subunits. This leads to the stimulation of signal transduction pathways resulting in T cell, B cell, and natural killer cell proliferation and clonal expansion. Biochemical studies have shown that the assembly of the IL-2 receptor complex is initiated by the interaction of IL-2 with IL-2Rα [123, 124]. Since IL-2Rα

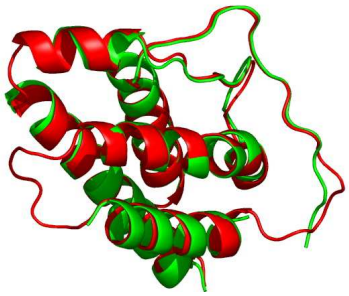


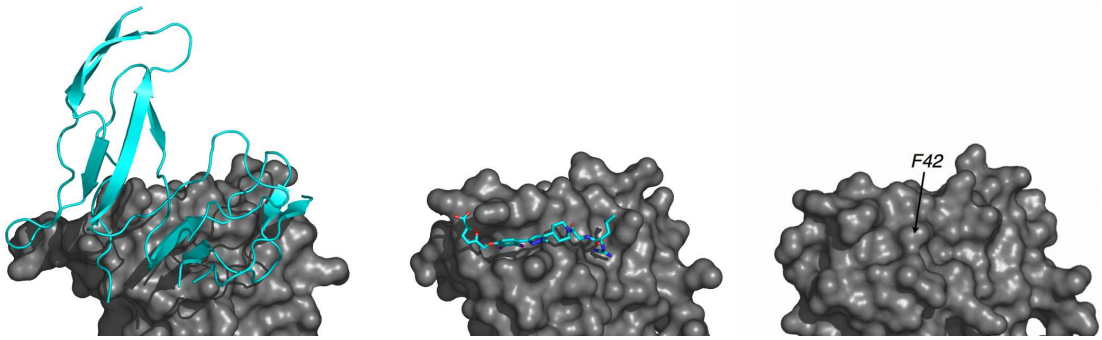
Figure 3.3: Cartoon representation of IL-2 in its apo (shown in red) and holo (shown in green) conformations. The backbone RMSD between the two structures is 0.5 Å.

is not expressed on resting T and B cells but continuously expressed by the abnormal T cells of patients with forms of leukemia, autoimmunity, and organ transplant rejection [125, 126], its interaction with IL-2 is a widely studied drug target.

For this system, the following crystal structures were selected from the PDB:

- apo IL-2 (PDB code 1M47, X-ray structure with a resolution of 1.99 Å) [127]
- complex between IL-2Rα and IL-2 (PDB code 1Z92, X-ray structure with a resolution of 2.8 Å) [123]
- complex between the SMPPII FRH (5-[[2,3-dichloro-4- [5- [1- [2- [[(2R)-2- (diaminomethylideneamino)-4-methylpentanoyl] amino] acetyl] piperidin-4-yl] -1-methylpyrazol-3-yl] phenoxy] methyl] furan-2-carboxylic acid) and IL-2 (PDB code 1PY2, X-ray structure with a resolution of 2.8 Å) [128]

Upon binding, IL-2 undergoes only minor structural adaptions in the backbone (see Fig. 3.3). When binding to IL-2Rα, the interface buries a total surface area of 1,868 Å^2 comprising two prominent hydrophobic patches on IL-2. The first one is composed of Tyr^{45} that packs into a pocket on IL-2Rα formed by Arg^{35} and Arg^{36}. This patch is surrounded by hydrogen bonds between the backbone of Glu^{106} (IL-2) and the side-chain of Arg^{35} (IL-2Rα) and between the side-chains of Glu^{62} (IL-2) and Arg^{36} (IL-2Rα). The second patch is composed of Phe^{42} and Leu^{72} of IL-2 that pack into a recessed pocket on IL-2Rα formed by Leu^{42}, Tyr^{43}, and Met^{25}. Just like the first patch, this patch is surrounded by hydrogen bonds between Lys^{35}, Arg^{38}, and Glu^{68} of IL-2 and Asp^{4}, Asp^{6}, Tyr^{43}, and Asn^{57} of IL-2Rα (Fig. 3.4 (a)). Thermodynamic measurements indicated that the desolvation of the nonpolar surface is the primary energetic driving force of this interaction [123]. Small-molecule inhibitors have been identified that bind to the hot spots on IL-2



(a) (b) (c)

Figure 3.4: Surface representation of the binding interface of IL-2 complexed with its α-receptor (a), the SMPPII FRH (b), or in the apo state (c). All complexes are shown in the same orientation.

[127]. The crystal structures of these inhibitor bound complexes revealed that the inhibitors bind buried into a groove composed of two subsites, a rigid hydrophilic subpocket where hydrogen bonds are possible to Lys^{43} and Glu^{62} and a highly adaptive and hydrophobic narrow channel created by Arg^{38}, Met^{39}, Phe^{42}, Leu^{72}, and Lys^{76}. We focus on the inhibitor FRH whose native binding mode is shown in Figures 3.4 (b) and A.2 (b), Appendix. Phe^{42} acts as a gatekeeper that exposes a hydrophobic channel which is blocked in the apo structure (Fig. 3.4 (c)).
So far, this system has not been extensively studied using *in silico* methods. In their review on targeting protein-protein interactions [26], Gonzalez-Ruiz and Gohlke showed that starting from apo IL-2, conformations can be generated that resemble the holo protein structure suggesting the existence of transient pockets at the binding interface.

3.2.3 MDM2 - p53

The oncoprotein mouse double minute 2 (MDM2) regulates cell growth processes like cell cycling, DNA repair, and apoptosis [129]. It acts by binding to the transcription domain of the tumor suppressor protein p53. This protein is thereby blocked and the transcription of the p53 target genes is prevented. Furthermore, MDM2 serves as specific E3 ligase and promotes the degradation of p53 [130]. p53 is the most frequently inactivated protein in cancer cells because MDM2 is overexpressed in many human tumors. Thus, restoring p53 function by inhibiting its binding to MDM2 is a promising anticancer strategy and MDM2 is, therefore, an important drug target [131, 132].
For our study, we selected the following crystal structures:

- apo MDM2 (PDB code 1Z1M: 24 NMR models) [133]
- complex between a peptide derived from the transactivation domain of p53 and MDM2 (PDB code 1YCR, X-ray structure with a resolution of 2.6 Å) [134]
- complex between the SMPPII DIZ ((2S)-2- (4-chlorophenyl) -2- [(3S)-3- (4-chlorophenyl) -7-iodo-2,5-dioxo-1,3- dihydro-1,4- benzodiazepin-4-yl] acetic acid) and MDM2 (PDB code 1T4E, X-ray structure with a resolution of 2.6 Å) [135]

MDM2 contains two globular repeats that bind to each other via their hydrophobic faces and so form a cleft as shown in Figure 3.5. About one-quarter of this cleft is narrow and shallow, while the remaining portion is wide and deep. The p53-derived peptide adapts an amphipathic α-helical conformation and binds into the deeper and wider portion of this cleft (compare Fig. 3.6 (a)). It buries a total surface area of 1,493 $Å^2$. The peptide side-chains Phe^{19}, Trp^{23}, and Leu^{26} of p53 fit tightly in pockets within this cleft formed by the MDM2 residues Met^{62}, Tyr^{67}, and

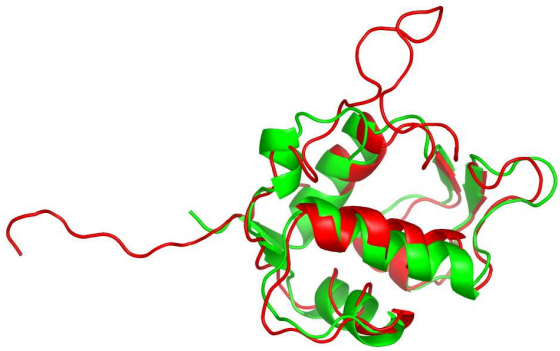
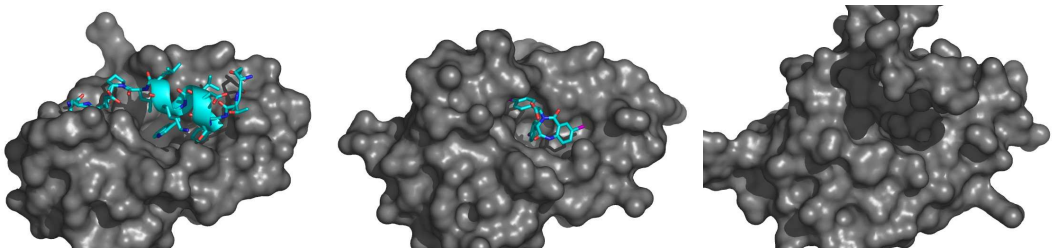


Figure 3.5: Cartoon representation of MDM2 in its apo (shown in red) and holo (shown in green) conformations. The backbone RMSD between the two structures is 1.6 Å. Note that the N- and the C-terminal regions are missing in the holo structure.



(a) (b) (c)

Figure 3.6: Surface representation of the binding interface of MDM2 complexed with p53 (a), the SMPPII DIZ (b), or in the apo state (c). All complexes are shown in the same orientation. Note that (c) shows extra portions of the protein that could not be resolved in the other two structures.

Ile^{61}, Phe^{91}, and Leu^{54}, Ile^{99}, respectively [133]. The crystal structure of the SMPPII DIZ bound to MDM2 demonstrates that the inhibitor binds to the same hydrophobic cleft and mimics the α-helical structure of the p53-derived peptide. It occupies the same pockets as the peptide side-chains Phe^{19}, Trp^{23} and Leu^{26} as shown in Figure 3.6 (b). The interactions are mainly nonspecific van der Waals contacts, just like the interactions with p53 (see also Fig. A.2 (c), Appendix) [135]. The MDM2 structure is quite unstable, but gets more stable (with respect to unfolding) upon ligand binding. The 24 NMR models representing this conformations differ mainly in the terminal loop regions. These models show that large conformational changes accompany ligand binding. In these conformations the shallow end of the binding cleft is partially occupied by an N-terminal segment (residues 19-25) and the cleft is generally less ordered and less wide (see Fig. 3.5 and 3.6 (c)) [133].
Besides the virtual screening study of Bowman et al. [36] already cited in Section 3.1, this system has been investigated by several other *in silico* approaches. For example, Barrett et al. used the peptide-bound MDM2 protein to show that principal component analysis of a conformational ensemble generated using CONCOORD predicted quite similar concerted motions as an MD simulation [136]. More interestingly, they observed that the first eigenvector was coupled to an opening and closing motion of the native binding pocket that was even more pronounced when the peptide was removed. A later study by Espinoza-Fonseca and Trujillo-Ferrara who conducted MD simulations of the same crystal structure of MDM2 with and without the bound peptide underpinned this finding [137]. They found that most motions of this binding site were accounted for by the first eigenvector for the holo protein and by the first two eigenvectors for the apo protein. Furthermore, they observed that the binding cleft was wider and more stable with the peptide bound. Dastidar et al. studied the complex between MDM2 and different p53-derived peptides (wildtype and mutants) by MD simulations as well and reported, for example, that the surface of MDM2 adapted optimally to the various peptides [138]. These findings suggest that the binding pocket is also accessible in apo MDM2, but less stable than with a ligand bound.

3.3 Methods and Materials

As this chapter describes our initial pocket detection protocol, this will now be introduced in detail. The subsequent chapters use the same structures and the docking procedure as presented here.

3.3.1 Preparation of the Experimental Structures

The apo and holo protein structures of BCL-X$_L$, IL-2, and MDM2 mentioned above were taken from the PDB. If multiple chains were available in the PDB file, either the one with the least

number of missing residues or atoms was chosen or chain A. All hetero atoms (including the ligands) were manually removed and the apo structures were superimposed on the holo structures based on the C_α atoms using the VMD program [139]. As residues 28-81 of BCL-X_L are missing in the apo crystal structure, the two parts of the protein were modeled as two distinct chains. In the apo structure of IL-2, the missing residues 75, 76, and 99-102 were modeled as loops of the lowest AMBER/GBSA potential energy generated by the program RAPPER [140]. We note that for both systems the missing residues are far away from the native binding pocket. The apo structure of MDM2 is represented by 24 NMR models that differ mainly in the loop regions. Since no model is defined as most representative, the first model was chosen.

3.3.2 Molecular Dynamics Simulations

For the MD simulations of the proteins, the GROMACS 3.3.1 package [141] was used along with the OPLS-AA force field [46]. The prepared apo structures of the proteins were placed in cubic boxes of 6.2-8.3 nm box dimensions with periodic boundary conditions and explicit TIP4P solvent molecules [142] were added. The system was then pre-equilibrated by 500 steps of steepest-descent energy minimization keeping the heavy proteins atoms harmonically restrained using a force constant of 1,000 kJ mol^{-1} nm^{-2}. Na^+ or Cl^- counter-ions were added to ensure a net neutral charge of the simulation system and the energy minimization was repeated. The system was treated as a NPT ensemble and further equilibrated during a 100 ps MD simulation. The Berendsen method [143] was used to ensure a constant temperature of 300 K and pressure of 10^5 Pa. Protein, solvent, and counter-ions were coupled to separate baths with coupling constants of 0.1 ps for the temperature and 1 ps for the pressure coupling. A cutoff of 0.9 nm was used to compute van der Waals interactions and electrostatic interactions beyond the short-range cutoff of 0.9 nm were treated by the Particle-Mesh-Ewald method [47]. The covalent bonds were constrained by the LINCS procedure [144]. Simulation snapshots were collected every 2.5 ps during the subsequent 10 ns production run, yielding a total of 4,001 MD snapshots. Before they were further processed, they were superimposed on the apo structure based on the C_α atoms as described above. The MD simulations were repeated once for every system to check the results for reproducibility.

3.3.3 Pocket Detection Using the PASS Algorithm

The PASS program was applied to every apo and holo structure and to each MD snapshot after removal of all hetero and hydrogen atoms. As surface pockets tend to be quite flat, the default burial count threshold of 55 protein atoms within 8 Å was too high. Therefore, the "-more" option was used to reduce this threshold to 45 protein atoms to obtain more probes and ASPs. Note that the output of PASS is a file containing the ASPs and a file containing all probes. There is no assignment which probe belongs to which ASP and, thus, no information about the pocket volume. We developed a C++ program called *PocketId* that solves this problem: It associates probes to ASPs to obtain contiguous *patches*. Within this context, a patch is a set of PASS probes used to represent the pocket as a contiguous volume. This volume can be considered as the negative image of the binding pocket as identified by the PASS algorithm. Then the number of patches in a given protein structure is given by the number of ASPs. By assigning each probe to the nearest ASP if it overlaps with any probe already assigned to this patch so far, it is guaranteed that each patch is contiguous. This procedure is listed in Algorithm 1 and has a run time that is quadratic in the number of probes per structure ($O(n^2_{probes} \cdot n_{asps})$). Note that defining each ASP to represent one pocket leads to the subdivision of very large cavities into two or even more pockets, when they consist of more than one ASP.

Algorithm 1 Algorithm for the assignment of PASS probes and ASPs to patches

Input: $asp \leftarrow ASPs$ given in ASPs file from PASS run for this structure
Input: $probes \leftarrow probes$ given in probe file from PASS run for this structure
for $i = 0$ to $|asp|$ **do**
  $patch[i] \leftarrow asp[i]$ {initialize each patch with an ASP}
end for
$last\_probes\_size \leftarrow |probes| + 1$
{repeat until no new probes can be added to patches}
while $|probes| < last\_probes\_size$ **do**
  $last\_probes\_size \leftarrow |probes|$
  {search for an appropriate patch for each not yet assigned probe}
  for each $probe \in probes$ **do**
    $min\_dist \leftarrow \infty$
    $asp\_index \leftarrow -1$
    for $i = 0$ to $|asp|$ **do**
      if $distance(asp[i], probe) < min\_dist\ AND\ connected(patch[i], probe)$ **then**
        $min\_dist \leftarrow distance(asp[i], probe)$
        $asp\_index \leftarrow i$
      end if
    end for
    if $asp\_index \geq 0$ **then**
      {probe can now be assigned to a patch and removed from set of not yet assigned probes}
      $patch[asp\_index] \leftarrow patch[asp\_index] \cup probe$
      $probes \leftarrow probes \setminus probe$
    end if
  end for
end while

return $patch[0]$, ... , $patch[|asp|]$

3.3.4 Calculation of Pocket Properties and Dynamics

For each patch, its *pocket lining atoms* (PLAs) were determined, which are all protein atoms found within a distance of 5 Å. Based on the chemical properties of the PLAs, the polarity of the pocket was approximated by

$$polarity = \frac{|PLA \setminus C_{atoms}|}{|PLA|} \tag{3.1}$$

The probes used by the PASS algorithm have two different radii: the probes in the first layer are of 1.8 Å radius and the ones from the subsequent layer have a radius of 0.7 Å. Hence, the pocket volume was estimated by the following formula:

$$volume = |probes_{layer=1}| \cdot \frac{4\pi}{3} \cdot (1.8)^3 + |probes_{layer>1}| \cdot \frac{4\pi}{3} \cdot (0.7)^3 \tag{3.2}$$

Having detected all pockets occurring in the 4,001 snapshots (about 11,000-20,000), it is of interest to investigate which pockets identified in different conformations correspond to each other. To this end, all pockets were clustered using an agglomerative complete linkage approach. The similarity between two pockets was defined by the similarity of their PLAs:

$$similarity(PLAs_i, PLAs_j) = \frac{|PLAs_i \cap PLAs_j|}{min(|PLAs_i|, |PLAs_j|)} \tag{3.3}$$

During the clustering, it was taken care that the similarity of two pockets was at least 85% and no cluster contained more than one pocket taken from the same MD snapshot. The clustering step is illustrated in Algorithm 2. After clustering, all pockets within the same cluster were labeled by the same unique pocket identifier (PID). We use the term *PID* to refer to a transient pocket. Thus, the dynamics of a transient pocket can be observed via the pockets belonging to this cluster that represent the different states taken from subsequent MD snapshots. Furthermore, for comparing two PIDs, a *subpocket* was determined for each transient pocket. These subpockets are characterized by those PLAs that line the pocket in at least 33% of all its occurrences. Moreover, two

Algorithm 2 Algorithm for the identification of analogous patches within different conformations

$clusterSet \leftarrow \emptyset$
while $(similarities \neq \emptyset)$ AND $(max(similarities) \geq 85\%)$ **do**
  $entry \leftarrow max(similarities)$
  if $isNew[entry.patch_1]$ AND $isNew[entry.patch_2]$ **then**
    {both patches are new}
    $clusterSet[entry.patch_1] \leftarrow entry.patch_1, entry.patch_2$
    $cluster[entry.patch_2] \leftarrow entry.patch_1$
    $isNew[entry.patch_1] \leftarrow false$
    $isNew[entry.patch_2] \leftarrow false$
  else if $isNew[entry.patch_1]$ **then**
    {only $entry.patch_1$ is new, if there is no patch within $clusterSet[cluster[entry.patch_2]]$ with the same structure ID, add $entry.patch_1$ to this cluster}
    if $NOThasCommonStrID(clusterSet[cluster[entry.patch_2]], entry.patch_1)$ **then**
      $clusterSet[cluster[entry.patch_2]] \leftarrow clusterSet[cluster[entry.patch_2]] \cup entry.patch_1$
      $cluster[entry.patch_1] \leftarrow cluster[entry.patch_2]$
      $isNew[entry.patch_1] \leftarrow false$
    end if
  else if $isNew[entry.patch_2]$ **then**
    {only $entry.patch_2$ is new, if there is no patch within $clusterSet[cluster[entry.patch_1]]$ with the same structure ID, add $entry.patch_2$ to this cluster}
    if NOT $hasCommonStrID(clusterSet[cluster[entry.patch_1]], entry.patch_2)$ **then**
      $clusterSet[cluster[entry.patch_1]] \leftarrow clusterSet[cluster[entry.patch_1]] \cup entry.patch_2$
      $cluster[entry.patch_2] \leftarrow cluster[entry.patch_1]$
      $isNew[entry.patch_2] \leftarrow false$
    end if
  else
    {both have already assigned to a cluster. If there is no patch within $clusterSet[cluster[entry.patch_1]]$ with the same structure ID as $entry.patch_2$ or the other way around, merge the clusters}
    if NOT $hasCommonStrID(clusterSet[cluster[entry.patch_1]], clusterSet[cluster[entry.patch_2]])$ **then**
      $clusterSet[cluster[entry.patch_1]] \leftarrow clusterSet[cluster[entry.patch_1]] \cup clusterSet[cluster[entry.patch_2]]$
      $clusterSet[cluster[entry.patch_2]] \leftarrow \emptyset$
      $cluster[entry.patch_2] \leftarrow cluster[entry.patch_1]$
    end if
  end if
  $similarities \leftarrow similarities \setminus entry$
end while
return $clusterSet$

sets of transient pockets (e.g. resulting from two different MD runs) can then be compared to each other by determining the fraction of PIDs of one set that have a similarity of at least 50% to any PID from the other set. The complete analysis of the transient pockets including application of the PASS algorithm, clustering, and calculation of the properties took 16-20 h for each set of MD snapshots on one 2.8 GHz Xeon CPU.

3.3.5 Docking Setup

All docking experiments were performed with AutoDock 3.0.5. The inhibitors were taken from the holo structures. The AutoDockTools (version 1.4.3) modules of the Python Molecular Viewer software [145] was used to add hydrogens and to compute the Gasteiger atomic charges [146]. The rotatable bonds (10 for N3B, 17 for FRH, and 5 for DIZ) were assigned with AutoTors.
Four different docking experiments were performed: (1) re-docking into the holo structure, (2) docking into the apo structure, (3) docking into all MD snapshots, and (4) docking into all transient pockets located at the interface. As polar hydrogen atoms are needed for a successful docking, they were added to the crystal structures and the nonpolar hydrogens were removed in the MD snapshots. Kollman united-atom partial charges and solvation parameter were assigned by the AutoDockTools utility. All grid maps were calculated with AutoGrid3 using the default spacing of 0.375 Å between the grid points. In the docking experiments (1) - (3), the grid center was chosen to coincide with the center of mass of the ligand in its bound conformation and the default

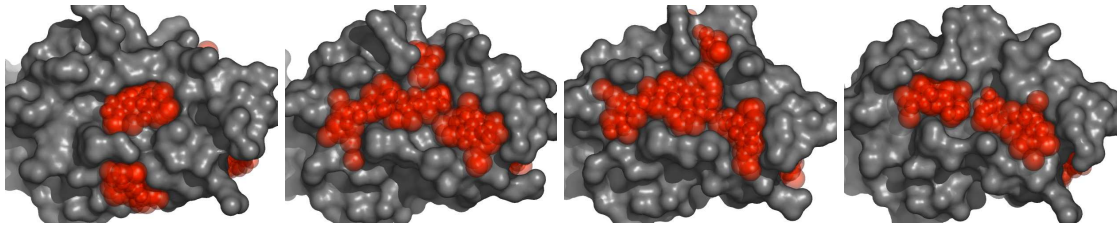
grid dimensions of 21 Å x 21 Å x 21 Å were used. In the docking experiment (4), no prior information about the bound ligand conformation was used and, thus, the grid center was set to the center of mass of the transient pocket. The holo structures of IL-2 and BCL-X_L reveal that only a terminal moiety of the ligands may be placed into a pocket. Hence, the grid dimensions were expanded to 30 Å x 30 Å x 30 Å. Whereas for MDM2, the grid dimensions were reduced to 16.125 Å x 16.125 Å x 16.125 Å to confine the position of the smaller ligand to the transient pocket.
Docking was performed using the standard LGA protocol with default parameters, i.e. an initial population of 50 randomly placed individuals, a maximum number of 2,500,000 energy evaluations, a mutation rate of 0.02, a crossover rate of 0.80, and an elitism value of 1. The probability of performing a local search using the Solis and Wets algorithm with a maximum of 300 iterations was set to 0.06, and the maximum number of consecutive successes or failures before doubling or halving the local search step size was 4. The complete docking step consisting of calculating the grid maps and 10 independent docking runs took 1-3 min per protein conformation on one 2.8 GHz Xeon CPU depending on the flexibility of the ligand and the size of the grid box.

3.4 Results

Before running the pocket detection protocol, the PASS algorithm was applied to the selected holo structures in order to validate its ability of detecting the native binding pockets. In addition, a reference value for the volume of these pockets in the inhibitor-bound state was obtained. For all tested structures, the native binding pocket could be identified. But for the holo structure of IL-2, only one of the two subsites of the binding pocket was detected. Applying the PASS algorithm to the apo structures revealed that in the absence of a ligand the native binding pocket is partly open in BCL-X_L (36% of the calculated pocket volume of the holo structure) as well as in MDM2 (42%). In the structure of apo IL-2 the binding pocket could not be detected at all.

3.4.1 Transient Pockets Detected in the MD Snapshots

The proteins were stable over the simulation time. As an example, the stability of the secondary structure of the first MD simulation runs is discussed in Section B.1 - B.3 (Appendix). The detection of pockets in the MD snapshots and the subsequent clustering revealed surprising results. For BCL-X_L, 23 distinct transient pockets were detected in the first run and 20 in the second run. For IL-2, 23 (respectively, 31), and for MDM2, 33 (respectively, 36) transient pockets were detected. In comparison, the total numbers of pockets detected for the apo structures were four for BCL-X_L, two for IL-2, and five for MDM2.



(a) 100 ps (b) 200 ps (c) 300 ps (d) 400 ps

Figure 3.7: Protein surfaces are fluid-like as illustrated at the example of MDM2. The molecular surface of the protein after (a) 100 ps, (b) 200 ps, (c) 300 ps, and (d) 400 ps of simulation time is shown. The PASS probes used to detect pockets in these MD snapshots are represented by red spheres. Note that due to the mobile surface the number of probes and the shape of the pocket changes significantly from snapshot to snapshot.

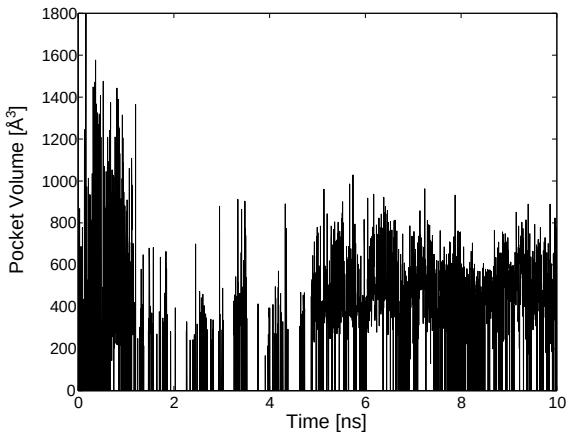


Figure 3.8: The fast opening and closing behavior of a transient pocket shown at the example of PID 5 of MDM2 (run 1).

Properties of Transient Pockets Analyzing the frequency of occurrences and the average pocket volumes gave similar results for all three systems (see Tables 3.1 and 3.2). The largest fraction (35.0-52.2%) of the transient pockets were rare events that were only present in less than 1% of all MD snapshots with mean volumes between 335.5 and 365.3 Å^3. Thus, in general, they represent the smallest cavities for each system, whereas the highest populated pockets (detectable in more than 50% of all MD snapshots) tended to belong to the largest ones of the respective system. However, there are just a small number of such favorable pockets (3.2-13.0% of all transient pockets) in each system. Especially the dynamics of the transient pockets was surprising. Instead of opening slowly, the pockets suddenly opened to volumes up to 500 Å^3 within 2.5 ps, stayed open for some time, vanished, and reappeared again several times. The mobility of molecular surfaces is due to its fluid-like properties that can be observed during MD simulations. An example of how the flexible surface effects the formation of pockets is shown in Figure 3.7 and an example illustrating the fast opening and closing behavior characteristic for the transient pockets is shown in Figure 3.8.

	mean volume [Å^3]							
system	**freq.: <1%**		**freq.: 1-10%**		**freq.: 10-50%**		**freq.: >50%**	
	run 1	run 2	run 1	run 2	run 1	run 2	run 1	run 2
BCL-X$_L$	361.4	340.2	405.1	384.4	451.5	469.9	527.7	423.8
IL-2	346.2	365.3	338.2	399.7	355.1	401.0	452.7	398.9
MDM2	335.5	354.3	400.7	365.3	422.3	405.9	468.7	639.1

Table 3.1: Mean volumes of the transient pockets according to their frequency per system for the two independent MD runs.

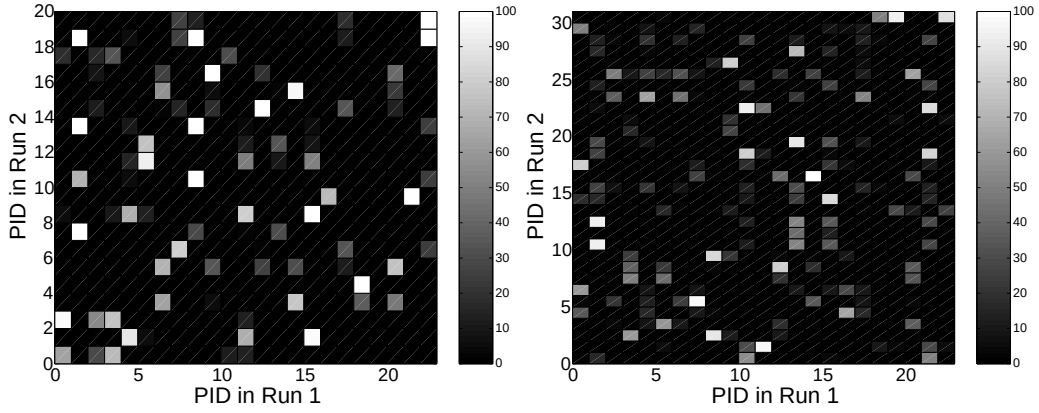
	relative number [%]							
system	**freq.: <1%**		**freq.: 1-10%**		**freq.: 10-50%**		**freq.: >50%**	
	run 1	run 2	run 1	run 2	run 1	run 2	run 1	run 2
BCL-X$_L$	52.2	35.0	13.0	25.0	21.8	30.0	13.0	10.0
IL-2	47.8	51.6	26.1	19.4	17.4	25.8	8.7	3.2
MDM2	45.5	47.2	24.2	19.4	21.2	27.8	9.1	5.6

Table 3.2: Relative number of transient pockets with different frequencies per system for the two independent MD runs.

system	reproducibility [%]									
	freq.: <1%		freq.: 1-10%		freq.: 10-50%		freq.: >50%		total	
	run 1	run 2	run 1	run 2	run 1	run 2	run 1	run 2	run 1	run 2
BCL-X_L	66.7	71.4	100.0	100.0	100.0	100.0	100.0	100.0	82.6	90.0
IL-2	81.8	62.5	100.0	83.3	100.0	100.0	100.0	100.0	91.3	77.4
MDM2	80.0	100.0	100.0	71.4	85.7	90.0	100.0	100.0	87.9	91.7

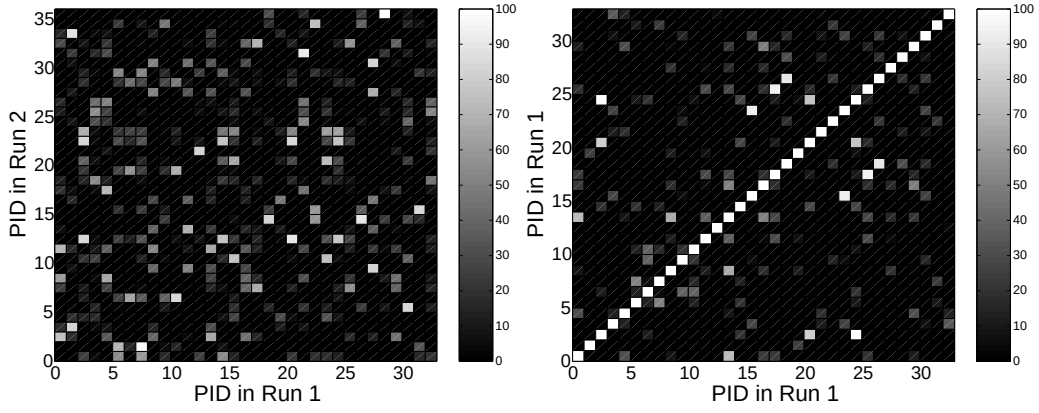
Table 3.3: Reproducibility of the PIDs according to their frequency for the two independent MD runs.

Reproducibility of Transient Pockets Figure 3.9 (a)-(c) shows that most transient pockets detected in a MD run are also found in an other MD run. This indicated that most transient pockets are reproducible. Note that a PID may correspond to more than one PID and that many PIDs found in the same run overlap as well as depicted in Figure 3.9 (d). Table 3.3 lists the reproducibility of the transient pockets by their frequency. In general, the more frequent a pocket occurs in one MD simulation, the more probable is its appearance in another MD simulations. Consequently, rare event pockets are often non-reproducible.



(a) BCL-X_L (run 1 vs. run 2)

(b) IL-2 (run 1 vs. run 2)



(c) MDM2 (run 1 vs. run 2)

(d) MDM2 (run 1 vs. run 1)

Figure 3.9: Pairwise similarities of the transient pockets. (a) - (c) show that most PIDs are reproducible by another MD run, (d) shows the similarity of PIDs obtained within the same run (run 1) for MDM2. The shading scheme indicates the level of similarity.

system	apo overlap vol. [%]	MD mean overlap vol. [%]		MD max overlap vol. [%]		no. MD snapshots with overlap		overlapping PIDs[a]	
		run 1	run 2	run 1	run 2	run 1	run 2	run 1	run 2
BCL-X_L	35.6	33.2	22.7	84.2	73.6	2,716	1,924	15	8, 11
IL-2	0	45.3	31.5	115.1[b]	130.7[b]	1,440	1,992	8	2, 4
MDM2	2.2	53.7	39.4	136.2[b]	99.2	2,716	3,883	2, 5	0, 22

[a] Shown are only the PIDs involved in the maximum overlap

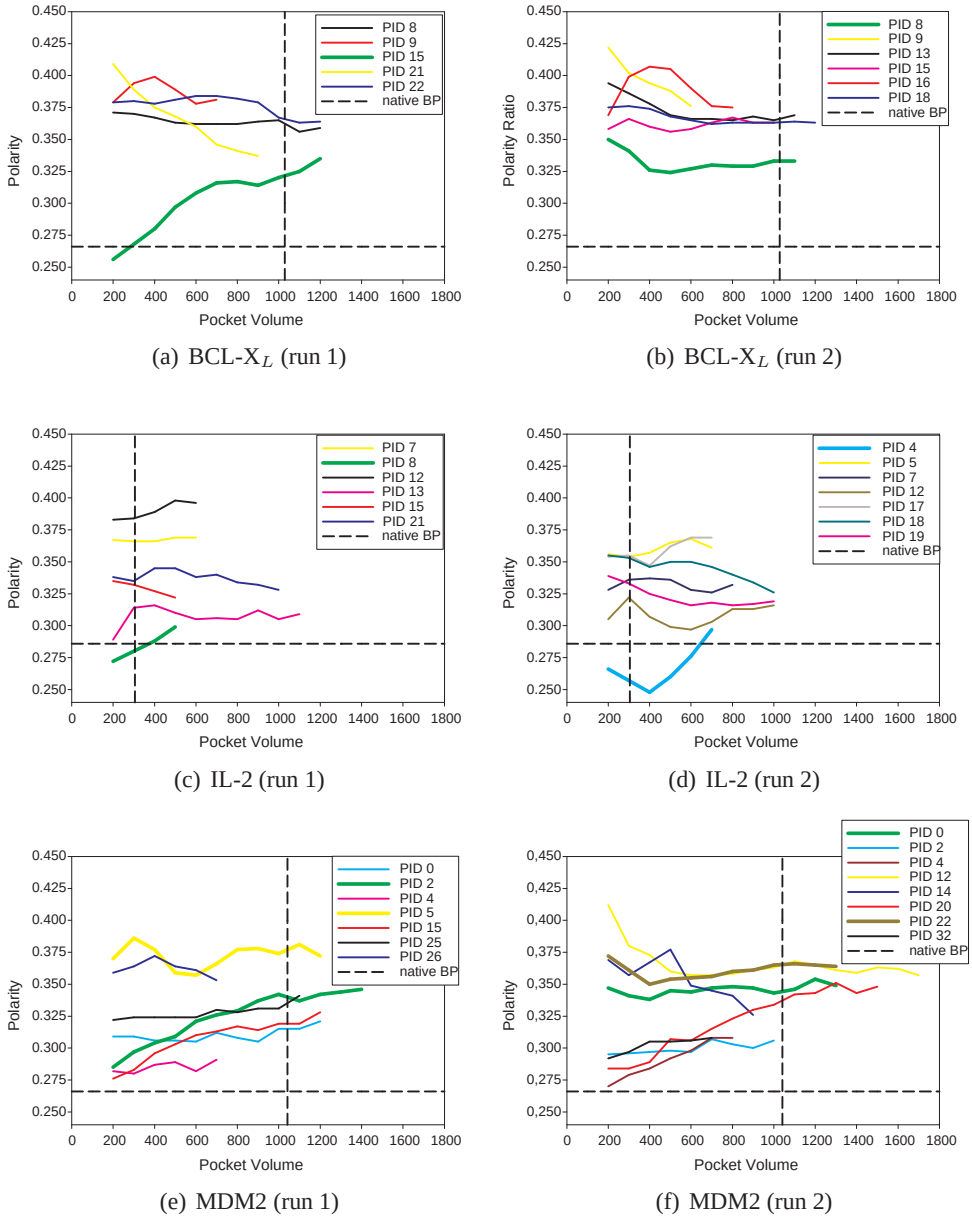
[b] Overlap volume is larger than in the holo structure

Table 3.4: Volume of the PASS probes overlapping with the atoms of the superimposed ligand relative to the overlap volume for the ligand bound structure per system for the two independent MD runs.

Did the native binding pocket open during the MD simulations? To test whether the native binding pocket is among these transient pockets, we superimposed the holo structures onto the MD snapshots and onto the apo structures and then determined for each conformation the PASS probes overlapping the inhibitor atoms. New pocket volumes were calculated by considering only the overlapping PASS probes and comparing their volumes to those obtained for the inhibitor bound structures. Furthermore, the PIDs possessing the largest overlap with the native inhibitor were identified and defined to correspond to (subpocket of) the native binding pocket. (As the PIDs within the same run may overlap as well resulting from the definition of the ASPs by the PASS algorithm, more than one PID may correspond to the native binding pocket.) The results shown in Table 3.4 indicate that for all three systems the native binding pocket opened up during MD simulations. For BCL-X_L and MDM2, where the native binding pocket was already detectable in the apo structure, the mean overlap volume determined for the MD snapshots was more or less comparable to the overlap volume of the apo form of the protein (35.6% for BCL-X_L and 42.2% for MDM2). However, in some MD snapshots of BCL-X_L the native binding pocket was more than twice as large as in the apo form, although not quite as large as in the holo structure. In some MD snapshots of MDM2, the PASS volumes overlapping with the superimposed inhibitor were even of equal or larger size than in the holo structure, indicating that the native binding pocket was sometimes large enough to fully accommodate the native inhibitor. Notably, for IL-2, the native binding pocket, which was not detectable in the apo form, was also found to fully open during the 10 ns MD simulations. Similar to MDM2, the overlapping volume was larger than in the inhibitor bound complex. Considering that the native binding pocket consists of two subsites, this result shows that the other subsite which was not detected in the inhibitor bound complex was detectable in some MD snapshots. This means that both subsites of the native binding pocket opened and the binding pocket may be fully accessible in some of the MD snapshots.

Polarity of Transient Pockets In addition to the volume, the polarities of transient pockets were studied. While the polarity of the entire protein surface of the apo structure is 0.37 for BCL-X_L and MDM2 and 0.38 for IL-2, the polarity of the transient pockets ranged from 0.25 to 0.45. The pocket volumes and the corresponding polarity for all PIDs with frequency greater than 20% are plotted in Figure 3.10. This analysis reveals that, in general, the largest pockets (volumes of $\geq$ 800 Å^3) have a smaller polarity than the overall protein surface, suggesting that the protein interior partly opens up and, thus, these pockets may be "sticky" enough to bind ligands. But except for IL-2, all detected transient pockets are more polar than the native binding pocket in the holo structure. This result suggests that only quite polar pockets open up during MD simulations in a polar solvent. Less polar pockets may require a less polar environment or the presence of a ligand to open, suggesting an induced-fit mechanism. A comparison of the polarity ratios of those

PIDs corresponding to the native binding pockets to all others indicates that for BCL-X_L and IL-2, these PIDs represent the most nonpolar pockets. Note that in the polarity plots for IL-2, the PIDs representing the native binding pocket (PID 8 in run 1 and PID 4 in run 2) correspond to different subpockets. In the first run of IL-2, PID 8 corresponds to the subpocket identified in the holo structure. Hence, it almost possesses the same polarity ratio as the native binding pocket at the reference volume. In the second run, PID 4 corresponds to the less developed subpocket missed by the PASS algorithm in the holo structure.



(a) BCL-X_L (run 1)

(b) BCL-X_L (run 2)

(c) IL-2 (run 1)

(d) IL-2 (run 2)

(e) MDM2 (run 1)

(f) MDM2 (run 2)

Figure 3.10: Changes in the mean pocket polarity depending on the pocket volume. For each PID, the polarity of pocket states having the same volume were averaged to smooth the curve. In order to obtain reliable values, only PIDs with a frequency greater than 20% were used, resulting in a different number of PIDs for the different runs of the same system. The dashed lines indicate the polarity and volume of the native binding pocket (BP). PIDs from different runs corresponding to each other are shown in the same color. The PIDs representing the native binding pocket are shown as thick lines.

system	re-docking RMSD [Å]	re-docking score [kcal/mol]	re-docking rank[a]	apo-docking RMSD [Å]	apo-docking score [kcal/mol]	apo-docking rank[a]
BCL-X_L - N3B	0.9	-10.5	2	3.3	-6.2	5
IL-2 - FRH	1.1	-10.8	1	2.9	-6.2	1
MDM2 - DIZ	1.1	-13.1	2	3.4	-6.7	5

[a]Rank of docking solution among 10 docking runs.

Table 3.5: Best docking results for (re-) docking into the holo and into the apo structures.

3.4.2 Docking into MD Snapshots

Extensive docking studies were performed to validate whether the transient pockets detected by this method are suitable to bind the known inhibitors and, hence, may be used for structure-based drug design. To validate whether AutoDock3 is capable of handling these kinds of ligands that are very flexible and do not bind into deep pockets, we re-docked the known inhibitors into the holo structures of the proteins. Furthermore, the apo structures were used for docking to estimate the extent of conformational changes necessary to accommodate the inhibitors. The results of these two docking experiments shown in Table 3.5 emphasize that the apo structure would not be suitable at all for an *in silico* drug design project. However, when using the MD snapshots much better results are obtained as listed in Table 3.6. We are aware that these results are somehow biased toward the native bound ligand conformation because these ligand conformations were used to define the center of the search grid. Without prior knowledge, it would not be possible to identify the correct docking solutions as the ranks of these results may be quite high. Let us, for example, consider the case of N3B binding to BCL-X_L. When using the known center of mass of the ligand in the docked complex as the grid center of the docking run (see column termed "snapshot docking"), 4.7% of all docking poses have a better score than the docking pose with the smallest RMSD of 1.4 Å. In an *in silico* drug design project, this center of mass would of course not be known and the center of mass of the transient pocket would be used to define the binding site (termed "PID-docking"). In this case, the best solution would only belong to the upper half of all docking poses. However, one should not exclusively focus on the docking pose with the smallest RMSD. The highest ranked docking solutions that can be classified as "correct" (RMSD ≤ 2.0) are listed in Table 3.7 and the corresponding docking poses are shown in Figure 3.11. Here, at least one correct docking solution is always ranked among the best 5% of all docking results. When taking the fraction of buried nonpolar ligand atoms into account, the ranks can be reduced to less than 1% for IL-2 and MDM2 and to less than 3% for BCL-X_L. Thus, even without prior knowledge this docking result would be selected for further investigation.

system	snapshot-docking[a] RMSD [Å]	snapshot-docking[a] score [kcal/mol]	snapshot-docking[a] rank[c] [%]	PID-docking[b] RMSD [Å]	PID-docking[b] score [kcal/mol]	PID-docking[b] rank[c] [%]
BCL-X_L - N3B	1.4	-8.7	4.7	1.5	-7.3	48.3
IL-2 - FRH	1.5	-6.6	20.6	1.9	-6.5	14.1
MDM2 - DIZ	1.9	-11.5	1.1	1.9	-11.5	0.7

[a]Docking into all MD snapshots (grid center coincident with center of mass of superimposed ligand).

[b]Docking into transient pockets (grid center coincident with center of mass of transient pocket).

[c]Number of docking results with better docking score in relation to the total number of docking results.

Table 3.6: Docking results with lowest RMSD for docking into the MD snapshots and transient pockets.

	PID-docking[a]			
system	**RMSD [Å]**	**score [kcal/mol]**	**score rank[b] [%]**	**final rank[c] [%]**
BCL-X_L - N3B	1.8	-9.2	4.8	2.7
IL-2 - FRH	2.0	-7.6	2.8	0.9
MDM2 - DIZ	1.9	-11.5	0.7	0.7

[a] Docking into transient pockets (grid center coincident with center of mass of transient pocket).

[b] Number of docking results with better docking score in relation to the total number of docking results.

[c] Number of docking results with better docking score and higher fraction of buried nonpolar ligand atoms (i.e. relative number of nonpolar ligand atoms overlapping with PASS probes) in relation to the total number of docking results.

Table 3.7: Highest ranked correct (RMSD ≤ 2.0) docking results for docking into transient pockets.

3.5 Discussion

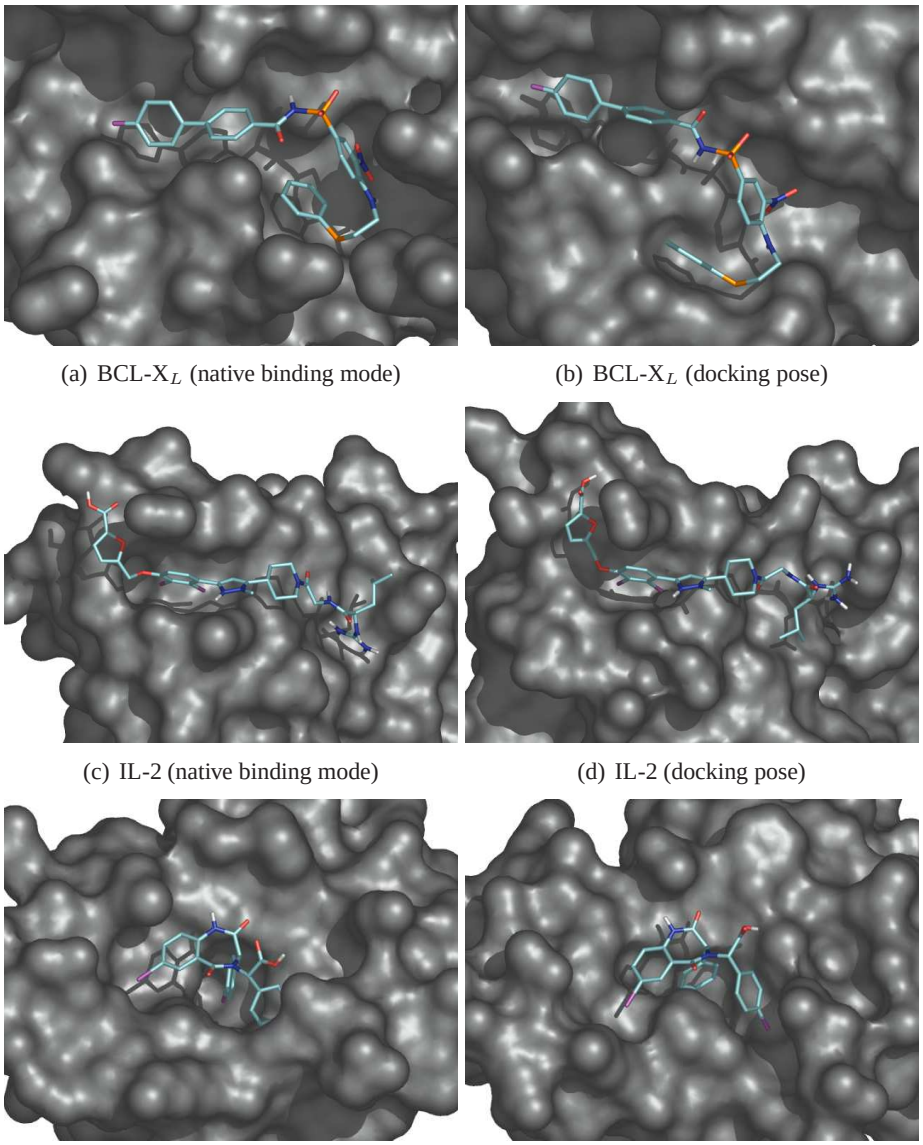
Docking into the apo structures of the proteins revealed that these conformations cannot accommodate the known inhibitors. So far, it is not known whether the native binding pockets only open in the presence of a nearby ligand or whether they also exist in conformations of the apo form of the protein. Even in the latter case, these openings could be rare events that do not occur on the typical nanosecond time scales of molecular dynamics simulations performed at room temperature. But to our surprise, even at a temperature of 300 K, large pockets opened frequently on the protein surface and when docking into these transient pockets, we obtained conformations that were quite close to the native binding modes. However, the shapes of the protein surfaces are somehow different such that some deviations are to be expected. The docking scores in Table 3.6 and 3.7 indicate that pockets of appropriate shapes form spontaneously during MD simulations of the apo BCL-X_L and MDM2 proteins suggesting that for these systems the inhibitor selects an appropriate protein conformation (conformational selection). Whereas for IL-2, the docking score decreased significantly compared to the re-docking score. This may be a sign that the formation of the native binding pocket requires the presence of the ligand with subsequent induced-fit effects.

3.5.1 Comparison to the "(Improved) Relaxed Complex Scheme"

Our approach is very similar to the (improved) RCS approach of Amaro and co-workers [92]. We also generate a conformational ensemble of the protein by MD simulations in explicit water, select eligible snapshots, and use them to dock the inhibitors. In fact, it can be considered as an application of this method to study the problem of detecting transient pockets opening at protein-protein interaction interfaces. Our approach has the advantage that no *a priori* knowledge is required about the location of the binding site. Instead of clustering the MD snapshots by RMSD of the binding site and selecting non-redundant receptor conformations, we cluster the detected pockets and select those located at promising positions (i.e. at the interaction interface). In contrast, the (improved) RCS protocol does not consider pockets explicitly. Although the authors mention that their technique successfully produced true positive docking poses when the binding site was undefined and the docking grid encompassed the entire protein [92], they tested only one enzyme that was in the holo state (and thus already contained a deep cavity) and gave no details about the docking results. Therefore, it is not sure whether the (improved) RCS method would be capable of identifying the native binding site when it corresponds to a transient pocket.

3.5.2 Critical Assessment of the Approach

Although this protocol is quite time-consuming due to the underlying MD simulations, the results are very promising. As an initial criterion for the suitability of this approach for detecting tran-



(a) BCL-X_L (native binding mode) (b) BCL-X_L (docking pose)

(c) IL-2 (native binding mode) (d) IL-2 (docking pose)

(e) MDM2 (native binding mode) (f) MDM2 (docking pose)

Figure 3.11: The best scored near-native docking poses (listed in Table 3.7, PID docking) when docking into the transient pockets (PID-docking) along with the native complexes. In (f), residues 1 to 16 were removed for better visibility.

sient binding pockets, we suggest considering the lowering of docking scores for MD snapshots versus apo structures. But this needs to be tested of course for a larger number of model systems. When considering the rank of the docking scores, one has to keep in mind that we only docked into pockets opening at the protein-protein interaction interface. In a real application scenario, the interface may be unknown and the docking score of a native binding pose may be ranked worse. Nevertheless, we recommend this pocket detection protocol as a starting point for structure-based drug design especially in cases when no appropriate binding pocket can be identified on the surface of the target protein. Then the regions in which transient pockets open may be used as potential binding sites for virtual screening with flexible docking methods. Besides sampling accessible pockets opening at the targeted interface, this approach also has the advantage that it detects pockets anywhere on the protein surface, e.g. one may identify new allosteric pockets. Moreover, these transient pockets and their properties may serve as a prefiltering tool to reduce the number of lig-

ands to be docked in virtual screening. But note that not all detected pockets are druggable. In order to study the properties that characterize a druggable transient pocket or to select MD snapshots that represent promising docking receptors, this protocol has to be tested on a larger number of model systems. Although quite similar results were obtained for all studied systems, one has to be cautious in generalizing these findings. Some ligand binding pockets may require the presence of the ligands, whereas our protocol is only capable of detecting cavities that open spontaneously. A transient pocket with a high frequency may indicate that the opening is energetically favorable and, thus, protein conformations containing such a pocket are sufficiently high populated *in vivo* to be recognized by a ligand. But on the other hand, a transient pocket with a low frequency is not necessarily "not-ligand binding" because the opening of a binding pocket may only be energetically unfavorable in the absence of the stabilizing ligand. Note that the calculated frequency has to be handled with care as the definition of a pocket is vague and crucially depends on the used pocket detection method. Especially defining pocket boundaries and the subdivision in distinct subpockets is subjective and, thus, the clustering is subjective, too.

3.6 Summary and Conclusion

We applied standard MD simulations to three protein systems to show that a surprisingly large number of transient pockets open up on protein surfaces on a 10 ns time scale. These pockets open and close very quickly due to the fluid-like properties of protein surfaces. For each system many transient pockets were detected that differed in volume, frequency, and polarity. On average, more frequently open pockets tend to have larger volumes. Furthermore, we observed that large pockets have a reduced polarity compared to the overall protein surface, suggesting that they may be suitable for ligand binding. Yet, these pockets are usually still more hydrophilic than the native binding pocket indicating that induced-fit effects are important during ligand binding, even if the receptor conformation contains a preexisting pocket. The opening of most transient pockets seems to be energetically favorable and, thus, the pockets are reproducible as evidenced by a second set of control simulations.

The identified transient pockets represent potential binding sites of new inhibitors. When focussing attention on the location of the native binding pocket, pockets of similar size compared with the known inhibitor bound could be observed in all three test systems. Flexible ligand docking into these pockets resulted in binding modes that differed only by 2 Å RMSD from the crystal structure conformation.

To our knowledge, this is the first protocol that accounts for conformational changes occurring on protein-protein interaction interfaces upon ligand binding. It clearly underlines the importance of incorporating protein flexibility in ligand design studies, particularly on the protein surface. This pocket detection protocol may therefore be an interesting starting point for structure-based drug design, especially on protein-protein interaction interfaces, when the crystal structure of the target protein lacks appropriate binding pockets.

Chapter 4

What Induces the Pocket Openings on Protein Surface Patches Involved in Protein-Protein Interactions?

With the pocket detection protocol introduced in Chapter 3, we were able to show that a surprisingly large number of pockets open during MD simulations of the apo structures. In this chapter, the underlying mechanisms of these pocket openings will be studied using modified versions of the initial protocol. In addition, MD simulations are replaced by three more efficient methods for generating conformational ensembles and their appropriateness for the sampling of transient pockets will be discussed. The findings were published in the *Journal of Computer-Aided Molecular Design* in 2009 [147].

4.1 Introduction

In Chapter 3, we have presented a pocket detection protocol that provides a starting point for *in silico* drug design in cases when no potential binding pocket could be identified so that standard screening methods would fail. This protocol is based on the finding that large pockets not detectable in the apo crystal structures of BCL-X_L, IL-2, and MDM2 opened frequently on the protein surfaces during standard MD simulations of 10 nanoseconds length at room temperature. These identified transient pockets represent potential binding sites of new inhibitors. Most of these pockets, especially the most frequent ones, were reproducible as evidenced by a second MD simulation run for each system. Furthermore, for all three systems, we observed that pockets of similar size as with a known inhibitor bound opened at the native binding site. When docking the inhibitors into these transient pockets with AutoDock3, docking poses with less than 2 Å RMSD from the native binding mode were predicted. However, the differences in these docking scores to the re-docking scores suggested that the physicochemical properties of the transient pockets were not as suitable for inhibitor binding as those of the native binding pocket. For example, most transient pockets were less polar than the overall protein surface but not as hydrophobic as the native binding pocket. Thus, we assumed that hydrophobic pockets are more appropriate as putative ligand binding sites. However, it is currently unclear whether the opening of such nonpolar pockets is energetically "forbidden" in water and requires the presence of a ligand. We speculated that this could be circumvented by simulating the protein in a nonpolar solvent that may allow for the opening of more and larger hydrophobic pockets, even in the absence of a ligand. Methanol appeared as a good candidate solvent as it may act as a hydrogen bond donor and acceptor and is less polar than water (its relative dielectric constant is 33 [148]). It has been used before as solvent for MD simulations of peptides and proteins. For example, Alonso and Daggett studied the unfolding and folding of ubiquitin by MD simulations in a mixture of methanol and water to mimic the cytosolic environment of biological cells [149]. Interestingly, they observed that partially unfolded

conformations with increased exposure of hydrophobic residues were only stable in the presence of methanol whereas the protein collapsed in water. So far, MD simulations in pure methanol have mostly been applied to membrane-bound peptides. Kovacs et al. compared simulations of an integral membrane helix of the surfactant protein C in chloroform, water, and methanol [150]. They observed a burial of aliphatic side-chains in water resulting in a decreased total accessible surface area of the peptide, whereas in chloroform more nonpolar side-chains became exposed and the total accessible surface area increased. In methanol, the total accessible surface area also increased because polar as well as nonpolar side-chains became exposed. In addition, they observed that the helical conformation was more stable in water and methanol than in chloroform.
In this chapter, we will try to understand the underlying dynamics of the opening of transient pockets on protein surfaces. To this end, we investigated the following points:

- How stable is the native binding pocket without bound ligand?
- Can pockets on the protein surface fully open in water and what is the additional benefit of simulating the proteins in a less polar solvent?
- Are backbone movements necessary for the opening of pockets or are side-chain rotations sufficient?

To answer these questions, we generated different conformational ensembles, applied the pocket detection protocol, and compared the properties of all detected transient pockets. As model systems, we used the three proteins (MDM2, BCL-X_L, and IL-2) introduced in Chapter 3. The MD simulations of the apo proteins discussed in Chapter 3 serve as a reference point. To answer the first question, MD simulations of the holo structures after removal of the inhibitor were conducted. For the second question, additional MD simulations in methanol were performed and the pockets found in these snapshots were compared to those detected from the simulations in water. The third question was addressed by comparing MD simulations with harmonic restraints on all heavy backbone atoms to the unrestrained simulation at the example of MDM2.
The main drawback of our previously introduced pocket detection protocol is the high computational demand of the underlying MD simulations. Thus, it would be desirable to replace them by a more efficient protocol. For this purpose we have also tested three established methods that generate conformational ensembles in a more efficient way: normal mode analysis, CONCOORD, and tCONCOORD.
In addition to investigating which aspects of the natural conformational dynamics of proteins (e.g. backbone movements, side-chain movements, or deformations along low-frequency normal modes) induce the formation of surface pockets, we tried to characterize for each method its appropriateness for detecting potential binding pockets. As before, this was realized by focusing on the binding pockets of known SMPPIIs because these are the only cavities with experimentally validated small-molecule ligand binding capabilities. By docking the known inhibitors with AutoDock3 into transient pockets that opened at the binding interface and by comparing the docking pose to the native binding mode, we could identify which methods are best suited for sampling putative ligand binding pockets.

4.2 Methods and Materials

The structure selection, preparation, equilibration, MD simulations, superposition of all conformations, and docking runs (using the “PID-docking” setup) were done as described in Chapter 3. All energy minimizations, MD simulations, and normal mode analysis were performed with the GROMACS 3.3.1 package using the OPLS-AA force field.

4.2.1 Molecular Dynamics Simulations

In addition to the MD simulation protocol described in Chapter 3, four variants were applied in this study. In the first variant, the simulation started from the holo (after removal of the bound inhibitor) instead of the apo structure. In the second variant, all heavy backbone atoms of the apo structures remained harmonically restrained (force constant of 1,000 kJ mol^{-1} nm^{-2}) during the production runs. The third and fourth variant of the MD simulation protocol are the simulations in methanol. Here, the apo structures were placed in cubic boxes filled with methanol molecules (using parameters from the OPLS-AA force field) and the equilibration was extended to 500 ps. The harmonic restraints were either removed in the 10 ns production run (third variant) or the harmonic restraints on the heavy backbone atoms were kept (fourth variant). Note that the second and fourth variants of the MD protocol were only applied to MDM2.

4.2.2 Generation of a Conformational Ensemble Using Normal Mode Analysis

The apo structures were minimized in vacuo without constraints using the L-BFGS algorithm until the maximum force on any atom was smaller than 0.001 kJ mol^{-1} nm^{-1}. Van der Waals interactions were calculated without cut-off and for calculating the electrostatic interactions, the relative dielectric constant ϵ was set to $4r$. The hessian matrix of the minimized structure was calculated in vacuo using the same parameters. 50 eigenvectors representing the normal modes with lowest frequencies were derived from the diagonalized mass weighted hessian matrix. As eigenvectors 1 to 6 correspond to the translational and rotational degrees of freedom of the system they were set to 0. Using the remaining 44 normal modes, 4,001 protein conformations were generated by random displacements along the eigenvectors at 300 K, where the position along each eigenvector was randomly taken from a Gaussian distribution with variance kT/eigenvalue.

4.2.3 Generation of a Conformational Ensemble Using (t)CONCOORD

The distance bounds for CONCOORD and tCONCOORD were determined from the energetically minimized structures generated for the normal mode analysis using Engh-Huber bonded parameters [151] and OPLS-AA van der Waals parameters. Based on these distance bounds 4,001 protein conformations were generated by CONCOORD and tCONCOORD. Note that the conformational ensembles generated by CONCOORD, tCONCOORD, and NMA used the same energy-minimized starting configuration as input but are otherwise unrelated.

4.2.4 Pocket Detection and Characterization Using EPOS^{BP}

In contrast to Chapter 3, we used *EPOS*^{BP}, a program that is based on BALLPass, which is the re-implemented version of the PASS algorithm that uses the BALL C++ library [152]. Note that although the PASS algorithm was implemented exactly as described in the publication [70], the number and positions of the ASPs and the PASS probes may differ from the original PASS program. The patches are calculated and clustered as described in Algorithms 1 and 2. The advantage of using EPOS^{BP} is that many procedures adjuvant when dealing with molecular structures are already implemented in the BALL library. For example, the volume of the pockets can now be determined more accurately by calculating the solvent excluded surface volume of the patches.

4.3 Results

Six conformational ensembles (each one consisting of 4,001 structures) were generated for each system and two additional ones for MDM2:

- *apo MD snapshots (water)*: snapshots extracted from MD simulations of the apo structure in water
- *apo MD snapshots (methanol)*: snapshots extracted from MD simulations of the apo structure in methanol
- *holo MD snapshots (water)*: snapshots extracted from MD simulations of the holo structure in water
- *restrained apo MD snapshots (water)*: snapshots extracted from MD simulations of the apo structure in water with harmonic restraints on all heavy backbone atoms (only for MDM2)
- *restrained apo MD snapshots (methanol)*: snapshots extracted from MD simulations of the apo structure in methanol with harmonic restraints on all heavy backbone atoms (only for MDM2)
- *CONCOORD* conformations: conformations generated by CONCOORD
- *tCONCOORD* conformations: conformations generated by tCONCOORD
- *NMA* conformations: conformations generated by deformations along normal modes

Whereas all MD simulations started from the apo or holo protein conformation taken from the crystal structures, the calculations of the CONCOORD, tCONCOORD, and NMA conformations were based on the energy-minimized conformation of the apo structure.
At first, we will present the findings for the ensembles from the various MD simulations, and then compare the results to those obtained for the CONCOORD, tCONCOORD, and NMA ensembles. However, when comparing the properties of the conformational ensembles, one should keep in mind that the apo and holo structures of BCL-X_L and MDM2 did not contain the same number of residues.

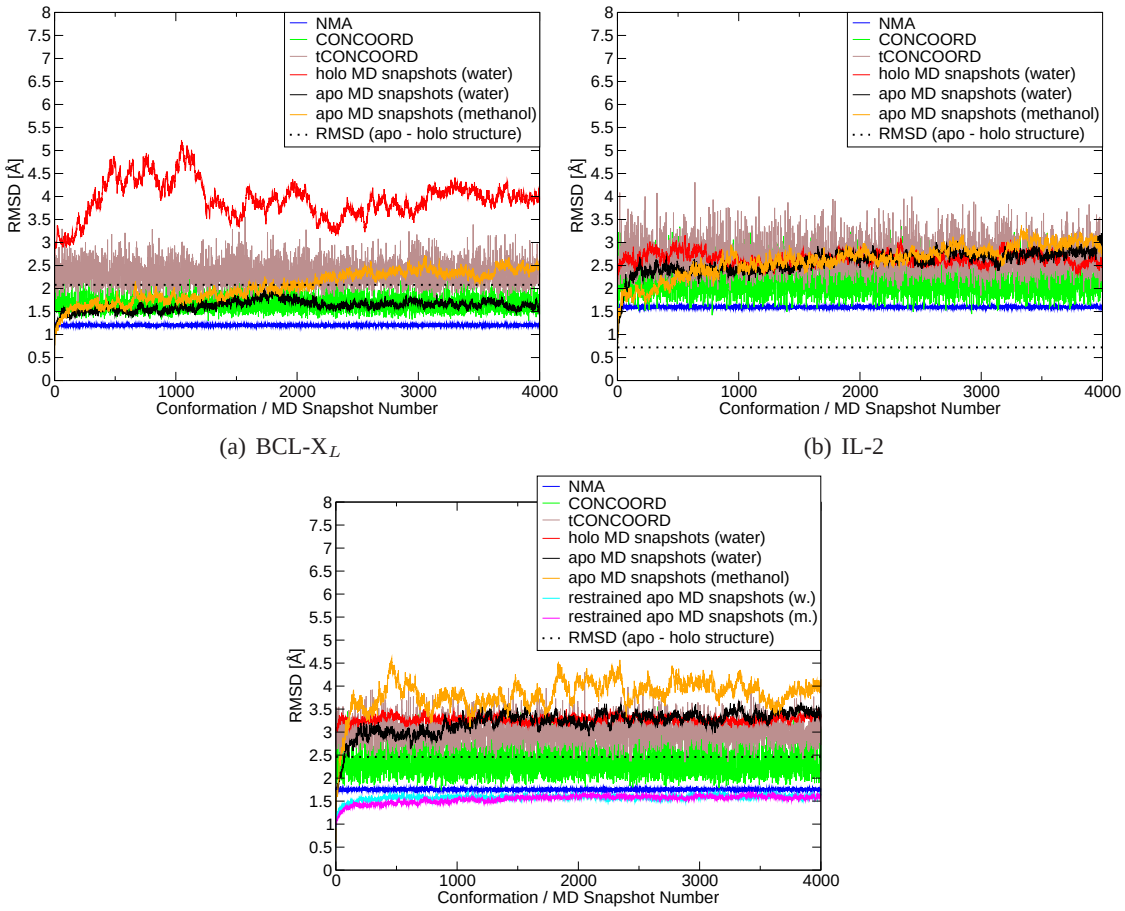
4.3.1 Pockets Detected in the Starting Structures

As we now used EPOS^{BP} instead of the original PASS program, we had to recalculate the pocket volumes and polarities of the native binding pockets in the holo and the apo structures to get new reference values. For the holo structures, we determined pocket volumes and polarities of 493.1 Å^3 and 0.26 for BCL-X_L, 400.3 Å^3 and 0.27 for IL-2, and 445.9 Å^3 and 0.25 for MDM2. In the apo structures of BCL-X_L and IL-2, the native binding pocket was only partly detectable (23.1% and 25.9% of the pocket volume relative to that of the holo structure). For apo MDM2, the native binding pocket was too compact to be detected in any NMR model.
As the CONCOORD, tCONCOORD, and NMA conformational ensembles were generated from the minimized structures of the apo proteins, it is of interest to know whether and to which extent the native binding pocket was already open in these structures. After the minimization in vacuo, the native binding pocket of BCL-X_L was closed, whereas a further opening to 74.8% was detected for IL-2. In the minimized structure of full-length MDM2, the N-terminal loop buries the native binding site. In the minimized structure of truncated MDM2, the native binding pocket opened to 50.8%.

4.3.2 Properties of the Conformational Ensembles

Before the CONCOORD, tCONCOORD, and NMA conformational ensembles were generated, the apo structures were energetically minimized. This resulted in conformations with backbone RMSDs from the apo structure of 1.0 Å for BCL-X_L, 1.4 Å for IL-2, and 6.6 Å for MDM2. For MDM2, this high value was caused by the floppy terminal loops of the NMR structure (compare Fig. 3.5) that folded back on the protein surface during the minimization in vacuo and so obstructed the p53 binding groove. This conformation appeared too compact as a starting structure. Note that such a conformation was suggested earlier [153], but in the publication describing the NMR models we used for this study, Uhrinova et al. stated that no long-range NOEs were observed for residues 2-17 [133]. This rules out the possibility that this loop occupies the p53 binding groove in a stable fashion and suggests that truncating these residues is valid. We repeated the minimization and the subsequent generation of the conformational ensembles with the stable part of the MDM2 protein (residues 17-111), to which we will refer as "truncated MDM2". This resulted in an RMSD of 1.6 Å from the apo structure.

The proteins remained stable in all MD simulations. The RMSD profiles of the different conformational ensembles of the test systems are shown in Figure 4.1. The stability of the secondary structures is discussed in Section B.1 to B.3, Appendix. The holo structure of BCL-X_L includes more loops than the apo structure (see Fig. 3.1), and, hence, the MD simulation of the holo



(a) BCL-X_L

(b) IL-2

(c) MDM2

Figure 4.1: All-atom RMSD of the six (eight, respectively) conformational ensembles from the apo structures. For MDM2, the RMSD was only calculated for residues 17 to 111.

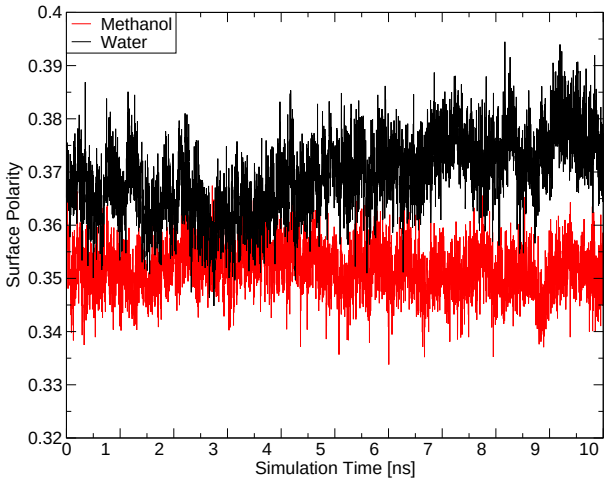


Figure 4.2: Surface polarity of MDM2 during the MD simulation in water and methanol.

structure displayed a slightly different dynamics resulting in larger RMSDs (up to 5.2 Å) than the other conformational ensembles which are based on the apo structure. The RMSD of the apo MD snapshots taken from the simulation in water stayed constantly below 2.0 Å and those taken from the simulation in methanol did not exceed 2.8 Å. Similarly, the unstable terminal loops of the apo structure are missing in the holo structure of MDM2. In the apo MD snapshots, they caused RMSDs of up to 8.0 Å for the simulation in water and up to 9.5 Å for the simulation in methanol, whereas the RMSD of residues 17 - 111 was 3.6 Å for the simulation in water and 4.6 Å for the simulation in methanol. When using harmonic restraints on the backbone atoms, the sampling was restricted to a small range around the conformation of the starting structure. Furthermore, for BCL-X_L and MDM2, the larger RMSDs observed for the snapshots extracted from the simulations in methanol suggest that the less polar solvent allowed for transitions to regions of the conformational space that were not sampled when using an aqueous solution at room temperature. In contrast, for IL-2, all MD simulations gave similar RMSD profiles.

As expected, the simulation in methanol leads to a more pronounced exposure of hydrophobic side-chains and, thus, a lowering of the overall surface polarity compared to the simulation in water (see Figure 4.2 for an example). For all three systems, CONCOORD and NMA generated conformational ensembles with much smaller RMSDs from the apo structure than the MD snapshots. The RMSD values for the NMA conformations were nearly constant (RMSD variation 0.1-0.2 Å) and only slightly larger than the RMSD of the minimized structure, whereas the RMSDs of the CONCOORD conformations varied up to 2.0 Å. In contrast, due to the enhanced conformational sampling of tCONCOORD, it generated conformations that differed up to 4.2 Å from the apo structures and, thus, are of comparable magnitude as most MD snapshots.

4.3.3 Transient Pockets Detected in the MD Snapshots

The pocket detection protocol was applied to all conformational ensembles and the properties of the identified transient pockets were analyzed. Their main properties are listed in Table 4.1. Note that it is not possible to draw any conclusions when comparing the total number of pockets (the number of pockets before clustering) and the number of distinct transient pockets (the number of pockets after clustering) detected for apo and holo MDM2 and BCL-X_L because the simulated proteins were not of equal size.

system	no. pockets before clustering	no. pockets after clustering	mean pocket volume [Å^3]	max. pocket volume [Å^3]	mean overlap volume [%]	max. overlap volume [%]
BCL-X_L						
apo MD snapshots (water)	17,079	24	375.1	1,363.9	43.5	91.3
apo MD snapshots (methanol)	22,818	23	380.7	1,357.9	48.3	95.0
holo MD snapshots (water)	39,596	46	395.2	1,606.9	38.7	94.6
CONCOORD	8,226	11	348.1	1,013.1	34.3	67.1
tCONCOORD	17,135	23	378.7	1,761.3	42.5	99.9
NMA	7,774	6	342.7	889.8	32.0	50.8
IL-2						
apo MD snapshots (water)	14,721	29	335.2	1,013.0	35.3	89.2
apo MD snapshots (methanol)	24,513	24	352.7	1,633.9	35.0	94.3
holo MD snapshots (water)	18,412	33	394.4	1,340.7	32.6	128.1[a]
CONCOORD	14,096	18	320.1	965.2	34.8	88.5
tCONCOORD	17,356	28	365.4	1,422.9	33.3	92.3
NMA	15,345	11	261.1	722.4	50.1	87.9
MDM2						
apo MD snapshots (water)	26,419	35	372.7	1,641.3	60.1	106.8[a]
apo MD snapshots (methanol)	34,022	42	407.6	2,204.5	47.6	108.0[a]
holo MD snapshots (water)	11,737	14	384.6	1,213.8	57.5	114.3[a]
restrained apo MD snapshots (w.)	21,568	17	330.5	839.9	52.9	81.2
restrained apo MD snapshots (m.)	25,204	22	351.8	1,024.3	52.2	82.0
CONCOORD[b]	13,519	10	348.6	828.2	48.0	70.1
tCONCOORD[b]	16,090	14	388.6	1,237.7	2.4	8.4
NMA[b]	14,776	7	322.2	670.2	45.4	61.4

[a] Overlap volume is larger than in the holo structure

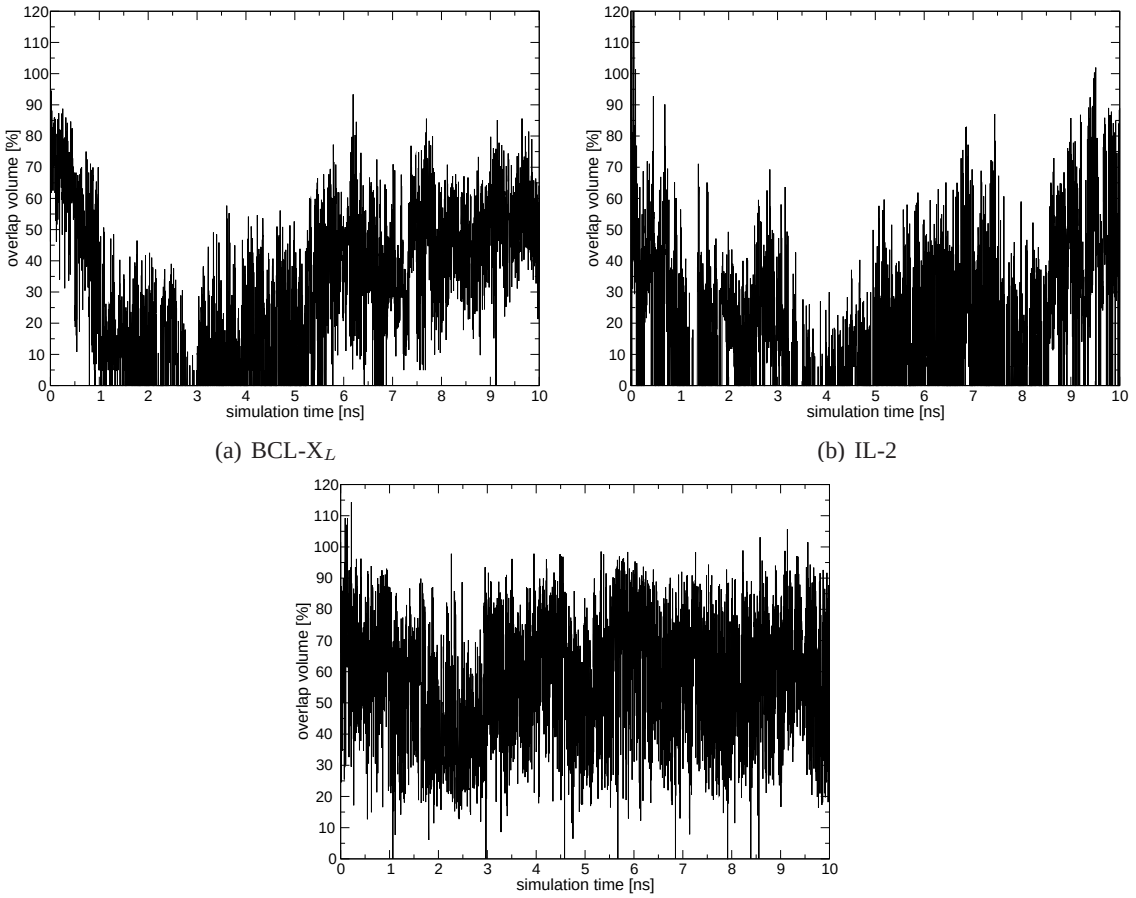
[b] minimized structure of truncated MDM2 (residues 17-111) used as starting structure

Table 4.1: Properties of the pockets detected in the conformational ensembles for each system.

Influence of the Simulation Solvent and the Backbone Flexibility on the Pocket Properties Another question we wanted to address is the influence of a less polar solvent on the pocket openings. Comparing the properties of the pockets detected in the apo MD snapshots in water and methanol reveals that for all three systems the opening of the native binding pocket seems to be eased in methanol. Besides, more (in terms of total number) and on average larger pockets opened, suggesting that the less hydrophilic methanol solvent facilitates the formation of cavities in general.

Further, we asked whether side-chain movements are sufficient for pocket openings. In all MD simulations presented so far, the whole protein was flexible so this question was hard to answer. Hence, we analyzed the pockets detected in the restrained apo MD snapshots in water and methanol of MDM2. As expected, the number of pockets (before and after clustering) and their volumes were reduced when their formation depends exclusively on side-chain movements. Interestingly, even in the restrained MD simulations the methanol solvent had the same effect on the pocket openings as in the unrestrained MD simulations. An opening of the native binding pocket was observed in all MD simulations, but its native volume was only reached when simulating without restraints in water or methanol.

Stability of the Native Binding Pocket The main reason for the simulations of the holo structures was studying the stability of the native binding pocket in the absence of a ligand. The overlap volumes indicate that in all three systems the native binding pocket fluctuated a lot during the simulation (see Figure 4.3). While the mean overlap volumes were in the same order of magnitude as those of the apo MD snapshots in water, for IL-2 and MDM2 the volume of the native binding pocket exceeded at least once its volume with the inhibitor bound. These findings indicate that the presence of a ligand is required to keep the native binding pocket fully open.



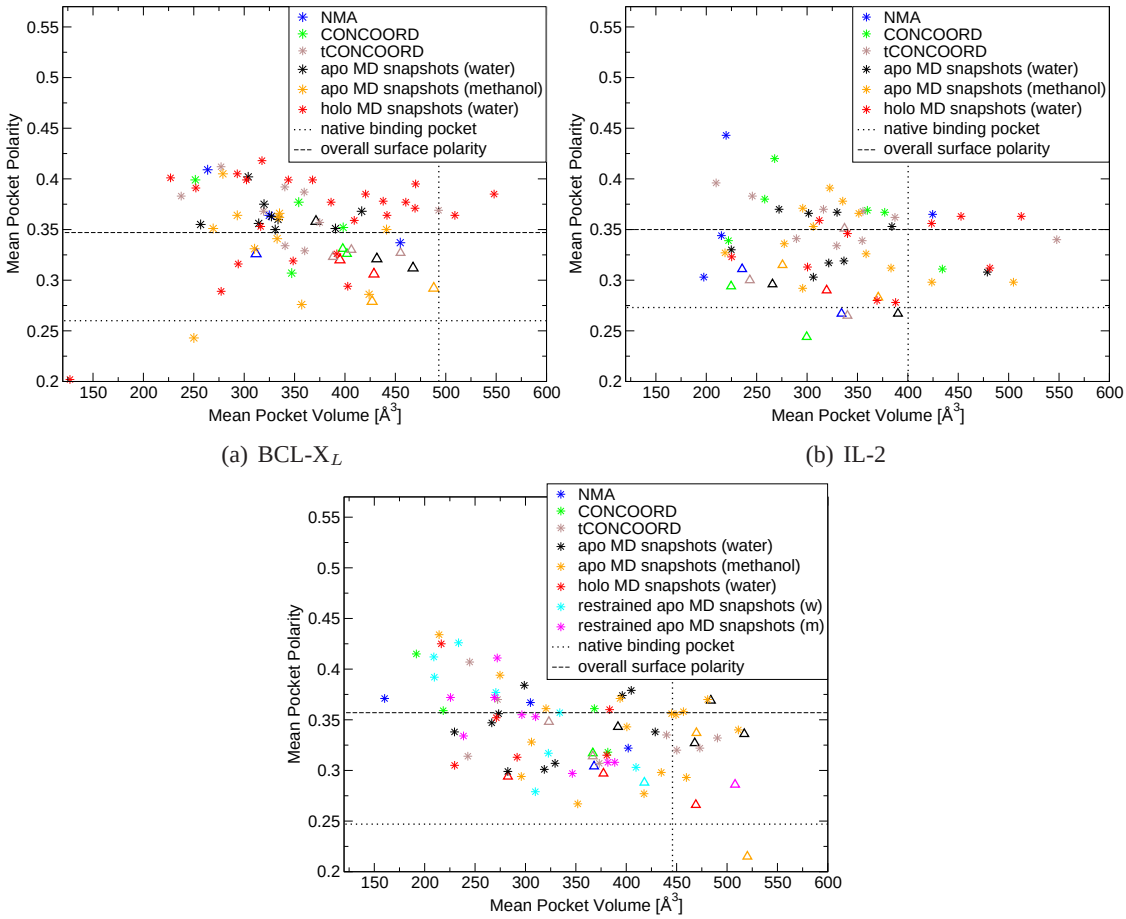
(a) BCL-X_L (b) IL-2

(c) MDM2

Figure 4.3: The stability of the native binding pockets during the MD simulations as represented by the relative overlap volume.

Do the transient pockets opening at the native binding site differ from the others? Except for the overlap volumes, all properties listed in Table 4.1 refer to pockets opening anywhere on the protein surface. A further analysis shown in Figure 4.4 addresses the differences in the properties between the pockets opening at the native binding site and those opening somewhere else on the protein surface. Note that several distinct transient pockets are opening at the protein-protein interaction interface. As already mentioned, they represent different subpockets of the native binding pocket and thus may possess different polarities. In most cases, they are on average larger than the pockets opening anywhere else and have a slightly reduced polarity. For BCL-X_L and MDM2, all transient pockets opening at the native binding site are on average more polar than the native binding pocket, except for those opening during the MD simulation in methanol. While for IL-2, transient pockets with an average volume and polarity comparable to the native binding pocket can be identified in both MD simulations of the apo structure. Moreover, this analysis demonstrates that, in general, the cavities identified in the apo MD snapshots of the simulation in methanol belong to the least polar pockets, especially when focusing on the largest pockets. Very small pockets (mean volume ≤ 200 Å^3) tend to be either very polar or very nonpolar.

When focussing on the differences between the restrained and unrestrained MD simulations of apo MDM2 shown in Figure 4.4 (c), the influence of the backbone movements on the pocket openings is evident. As already shown in Table 4.1, the transient pockets observed during the restrained MD simulations are relatively small. Pockets with mean volumes that exceed the volume



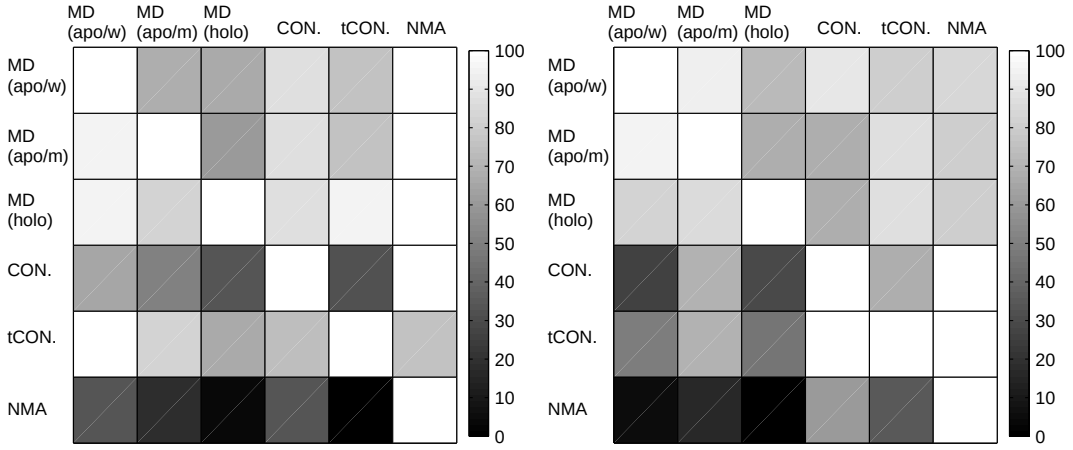
(a) BCL-X_L (b) IL-2

(c) MDM2

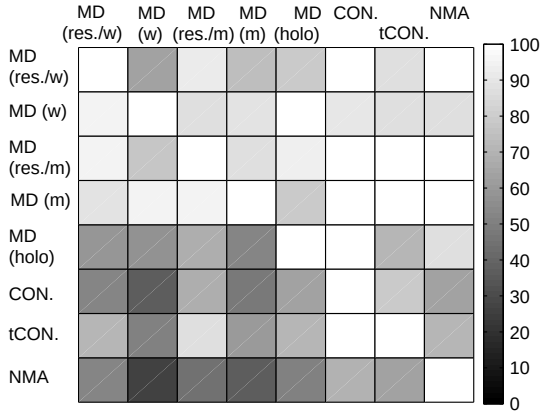
Figure 4.4: Mean volume of transient pockets plotted against their mean polarity. Only transient pockets with frequency $\geq$ 5% are shown. Triangles represent pockets opening at the native binding site, stars represent pockets opening anywhere else on the protein surface. The dotted line represents the volume and the polarity of the native binding pocket, the broken line the overall surface polarity of the apo structure.

of the native binding pocket were only found in the MD snapshots of the restrained simulation in methanol. Although in all MD simulations the transient pockets with the largest mean volume opened at the native binding site, it is not clear whether these pockets are appropriate for ligand binding as they are more polar than the native binding pocket except for those found in the unrestrained MD simulation in methanol.

Similarity of the Transient Pockets Detected in the Different Conformational Ensembles Especially when studying the influence of the backbone movements an important question arises: Are the transient pockets detected in conformational ensemble i also detected in conformational ensemble j? To investigate this question, we calculated the reproducibility of the transient pockets from different conformational ensembles as described in section 3.3.4. The results per system are shown in Figure 4.5. This analysis approves the assumption that during the simulation in methanol additional pockets open that are not observable during the simulation in water. Further, it reveals that some pockets can only open when backbone movements are allowed. This emphasizes the intrinsic influence of backbone movements on pocket openings.



(a) BCL-X_L (b) IL-2



(c) MDM2

Figure 4.5: Reproducibility of the detected transient pockets in one conformational ensemble by the transient pockets detected in another conformational ensemble (calculated as described in section 3.3.4). In each column, the percentage of reproduced transient pockets of one method by the other methods (rows) is shown. These plots show that, for example, the tCONCOORD ensemble contains more pockets also opening during MD simulations than the CONCOORD ensemble. This observation supports the importance of solvent effects because tCONCOORD is not considering hydrogen bonds that may be attacked by solvent molecules in the definition of the distance constraints.

4.3.4 Which transient pockets are suitable for accommodating known inhibitors?

So far, we used the overlap volume to estimate how far the native binding pocket opened. However, this measure is only a rough estimate and it is unclear whether an overlap value of 50-90% is sufficient to accommodate a ligand. Therefore, the docking experiments described in Chapter 3 in which the definition grid center was based on the center of mass of the transient pocket were repeated. All transient pockets that opened at the interface in all snapshots extracted from MD simulation of the apo structure in water or methanol were used as starting points. The best docking results listed in Table 4.2 emphasize that the bahavior of the native binding pocket of IL-2 differs from that of BCL-X_L and MDM2. While for the latter two systems much better docking results were obtained when docking into MD snapshots taken from the simulation in methanol (as

system	RMSD [Å]	score [kcal/mol]	score rank[a] [%]
BCL-X_L			
apo MD snapshots (water)	1.9	-10.2	1.5
apo MD snapshots (methanol)	1.7	-11.8	0.5
CONCOORD	1.8	-7.6	29.0
tCONCOORD	2.0	-11.1	1.5
NMA	2.3	-8.1	16.3
IL-2			
apo MD snapshots (water)	1.9	-8.5	0.7
apo MD snapshots (methanol)	2.0	-6.9	8.4
CONCOORD	2.0	-6.6	7.9
tCONCOORD	2.4	-5.4	25.8
NMA	1.8	-7.0	7.1
MDM2			
apo MD snapshots (water)	1.5	-11.8	8.0
apo MD snapshots (methanol)	1.7	-13.5	0.2
restrained apo MD snapshots (water)	2.0	-8.8	82.6
restrained apo MD snapshots (methanol)	2.0	-9.5	68.5
CONCOORD[b]	2.1	-8.7	61.6
tCONCOORD[b]	1.6	-11.2	0.1
NMA[b]	3.3	-9.7	13.0

[a] relative rank defined as the rank of this solution after sorting all results by increasing docking score in relation to the total number of docking results

[b] minimized structure of truncated MDM2 (residues 17-111) used as starting structure

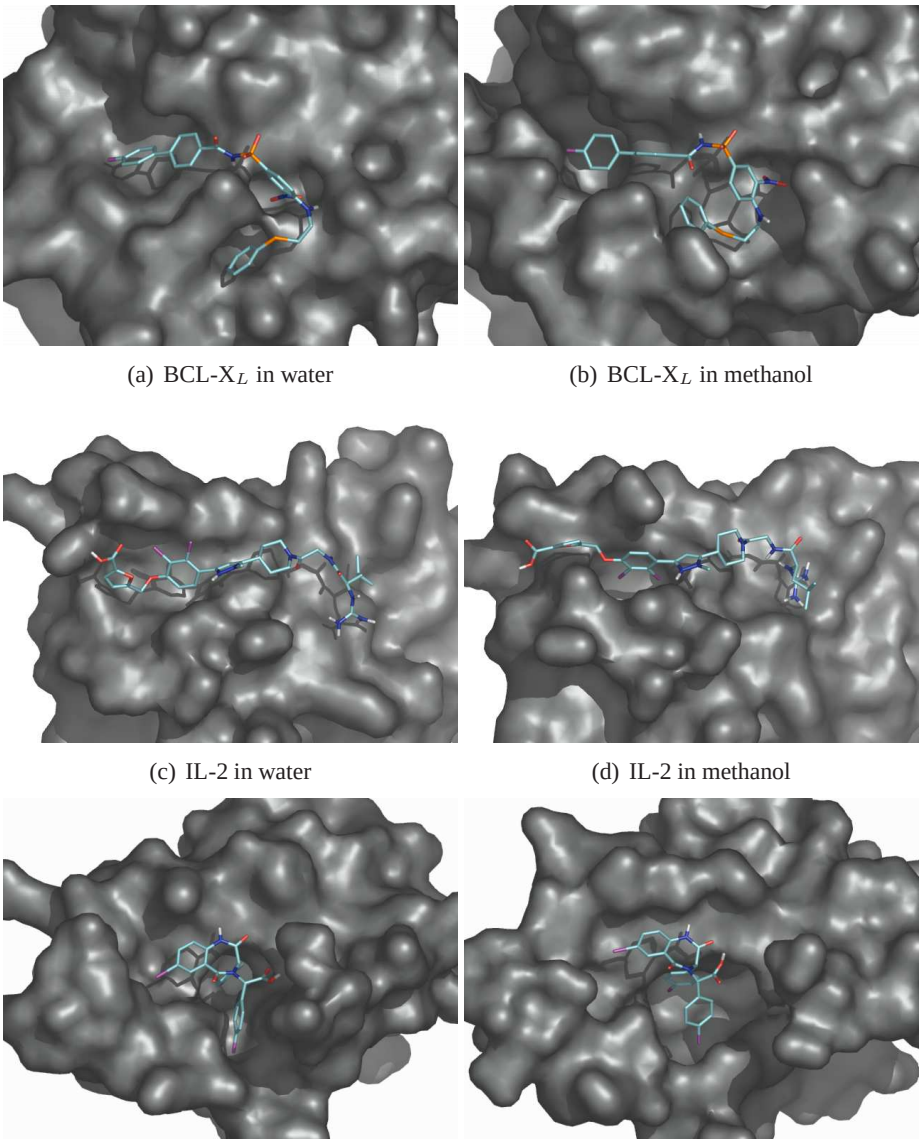
Table 4.2: Best ranked correct (RMSD $\leq$ 2 Å) docking results or docking results with lowest RMSD per conformational ensemble and system.

reflected by the better docking scores and the reduced score rank), docking into water snapshots gave better results for IL-2 (although the maximal overlap in Table 4.1 gives another impression). A possible explanation for this may be that the pockets opening in methanol at the native binding site are too small for the native ligand (see Fig. 4.4 (b)), even though the pockets are on average larger than those opening in water. The best scored docking poses for the MD simulations of the apo structures are shown in Figure 4.6.
Surprisingly, when docking the inhibitor into snapshots extracted from restrained MD simulations of MDM2 the native binding mode was correctly predicted. However, the docking scores and their ranks are worse compared to the unrestrained simulations indicating that although side-chain movements are sufficient to open new cavities, suitable backbone movements are also needed to achieve enough depth and plasticity. Again, the results got slightly better when docking into snapshots from the simulation in methanol. The impact of methanol as solvent is most striking for the snapshots taken from the unrestrained MD simulations of MDM2. Here, the relative score rank of a correct docking result improved from 8.0% to 0.2% when simulating in methanol instead of water. Moreover, the score improved by 1.7 kcal/mol suggesting that, in addition to backbone movements, the effect of a less polar solvent promotes the opening of pockets even further.

4.3.5 Are CONCOORD, tCONCOORD, or NMA conformations an alternative to MD snapshots?

Molecular dynamics simulations are quite time consuming. For this reason, it would be desirable to replace them by a more efficient method. Potential alternatives to MD snapshots are conformations generated by CONCOORD, tCONCOORD, or NMA.

Properties of Transient Pockets Detected in the Alternative Conformational Ensembles The properties of the transient pockets detected in these conformational ensembles are listed in Table 4.1. For IL-2, the total number of cavities found in the CONCOORD and the NMA conformational



(a) BCL-X_L in water (b) BCL-X_L in methanol

(c) IL-2 in water (d) IL-2 in methanol

(e) MDM2 in water (f) MDM2 in methanol

Figure 4.6: The best scored docking poses when docking into MD snapshots extracted from the simulation of the apo structure in water and in methanol (corresponding to those listed in Table 4.2). In (e) and (f), residues 1 to 16 were removed for better visibility.

ensemble are comparable to the total number of cavities found in the apo MD snapshots from the simulation in water. In contrast, for BCL-X_L and MDM2 the number of detected pocket openings is significantly reduced. Anyhow, the pockets found in these ensembles are not as diverse (indicated by the number of pockets after clustering) as those opening during MD simulations. Particularly when using NMA for generating a conformational ensemble one ends up with only a few distinct transient pockets. In addition, these pockets are smaller than those opening during the MD simulations. More importantly, they are also smaller than the native binding pockets and too polar, except for the pockets detected for IL-2. In summary, for all three systems, varying interatomic distances in the CONCOORD approach or random deformations along normal modes approach sometimes resulted in a certain opening or enlargement of the native binding pocket, but not to the same extent as observed in MD simulations. Only for IL-2, the maximum overlap volumes of the NMA and CONCOORD conformations are of the same magnitude as those of the

apo MD snapshots in water.
The sampling of transient pockets can be significantly improved when using the tCONCOORD method. This method gives a larger total number of detected cavities and more diverse transient pockets as indicated by the number of pockets after clustering. Besides, these pockets are considerably larger than those detected in the CONCOORD and NMA conformational ensembles. Their volumes are even of the same order of magnitude as the pockets opening during the MD simulations, but except for IL-2, they are too polar (see Fig. 4.4).

Are these conformations approriate to accommodate the known inhibitors? Here again, it is interesting to know whether the transient pockets identified for these conformational ensembles are the same as those observed during MD simulations. Figure 4.5 shows that although NMA, CONCOORD, and tCONCOORD conformations are based on the same starting structure, the tCONCOORD conformational ensemble performs best in reproducing pockets detected during MD simulations. Most pockets found in NMA structures were also found in all other conformational ensembles, because, as stated above, the NMA conformations tend to be quite similar and so possess only a small number of distinct pockets. This finding emphasizes that slow normal modes are only involved in the dynamics of a few pocket opening regions. Whether the native binding site belongs to these regions can be addressed by docking into the NMA conformations. Besides the transient pockets that opened at the interface in this conformational ensemble, those detected in the CONCOORD and tCONCOORD conformations were also used as starting points for docking experiments. When analyzing these docking results (see Table 4.2) it is obvious that the opening of the native binding pocket of IL-2 seems to be driven by other dynamics than that of BCL-X_L and MDM2. Namely, for IL-2, RMSD values below 2 Å could be achieved when docking into NMA conformations but not for the other two systems. In contrast, when using the CONCOORD conformations, the native binding mode of the known inhibitors could be more or less reproduced for each system. However, the scores of these results demonstrate that the structural and/or the biochemical environment is not as appropriate as it may get when using MD simulations without restraints. The docking results for the tCONCOORD conformational ensembles were quite surprising. While for IL-2, the docking results were significantly worse than those for docking into the CONCOORD conformations, the results for BCL-X_L and MDM2 were comparable to those obtained when docking into the apo MD snapshots extracted from the simulations in water. This emphasizes the ability of tCONCOORD of sampling ligand-bound conformations even if the unbound structure was used as input and suggests that at least for BCL-X_L and MDM2, tCONCOORD seems to be an efficient alternative to MD simulations.

4.4 Discussion

As we have shown in Chapter 3, transient pockets of similar size as when bound to an known inhibitor open during MD simulations of apo proteins at the native binding sites. These pockets are not only observed by chance, but they were reproducible in a second MD simulation under the same conditions. The results of this chapter indicate that most pockets are also reproducible in MD simulations under different conditions. Here again, some of these pockets opened at the native binding site and were appropriate for ligand binding. However, the properties of the detected cavities depended crucially on the complexation status of the starting structure (apo vs. holo) and the solvent. We calculated the pocket properties for three test systems and the general impressions were quite similar. When the holo structure was used in the MD simulations, the volume of the native binding pocket showed the largest fluctuations. Although the same starting structure was used, more and larger pockets opened on the protein surface during the simulation in methanol than during the simulation in water. Furthermore, the restrained simulation of MDM2 showed that

side-chain movements alone indeed lead to the formation of pockets, but their number and volume is reduced. In summary, these findings suggest that pocket openings are induced by movements of the protein backbone and side-chains that are coupled to the solvent.

4.4.1 Can MD simulations be replaced by a more efficient method?

Besides the transient pockets opening during MD simulations conducted under different conditions, we also analyzed pockets detected in conformational ensembles generated by CONCORD, tCONCOORD, and NMA. Almost all of these transient pockets were also found during MD simulations. However, the number of distinct cavities is limited in the CONCOORD and NMA conformations, and additionally they are relatively small. In most cases they are not appropriate for ligand binding as the quality of the docking results compared to those of the MD snapshots demonstrates. Hence, due to their neglect of solvent effects the applicability of CONCOORD or NMA for the purpose of inducing pocket openings appears to be limited. This problem seems to be overcome in tCONCOORD by not considering hydrogen bonds that may be attacked by solvent molecules in the definition of the distance constraints. As intended by the authors, this enables an enhanced conformational sampling compared to CONCOORD [59]. For our purpose, this means the detection of transient pockets of an increased variety and volumes that are even comparable to that of pockets opening during MD simulations.

4.4.2 Are pocket openings related to normal modes?

Although the overlap volumes in NMA conformations suggest that the opening of the native binding pocket is somehow related to deformations along slow normal modes, they are not sufficient to induce full pocket openings. However, compared to the docking results obtained when docking into the apo structures (see Table 3.5), using NMA conformations led to significant improvements. For MDM2, the docking score even improves by 3 kcal/mol. This finding indicates that the opening of the native binding pocket of MDM2 is weakly related to normal modes is in agreement with the results reported by Barrett et al. [136] and Espinoza-Fonseca and Trujillo-Ferrara [137]. Note that these authors only observed the opening and closing movement and did not measure whether the increase of the pocket volume is sufficient for ligand binding.
IL-2 was the only system for which deformations along the normal modes were sufficient to reproduce approximately the native binding pocket (see Table 4.2). One should keep in mind, however, that here the native binding pocket was already open to almost 75% in the starting structure used for generating the NMA conformations. Therefore, it is an educated guess that the opening of the native binding pocket appears energetically quite favorable and may be observed by a variety of methods that sample low-energy conformations. Indeed, this was also true for CONCOORD, but not for tCONCOORD. We assume that this is due to tCONCOORD's ability of sampling structural transitions. As the input structure was already quite similar to the holo structure, only regions of the conformational space that are further away from these structures were sampled. Thus, no native-like binding pose could be found when docking into the tCONCOORD conformational ensemble. On the other hand, in the case of BCL-X_L, minimization resulted in the closure of the previously partly open native binding pocket and here CONCOORD and tCONCOORD successfully produced conformations that were able to accommodate the known inhibitor.

4.4.3 Critical Assessment of the Approach

Besides studying what induces the opening of transient pockets, the aim of this chapter was to test whether the time-consuming MD simulations may be replaced by a more efficient method. Note that as already discussed in the previous chapter, only a few suitable model systems were available

for this investigation making it difficult to extract generic conclusions. The results were quite promising for tCONCOORD, but the MD simulation in water was the only method that performed equally well for all three systems. Although more, larger, and less polar pockets were detected when simulating in methanol, the results improved for only two of the three systems. These observations suggest that the openings of the native binding pockets of the three studied systems are driven by different mechanisms and none of the studied methods was capable of generating conformational ensembles of all systems that contain the native binding pocket in a druggable state.
Note that the score ranks listed in Table 4.2 are quite low because we only docked in those pockets opening at the interface. However, this study indicated that in general the native binding pocket differs from other pockets by its volume and polarity (see Figure 4.4) and, thus, may be identified by docking without prior knowledge about the location of the binding site when focussing on large, nonpolar pockets. But, of course, this also has to be validated using a larger number of model systems.

4.5 Summary and Conclusion

In this chapter we extended our investigation of transient pockets opening on protein surfaces and analyzed what induces these openings. A significant impact of backbone movements and of the solvent was identified. This was evident from the simulation in methanol were the total number of pocket openings and their volumes increased compared to the simulation in water. For two out of the three systems, this also led to the formation of nonpolar pockets at the interface what significantly improved the docking results. This suggests that a more hydrophobic solvent facilitates the opening of the native binding pocket. Comparing MD simulations with full flexibility or with harmonically restrained backbone atoms revealed that although side-chain movements alone lead to the formation of surface cavities, the required depth and plasticity for ligand binding can only be achieved by including the backbone movements. Additionally, we could show that the volume of native binding pockets fluctuates significantly, suggesting a decreased stability in the absence of a ligand. By calculating the reproducibility of the transient pockets detected in the different MD simulations, we could show that the opening of most pockets is independent from the starting structure and the solvent. Moreover, we tested more efficient methods for generating conformational ensembles, but although CONCOORD and NMA were capable of producing conformations with pockets not observable in the starting structure, their diversity and volume was limited. Though the formation of some pockets is coupled to low frequency normal modes, deformations along these modes were not sufficient to achieve full pocket openings. On the other hand, conformations generated by tCONCOORD possessed pockets with volumes and diversity that were comparable to those of pockets opening during MD simulations. For two out of the three test systems, this method was even able to generate conformations suitable to accommodate the native ligand. These findings open promising avenues for structure-based drug design on protein surfaces.

Chapter 5

Designing Binding Pockets on Protein Surfaces using the A* Algorithm

While the protocols discussed in Chapter 3 and 4 sampled the entire protein surface for pockets, we will now present a rigorous algorithmic approach that induces the opening of putative binding pockets at predefined surface regions. This initial study was published as a full paper in the Proceedings of the *German Conference on Bioinformatics* in 2008 [154].

5.1 Introduction

In the previous two chapters, we presented a protocol that detected pockets that opened on the protein surface. The advantage of this protocol is that no prior knowledge about the location of the binding site is required because the entire protein surface is sampled for transient pockets. However, only those pockets that open spontaneously will be identified by this protocol. Pockets whose openings are induced by a nearby ligand will remain undetected. Furthermore, the pockets are often too small and/or too polar for ligand binding. Fortunately, in many drug design projects, the approximate location of the binding site is already known. Hence, it is sufficient to sample only a part of the protein surface. This local instead of global search allows for a more accurate and directed sampling of accessible protein pockets and so also allows to "force" the opening of a pocket at a known location. The resulting protein conformations and their ligand binding pockets can then be used to optimize the interaction between the protein and the ligand or for virtual screening.

The problem of finding appropriate protein conformations can be solved efficiently using an informed search. For this purpose, several algorithms have been developed in artificial intelligence. A popular example implementing an informed graph search is the A* algorithm [155] that uses knowledge about the structure of the search space incorporated in heuristic functions to guide the search towards optimal solutions. The nodes of the graph represent states of the system. Given the initial state represented by the start node, the algorithm searches an optimal (i.e. minimal cost) path to a given goal node, representing the goal state. During this search, a graph is built up in which each node represents a partial solution. The generated nodes are maintained in a priority queue. The priority of a partial solution *x* is given by

$$f(x) = g(x) + h(x) \tag{5.1}$$

where *g(x)* is the cost of this partial solution so far, i.e. from the start node to *x* and *h(x)* is the heuristic estimate of the minimal cost to reach the goal node from *x*. If the heuristic function is admissible (i.e. it never overestimates the cost of reaching the goal node) and consistent (i.e. it fulfills the triangle inequality), it will always find a path with minimal cost from a given start node to a given goal node if such a goal node exists. As already mentioned in Chapter 2, Leach applied the A* search to the flexible docking and the side-chain placement problems [97]. After placing an

anchor region of the ligand into the binding site, all possible ligand conformations were generated. For each conformation that made no unfavorable interactions with the protein backbone and all rotameric states of a residue, the optimal combination of side-chain rotamers was determined by an A* search. In this approach, the initial node represented the structure without assigned rotamers for the residues at the binding site, while the goal nodes represented the optimal docking solutions where all residues had assigned rotamers.

We incorporated ideas from PASS, MCSS, and Leach's application of the A*-search into a new approach that algorithmically generates energetically favorable protein conformations with accessible binding pockets at defined locations on protein surfaces. Since conformations of minimal energy and conformations with large cavities may be incompatible, the user can control this compromise. Based on our findings about the importance of backbone movements presented in the previous chapter, this method considers full protein flexibility during the design of protein conformations. As before, the applicability of the approach was validated using the proteins BCL-X_L, IL-2, and MDM2.

During the implementation of this method, two approaches were published that modify the ligand binding sites. Bottegoni et al. developed SCARE, an induced-fit docking protocol that starts from a single (apo) input structure [156]. By mutating different pocket residues to alanine, multiple variants of the binding pocket are generated into which the flexible ligand is docked. The best scored poses are kept, the residues are mutated back, and the receptor pocket is optimized globally (thus allowing for backbone and side-chain flexibility) while the positions of the ligand atoms are restrained. After re-scoring the optimized docking complexes, the ligand binding pose was correctly predicted (RMSD $\leq$ 2 Å) in 80% of the best scored conformations. The authors emphasize that no prior knowledge about the binding site location is required because they run a pocket detection program and pick the largest pocket. This method would most probably fail for protein-protein interaction interfaces as it is limited to conformations that already contain accessible pockets. In another publication, Withers et al. presented "active site pressurization", an approach for predicting the deformability of protein pockets [157]. During a MD simulation a rectangular block of Lennard-Jones particles is injected into the ligand binding site and the number of particles interacting with the protein is gradually increased. Thereby new energetically reasonable protein conformations are generated that may be more appropriate for ligand binding than the starting structure. But rather than inducing the opening of new cavities, this method is designed to enlarge existing pockets.

5.2 Methods and Materials

Here, we introduce two programs for the generation of protein conformations that possess putative binding pockets: *PocketScanner* and *PocketBuilder*. An overview of this approach is depicted in Figure 5.1. PocketScanner scans a user-defined region of the protein surface for energetically favorable pocket positions by generating conformations with preformed pockets at these sites. Subsequently, PocketBuilder refines these intermediate conformations and designs a final set of conformations that best fulfill the search criteria, namely the desired trade-off between a protein conformation with low-energy side-chain rotamers and a pocket of defined volume. Both programs were implemented in C++ using the BALL library. All energies are computed using the CHARMM EEF1 force field [158] that treats the solvent as an implicit continuum because including such effects is crucial for designing binding pockets on protein surfaces. We added so-called *generic pocket spheres* (GPS) to the force field. Each pocket was represented by a GPS that only interacts with the protein atoms via van der Waals interactions (with a radius of 1, 2, or 3 Å and a well depth of 0.05 kcal/mol). Note that the volume of a designed pocket is controlled by the radius of the GPS while it is represented by the van der Waals interaction energy between the pro-

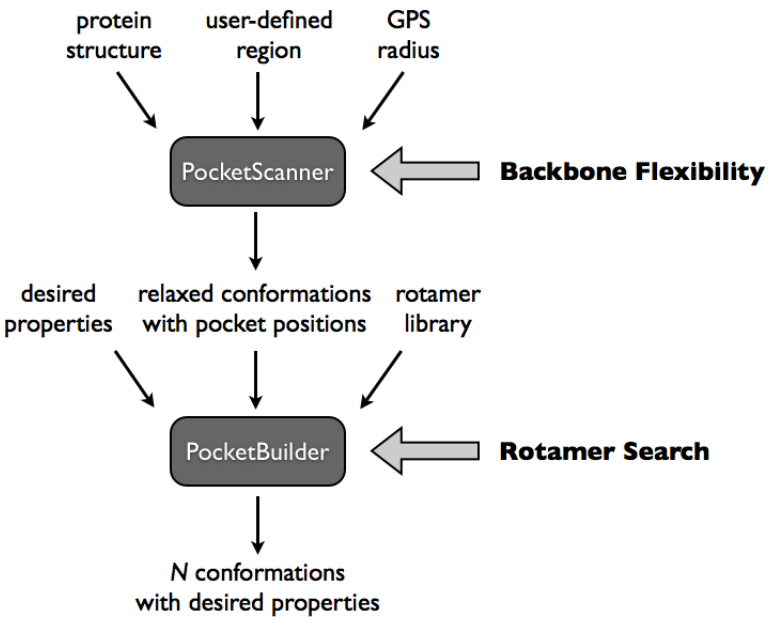


Figure 5.1: The PocketScanner / PocketBuilder approach. PocketScanner requires as input a starting structure, a region that should be scanned for pockets, and a radius of the GPS controlling the volume of the pocket. For each putative pocket position, the input structure is then energetically minimized in the presence of a GPS in order to adopt the protein conformation (including the backbone) to a pocket at this position. The generated conformations and the corresponding pocket positions are then used by PocketBuilder, along with a rotamer library and weights controlling the properties of the final conformations. The A* algorithm then searches for the best combination of side-chain rotamers such that the resulting final conformations possess the user-defined trade-off between low energy and large accessible pocket.

tein atoms and the GPS. For the analysis of the pockets, the pocket volumes and polarities were calculated by EPOS^{BP} as described on page 71.

We used exactly the same prepared apo, holo, and inhibitor structures as in Chapters 3 and 4. Docking experiments were performed as described in Chapter 3, but here, the positions of the GPSs were used as grid centers.

5.2.1 The PocketScanner Algorithm

In order to scan the protein surface for potential pocket positions, a grid with a user-defined center, dimensions, and edge length is placed on the protein surface. The z-axis of this grid is the solvent vector defined by the grid center and the center of gravity of the 10 nearest solvent exposed atoms. A GPS of given radius representing the pocket center is then successively placed on each grid point having a burial count (number of protein atoms within 8 Å) above a given threshold (here: 65). Thereby, we ensure that pockets are only induced at positions of high protein atom density. To exclude pocket positions that are deeply buried inside the protein, we additionally require that the minimal distance to any solvent exposed atom must be smaller than 2 Å. The protein is then energically minimized in the presence of the GPS using 500 steps of L-BFGS or until the RMS gradient is smaller than 0.01 kcal mol^{-1}Å^{-1}. During this energy minimization, the position of the GPS is fixed, so that the protein has to adopt its conformation. This relaxation may either result in the formation of a cavity or in a flattening of the protein surface. Thus, only if the burial count is still high enough after the energy minimization, this protein conformation in combination with this pocket position is written to an output file for using it as a starting conformation in PocketBuilder. The complete procedure is listed in Algorithm 3.

Algorithm 3 The PocketScanner algorithm

Input: $start\_conf \leftarrow$ apo protein structure
Input: $grid\_parameter \leftarrow$ center, x-, y-, and z-dimension, edge length
Input: $radius \leftarrow$ radius of the GPS
$generated\_confs \leftarrow \emptyset$
$grid \leftarrow makeGrid(grid\_parameter)$ {creates a grid such that the z-vector is approx. perpendicular to the protein surface}
for each $grid\_point \in grid$ **do**
  $BC \leftarrow getBurialCount(grid\_point, start\_conf)$ {calculate the burial count of this position}
  $dist\_to\_surface \leftarrow getMinimalDistanceToSEAtom(grid\_point, start\_conf)$ {calculate the distance to the nearest solvent exposed atom}
  if $(BC \geq 65)$ AND $(dist\_to\_surface \leq 2)$ **then**
    $GPS \leftarrow generateGPS(grid\_point, radius)$
    $conf\_with\_GPS \leftarrow start\_conf \cup GPS$
    $conf\_with\_GPS \leftarrow runEM(start\_conf, GPS)$ {fix the position of the GPS and energy minimize the protein}
    $BC_{minimized} \leftarrow getBurialCount(grid\_point, conf\_with\_GPS \setminus GPS)$ {recalculate the burial count}
    if $BC_{minimized} \geq 65$ **then**
      $generated\_confs \leftarrow generated\_confs \cup conf\_with\_GPS$ {save this conformation with the corresponding GPS}
    end if
  end if
end for
return $generated\_confs$

5.2.2 The PocketBuilder Algorithm

Starting from the protein conformations with preformed pockets generated by PocketScanner, PocketBuilder calculates a user-defined number of conformations that best represent the selected trade-off between a pocket of a given volume and a protein conformation of low energy. The algorithm consists of two stages: the initialization stage and the A*-search. The pseudocode of the program can be found in Algorithm 4. The initialization is performed separately for each starting conformation. It starts with defining all side-chains within 8 Å of the GPS as flexible. The remaining part of the protein is treated as rigid. For this part, the energy E_{rigid} and the van der Waals interaction energy with the GPS $E_{rigid,pocket}$ are calculated. For each of the flexible residues i, all rotamers j taken from Dunbrack's backbone independent rotamer library from 2002 [64] (including the original side-chain conformation), the van der Waals interaction energy with the pocket $E_{i_j,pocket}$, and the energy change ΔE_{i_j} resulting from including this side-chain rotamer in the calculation of E_{rigid} are determined. Unfavorable rotamers are deleted if

$$E_{i_j}^{weighted} = w_{energy} \cdot \Delta E_{i_j} + w_{pocket} \cdot E_{i_j,pocket} \geq 100 kcal/mol \tag{5.2}$$

Finally, the pairwise non-bonded interaction energies E_{i_j,k_l} between the remaining rotamers *j* and *l* of each pair of residues *i* and *k* are calculated and stored in a hash table.

Afterwards, the algorithm builds up a search tree where the nodes represent partial conformations, i.e. conformations in which rotameric states have only been assigned to some of the flexible residues. The nodes of the first level in the tree correspond to the rigid part of the different input conformations (0 assigned side-chains), then in each subsequent level $i+1$, rotamers are assigned to each flexible residue i until the protein conformation is complete (see also Figure 5.2). The different input conformations represent different subtrees in which the rigid part of the protein as well as the pocket position is identical. The order in which side-chain rotamers are assigned is fixed, so that in all partial solutions represented by nodes of level $i+1$ in subtree s, the side-chains of the same residues $0, ..., i$ are already defined. (The order in which side-chains are added has no effect on the final result.) Note that the level of the leave nodes are identical within the same subtree, but may differ within different subtrees depending on the number of flexible residues defined for this input conformation. The buildup of the tree is controlled by the A* algorithm. A priority $f(x)$ is assigned to each node x that evaluates the true cost $g(x)$ of this partial conformation so far

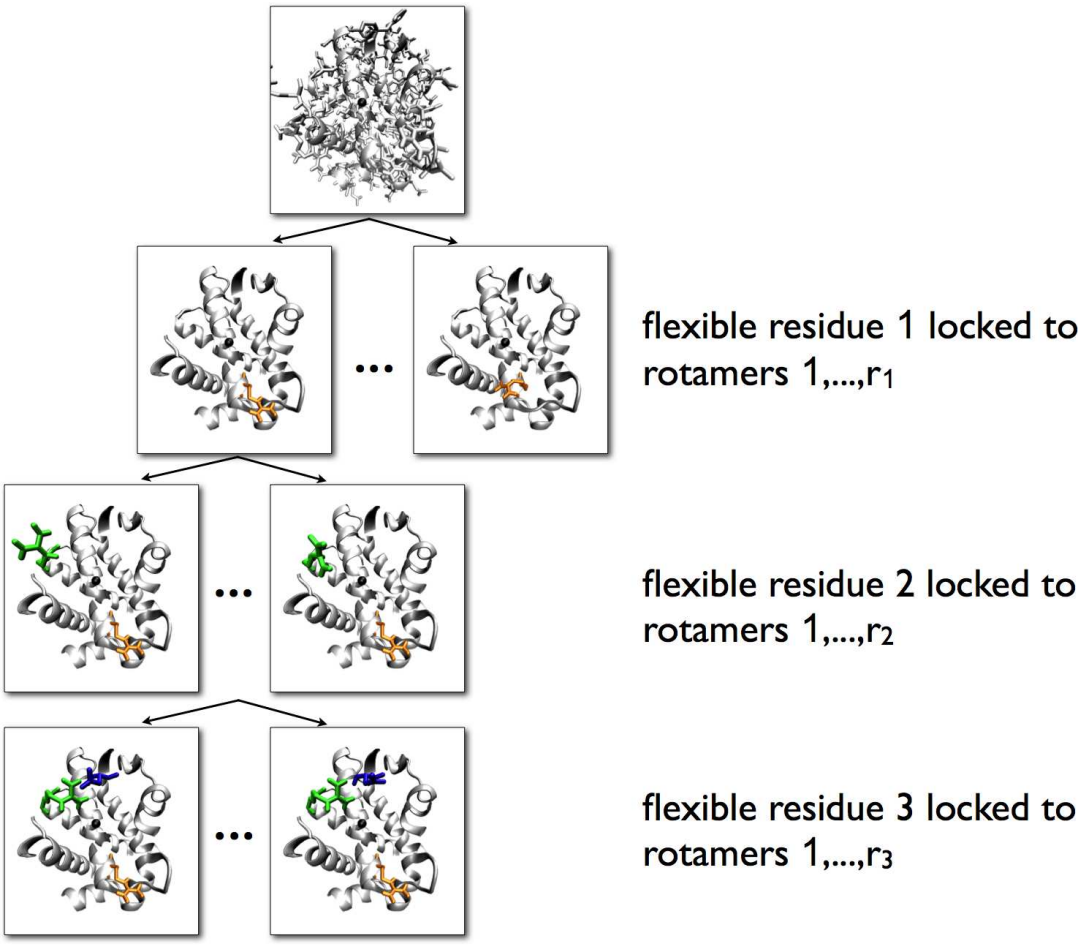


Figure 5.2: A part of the rotamer search tree generated during the A* search of PocketBuilder, shown here for a single starting conformation. The 1st level contains the starting conformation with all residues within 8 Å of the GPS (shown in black) mutated to glycine. If this node is chosen for expansion, r_1 nodes are added in the 2nd level, and each node represents a partial solution in which the first flexible residue is locked into the different rotameric states $1, ..., r_1$ (shown in orange). If a node in the 2nd level is expanded, r_2 nodes are added in the 3rd level per node and here, the second flexible residue gets locked into the different rotameric states $1, ..., r_2$ (shown in green). In the 4th level, the third flexible residue gets locked into the different rotameric states $1, ..., r_3$ (shown in blue), and so on until a leaf node is reached where rotamers are assigned to each flexible residue. For better visibility, the backbone of the partial solutions starting from the 2nd level is shown in white cartoon representation and only the flexible residues are shown in colors.

and estimates the minimal cost $h(x)$ for reaching a leaf node, where

$$\begin{aligned} g(x) &= w_{pocket} \cdot E_{rigid,pocket} + w_{energy} \cdot E_{rigid} + \\ &\quad \sum_{i=1}^{x} \left(w_{pocket} \cdot E_{i_r,pocket} + w_{energy} \cdot \left(\Delta E_{i_r} + \sum_{k=1}^{i-1} E_{i_r,k_r} \right) \right) \end{aligned} \tag{5.3}$$

$$\begin{aligned} h(x) &= \sum_{k=x+1}^{N} \min_{l} \left(w_{energy} \cdot \Delta E_{k_l} + w_{pocket} \cdot E_{k_l,pocket} \right) + \\ &\quad \sum_{k=x+1}^{N} \left(\left(\sum_{i=1}^{x} \min_{l} E_{i_r,k_l} \right) + \left(\sum_{n=x+2}^{N} \min_{l,m} E_{k_l,n_m} \right) \right) \end{aligned} \tag{5.4}$$

Algorithm 4 The PocketBuilder algorithm

Input: $start\_confs \leftarrow$ PocketScanner conformations with GPS
Input: $flexibility\_radius \leftarrow$ radius defining which side-chains around the GPS are to be optimized
Input: $w_{energy} \leftarrow$ weighting factor for scoring the internal protein energy
Input: $w_{pocket} \leftarrow$ weighting factor for scoring the pocket volume
Input: $N \leftarrow$ number of protein conformations to be generated
{initialization}
for each $conf \in start\_confs$ **do**
  $flexible\_residues \leftarrow getFlexibleResidues(conf, conf.GPS, flexibility\_radius)$ {get all residues within a certain distance of the GPS}
  $rigid\_part \leftarrow conf \setminus (flexible\_residues \cup conf.GPS)$ {get the rigid part of the protein}
  $E_{rigid} \leftarrow getEnergy(rigid\_part)$ {get the internal protein energy of the rigid part}
  $E_{rigid,pocket} \leftarrow getEnergy(rigid\_part, conf.GPS)$ {get the vdW interaction energy between the GPS and the rigid part}
  for each $i \in flexible\_residues$ **do**
    $rotamers_i \leftarrow getRotamers(i) \cup i$ {get all rotamers of i and also add the original side-chain conformation}
    for each $j \in rotamers_i$ **do**
      $E_{i_j,pocket} \leftarrow getEnergy(i_j, conf.GPS)$ {get the vdW interaction energy between the GPS and the rotamer}
      $\Delta E_{i_j} \leftarrow getEnergy(rigid\_part \cup i_j) - E_{rigid}$ {get the change in internal protein energy resulting from including this rotamer}
      if $w_{energy} \cdot \Delta E_{i_j} + w_{pocket} \cdot E_{i_j,pocket} \geq 100 kcal/mol$ **then**
        $rotamers_i \leftarrow rotamers_i \setminus j$ {remove unfavorable rotamers to speed up the calculations}
      end if
    end for
  end for
  for each $i \in flexible\_residues$ **do**
    for each $j \in rotamers_i$ **do**
      for each $k \neq i \in flexible\_residues$ **do**
        for each $l \in rotamers_k$ **do**
          $E_{i_j,k_l} \leftarrow getEnergy(i_j, k_l)$ {get the non-bonded interaction energy between the the two rotamers}
        end for
      end for
    end for
  end for
end for
{A* search}
$generated\_confs \leftarrow \emptyset$
$priority\_queue \leftarrow \emptyset$
$root \leftarrow Node(NULL, 0)$ {the root node is a dummy node}
for each $conf \in start\_confs$ **do**
  $x \leftarrow Node(rigid\_part, root)$ {1. level is the rigid part of each input conformation}
  $priority\_queue.push(f(x), x)$ {add x to the priority queue}
end for
while $(priority\_queue \neq \emptyset)$ AND $(|generated\_confs| < N)$ **do**
  $x \leftarrow priority\_queue.pop$ {get node x with lowest f(x)}
  if $isLeafNode(x)$ **then**
    $generated\_confs \leftarrow generated\_confs \cup x.conformation$ {add complete conformation of x to results}
  else
    {add a new node for each rotamers r of the next flexible residue $i+1$}
    for each $r \in rotamers_{i+1}$ **do**
      $y \leftarrow Node(x.conformation \cup (i+1)_r, x)$ {add rotamer r to the partial conformation of x}
      $priority\_queue.push(f(y), y)$ {add y to the priority queue}
    end for
  end if
end while
return $generated\_confs$

In the summations, i runs over all flexible residues with already assigned rotamers r, while k and n run over the remaining ones, and l and m denote different rotamers of a side-chain. In each step, the node x with lowest $f(x)$ (representing the partial conformation that seems most promising) is taken from the priority queue. If x is a leaf node, the corresponding conformation is written to an output file. Otherwise, x is expanded, i.e. a new node y is added for each possible rotamer of the succeeding flexible residue $i+1$ and the priorities of these new partial conformations are determined. The algorithm terminates as soon as the predefined number of output conformations is reached.

5.3 Results

PocketScanner and PocketBuilder were tested using different parameters controlling the volume of the induced pockets. PocketScanner was run twice, using a GPS radius of either 2 or 3 Å. PocketBuilder was tested with three different weighting schemes per PocketScanner setup, resulting in a total of six runs.

5.3.1 Properties of the Pockets Induced by PocketScanner

PocketScanner was used to scan the apo protein structures for positions of inducible pockets. For each system, the grid center was placed at the ligand center of mass, the dimension was 11 x 11 x 5, and the edge length 2 Å. That way, the grid covered the entire protein-protein interaction interface. Running PocketScanner took about 1 hour on a single CPU of an Intel Core 2 Duo processor which mainly resulted from the large number of energy minimizations. PocketScanner was run twice using the same settings for the grid but different radii for the GPS. The use of the larger GPS had a significant impact on the number of accepted pocket positions. Out of the 605 possible positions, 67 (66) were accepted for BCL-X_L, 25 (18) for IL-2, and 29 (20) for MDM2 when using a GPS radius of 2 Å (3 Å respectively). Note that these pocket positions may be located anywhere in the protein-protein interaction interface and are not limited to the inhibitor binding site. As an example, the grid and the accepted pocket positions of BCL-X_L are shown in Figure 5.3.

EPOS^{BP} was applied to the resulting PocketScanner conformations for detecting those pockets that were induced at the position of a GPS. An overview of the properties of these pockets is shown in Table 5.1. This analysis revealed that pockets were detected in more PocketScanner conformations when the larger GPS radius was used indicating that a radius of 2 Å may be not sufficient to induce the opening of accessible pockets. Furthermore, the polarity of the pockets resulting from larger GPSs was slightly reduced. One could expect that the mean pocket volume would significantly increase when using a larger GPS, but this is not the case. For MDM2, the mean pocket volume even decreased. However, using a larger GPS radius may also cause a cavity that is more flat and, thus, of reduced volume.

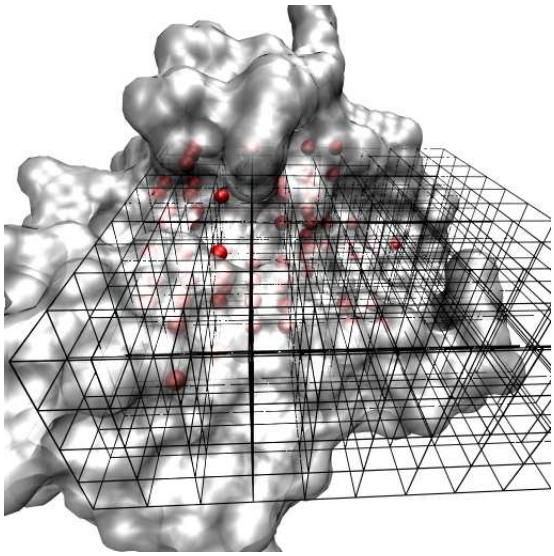


Figure 5.3: The interaction interface of apo BCL-X_L covered by the grid generated by PocketScanner. Accepted pocket positions are shown as red spheres.

system	GPS radius [Å]	detected pockets [%]	mean pocket volume [Å^3]	mean pocket polarity
BCL-X$_L$	2	43	381.3 ± 82.3	0.33 ± 0.03
	3	86	394.9 ± 109.7	0.29 ± 0.04
IL-2	2	52	311.8 ± 59.1	0.31 ± 0.04
	3	78	328.8 ± 58.9	0.29 ± 0.03
MDM2	2	31	376.8 ± 91.7	0.33 ± 0.02
	3	75	315.7 ± 98.6	0.31 ± 0.03

Table 5.1: Properties of the pockets induced by PocketScanner.

5.3.2 Properties of the Pockets Designed by PocketBuilder

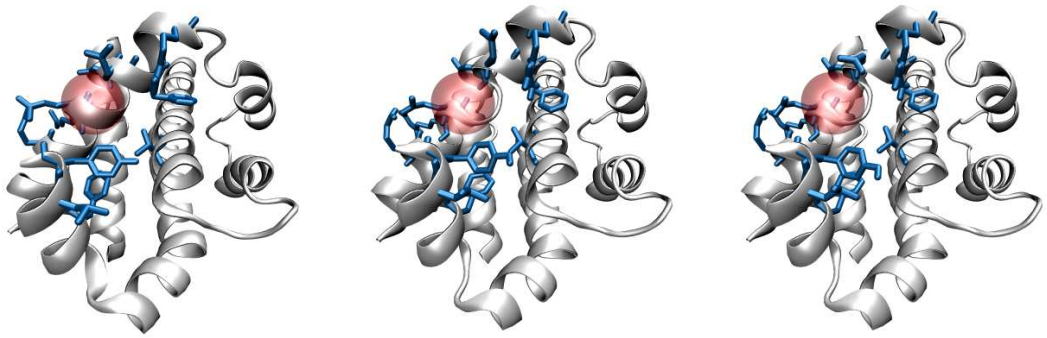
The conformations and the corresponding pocket positions generated by the two runs of PocketScanner were used as starting conformations for PocketBuilder. As the weighting factors for the internal protein energy and the protein-pocket interaction energy crucially influence the scores and, thus, the A* search, we calculated 500 conformations using three different weightings schemes for each GPS radius (resulting in six runs of PocketBuilder):

- internal protein energy and protein-pocket interaction energy weighted equally (0.5 and 0.5)
- a strong emphasis on the protein-pocket interaction energy (0.1 and 0.9)
- a dominance of the protein-pocket interaction energy (0.01 and 0.99)

In an initial test, we found that the initialization stage is the bottleneck for the run time of PocketBuilder with 6-10 minutes per starting conformation depending on the number of flexible residues (here, 8-18 flexible residues) and accepted rotamers. To speed-up the calculations, a greedy preselection of the starting conformations was added: For each conformation, the weighted sum of the internal protein energy and protein-pocket interaction energy was calculated and only the 20

system	GPS radius [Å]	w_{pocket}	total no. leaf nodes	efficiency	mean pocket volume [Å^3]	mean pocket polarity
BCL-X$_L$	2	0.50	$1.0 \cdot 10^{12}$	$8.3 \cdot 10^{6}$	715.3 ± 21.9	0.36 ± 0.01
	2	0.90	$1.9 \cdot 10^{12}$	$1.7 \cdot 10^{7}$	343.6 ± 31.7	0.27 ± 0.01
	2	0.99	$3.4 \cdot 10^{12}$	$1.6 \cdot 10^{9}$	337.4 ± 37.2	0.27 ± 0.01
	3	0.50	$1.7 \cdot 10^{11}$	$2.4 \cdot 10^{6}$	282.6 ± 34.2	0.30 ± 0.01
	3	0.90	$5.6 \cdot 10^{11}$	$7.1 \cdot 10^{6}$	276.1 ± 55.0	0.31 ± 0.01
	3	0.99	$4.5 \cdot 10^{14}$	$1.2 \cdot 10^{9}$	485.2 ± 92.7	0.37 ± 0.01
IL-2	2	0.50	$2.0 \cdot 10^{15}$	$1.0 \cdot 10^{11}$	291.7 ± 3.8	0.27 ± 0.01
	2	0.90	$2.7 \cdot 10^{16}$	$9.3 \cdot 10^{11}$	290.3 ± 4.8	0.27 ± 0.01
	2	0.99	$1.9 \cdot 10^{18}$	$4.1 \cdot 10^{13}$	359.6 ± 36.3	0.33 ± 0.01
	3	0.50	$1.2 \cdot 10^{14}$	$5.9 \cdot 10^{9}$	450.9 ± 80.2	0.31 ± 0.01
	3	0.90	$2.2 \cdot 10^{15}$	$6.9 \cdot 10^{10}$	507.4 ± 90.7	0.30 ± 0.01
	3	0.99	$4.6 \cdot 10^{16}$	$1.2 \cdot 10^{12}$	344.4 ± 23.3	0.33 ± 0.01
MDM2	2	0.50	$1.5 \cdot 10^{14}$	$1.4 \cdot 10^{10}$	314.0 ± 56.8	0.31 ± 0.02
	2	0.90	$1.4 \cdot 10^{15}$	$1.4 \cdot 10^{10}$	420.0 ± 50.2	0.33 ± 0.02
	2	0.99	$2.1 \cdot 10^{16}$	$2.4 \cdot 10^{7}$	277.9 ± 19.6	0.32 ± 0.01
	3	0.50	$2.6 \cdot 10^{12}$	$7.0 \cdot 10^{8}$	233.8 ± 26.3	0.32 ± 0.01
	3	0.90	$8.8 \cdot 10^{13}$	$7.6 \cdot 10^{9}$	235.3 ± 27.1	0.32 ± 0.01
	3	0.99	$2.0 \cdot 10^{15}$	$1.4 \cdot 10^{10}$	339.1 ± 89.1	0.31 ± 0.02

Table 5.2: Influence of the GPS radius and the weighting on the performance of PocketBuilder and the properties of the induced pockets.



(a) apo structure (b) after PocketScanner (c) after PocketBuilder

Figure 5.4: Conformational changes of the backbone (shown in cartoon representation) and the residues defined as flexible in PocketBuilder (shown in licorized representation) in a subpocket of BCL-X_L. The GPS (radius = 3 Å) is shown as transparent red sphere. The initial overlap of the protein atoms with the GPS (a) was reduced during the energy minimization in PocketScanner (b). Running PocketBuilder with w_{pocket} = 0.99 optimized the interaction with the GPS by changing the rotamers of some flexible side-chains (c).

starting conformations with lowest score were retained. On the one hand this preselection may delete conformations that would later on score better with altered side-chain rotamers, but on the other hand running the algorithm with too many starting conformations is nearly infeasible. The run time of the A* search took between 40 minutes and 4 hours depending on the number of possible nodes in the search tree and on how similar the scores of these nodes are. Here, we use the ratio between the number of possible nodes and the number of generated nodes as a measure for the efficiency of the algorithm. Analogously to the pockets induced by PocketScanner, the properties of the pockets designed by PocketBuilder were calculated. The results of this analysis, the number of different conformations (leaf nodes) that could be generated using this setup, as well as the measure for the efficiency of this PocketBuilder run are listed in Table 5.2. Although the total number of leaf nodes increased with augmenting w_{pocket}, the algorithm generally found the 500 best conformations more efficiently, suggesting that the interaction energy between the protein and the GPS was more diverse than the internal protein energy. No trend was apparent for the influence of the weighting and the GPS radius on the mean pocket volume and polarity. These mean volumes even seem to suggest that PocketBuilder reduced the volume of most pockets to snugly fit around the GPSs. An example of how PocketScanner and PocketBuilder changed the apo structure of BCL-X_L is shown in Figure 5.4.

5.3.3 Docking into Pockets Designed by PocketBuilder

The aim of this approach is the efficient design of ligand binding pockets on the protein surface. As before, the appropriateness of this protocol for drug design is validated by docking the known inhibitors into the designed binding pockets. The main questions are:

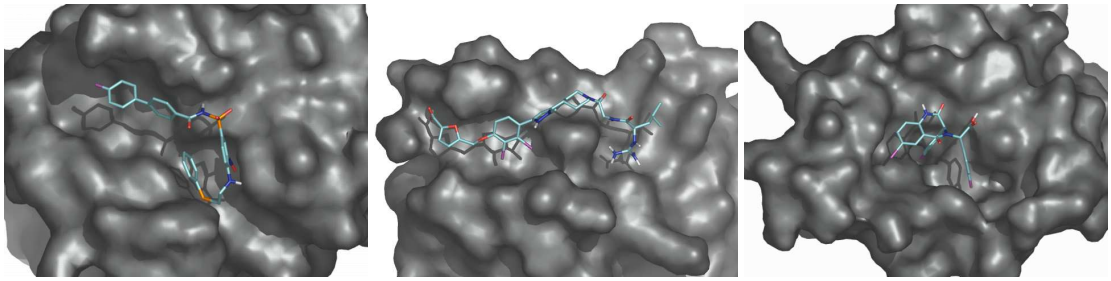
- Can docking into the designed pockets reproduce the native ligand binding mode?
- Which weighting and GPS radius requires the lowest number of generated conformations?

Table 5.3 lists the best scored docking results with RMSD $\leq$ 2 Å (or the docking result with lowest RMSD) for each PocketBuilder run. The corresponding docking complexes are shown in Figure 5.5. With each setup, PocketBuilder successfully induced the opening of native-like binding pockets on the surface of the BCL-X_L and the IL-2 protein as the RMSDs indicate. For BCL-X_L, the docking scores were even in the same order of magnitude than the re-docking scores (see Table 3.5) and the scores obtained when docking into snapshots taken from the MD simulation

system	GPS radius [Å]	w_{pocket}	RMSD [Å]	score [kcal/mol]	score rank [%]	solution no.
BCL-X_L - N3B	2	0.50	1.9	-10.0	42.0	213
	2	0.90	2.0	-10.1	34.4	169
	2	0.99	2.0	-10.2	33.6	241
	3	0.50	1.7	-10.2	55.4	82
	3	0.90	1.5	-10.4	58.2	376
	3	0.99	2.0	-11.3	6.2	29
IL-2 - FRH	2	0.50	1.8	-6.5	6.4	75
	2	0.90	1.8	-7.3	1.7	226
	2	0.99	2.0	-4.3	54.4	428
	3	0.50	2.0	-5.6	44.1	167
	3	0.90	2.0	-6.6	29.6	285
	3	0.99	2.0	-4.4	60.6	430
MDM2 - DIZ	2	0.50	2.6	-7.9	83.1	193
	2	0.90	2.6	-7.8	90.5	225
	2	0.99	2.9	-9.1	4.9	113
	3	0.50	3.2	-9.7	5.9	436
	3	0.90	3.1	-8.8	27.4	345
	3	0.99	2.2	-9.1	88.3	41

Table 5.3: Influence of the GPS radius and the weighting on the docking results. Shown are the best scored docking results with RMSD $\leq$ 2 Å or the docking result with lowest RMSD.

in water (see Table 4.2). Interestingly, when using the larger GPS and setting w_{pocket} to 0.99, the docking score is even lower than in the re-docking experiment. For IL-2, the docking scores were less satisfying. However, one should keep in mind that this binding site consists of two subpockets that lie about 15 Å apart and with this approach one can only induce the opening of one of these subpockets. Here, using more than one GPS would most probably improve the docking score. For MDM2, PocketBuilder was not able to generate binding pockets into which the ligand could bind in its native binding mode. But comparing these docking results to those of the apo-docking (listed in Table 3.5) indicates that an opening of the native binding was at least partly induced. In this example, the truncated structure used in Chapter 4 may be more appropriate for inducing pocket openings by energy minimizations. However, the large relative rank of most docking results indicates that all setups do not only lead to openings of pockets similar to those seen in the holo structures, but also to alternative pocket conformations. In fact, most setups seem to prefer these alternative pockets because the protein conformation in which the native binding mode is best reproduced was often generated quite late during the A* search.



(a) BCL-X_L (b) IL-2 (c) MDM2

Figure 5.5: The best scored docking poses obtained when docking into PocketBuilder conformations corresponding to those listed in Table 5.3. In (c), residues 1 to 16 were removed for better visibility.

5.4 Discussion

The docking experiments indicate that many conformations with cavities at locations different from that of the native binding pocket were computed. The putative pocket positions are suggested by PocketScanner and each generated protein conformation contains a pocket at a different position. However, not all of these pocket positions are retained in the 500 final conformations as the selection depends on the scoring function incorporated in PocketBuilder. Note that although the nodes of the first level of the A* search tree represent different PocketScanner conformations and pocket positions, it is not guaranteed that these nodes are all expanded or lie on a path to a leaf node (i.e. they represent the basis of a final conformation). So it may happen that PocketScanner induces an initial opening of the native binding pocket but this conformation is not further considered by PocketBuilder because at least one other conformation achieved a better score. Similarly, although PocketBuilder may compute conformations that are based on PocketScanner conformations with the pocket at the native binding site, they are not necessarily more favorable than others. Therefore, one should rather consider a number of solutions instead of just the optimal one (only the first conformation generated during the A* search).

5.4.1 Is the representation of a pocket by a single GPS reasonable?

The representation of a binding pocket by a single GPS is just a rough approximation and bears several disadvantages. As the protein is relaxed around a spherical obstacle, the resulting pocket tends to be of an artificial globular shape and the protein surface may be too smooth as all protruding side-chains elude unfavorable interactions with the GPS and huddle against the protein surface. Moreover, as the example of IL-2 demonstrated, binding pockets may consist of several subpockets. In such a case, inducing the opening of just one of them is not enough. However, when using PocketScanner as described above, scanning the interface with two or even more GPS at the same time would result in a combinatorial explosion. Here, a more accurate description of the binding site would be required. For example, when the exact location of the binding (sub-) pockets are known, one could place several GPS manually. We tested this alternative setup of PocketScanner for BCL-X_L, IL-2, and MDM2 and placed several GPS manually (based on the structure of the superimposed inhibitors) in the apo structures. Here, only one PocketScanner conformation was generated and used as input structure for PocketBuilder. When docking the inhibitors into the pockets of the 100 conformations refined by PocketBuilder, the native ligand binding pose could be reproduced for all three systems with docking scores of -9.6 kcal/mol for BCL-X_L and MDM2, and -6.5 kcal/mol for IL-2. Thus, even when incorporating detailed information about the native binding pose in the design process of the binding pocket, the docking results do not improve significantly. Although the unbiased approach that depends only on an approximate definition of the binding site location is not yet mature, it shows that scanning a protein-protein interaction interface computationally for inducible pockets is feasible and that the results are of comparable quality than those obtained when detailed *a priori* knowledge about the binding site was considered.

5.4.2 Critical Assessment of the Approach

The approach presented in this chapter shows that, in principle, it is a promising idea to induce pockets algorithmically by representing them by their negative image that interacts with the protein. But as discussed above, a single GPS is in many cases not sufficient to induce native-like binding pockets. This representation requires keeping the position of the GPS fixed during the energy minimization because otherwise it could be easily “pushed away” by the protein. But this procedure of “drilling holes” in the protein surface may be too hard and artificial. When looking

back, it appears more reasonable to harmonically restrain the position of the GPS and this may be tested in future work. By fixing the GPS, the protein may get distorted and, thus, be in a high-energy conformation that is useless for PocketBuilder. In many cases it may be sufficient to move the GPS just by less than 1 Å to resolve this strain. Alternatively, clashes between the GPS and the protein atoms would be less severely punished when using a soft-core potential for calculating the van der Waals interaction energy. A further alternative is sampling the pocket positions using a finer resolution (i.e. smaller edge length of the grid) but this would be computationally very demanding. Likewise, the quality of the PocketScanner conformations depends on the orientation and resolution of the grid as it defines the positions of the GPSs. As a result, PocketBuilder often considers only a few PocketScanner conformations as the others have an unfavorable internal protein energy. But when generating 500 final conformations with only up to 18 flexible residues, it is not surprising that the resulting conformational ensemble is very redundant with solutions differing often by only one side-chain.
Besides the orientation and resolution of the grid, the choice of the rotamer library and the force field may also have an impact on the performance of this approach. While Dunbrack's backbone independent rotamer library from 2002 was chosen as it is the newest one distributed with the BALL library, we decided to use the CHARMM EEF1 force field as it incorporates an implicit solvent term. Based on our findings discussed in Chapter 4, accounting for solvent effects seemed to be important when focussing on solvent exposed pockets. But this force field has the drawback that all amino acid side-chains have a neutral charge, even those which are charged at physiological conditions. So it is unclear whether this force field is really suited to determine the best combinations of rotamers in terms of internal protein energy and pocket volume.

5.5 Summary and Conclusion

The pocket detection protocols that were introduced before scanned the entire protein surface for transient pockets and so suggested putative binding sites. Here, we presented a new approach that assumes that the location of the binding site is approximately known. Thus, this surface region can be exhaustively scanned and tailored ligand binding pockets can be induced algorithmically by considering protein backbone and side-chain flexibility. While PocketScanner relaxes protein conformations in the presence of generic pocket spheres, PocketBuilder induces pockets of desired properties by searching for the best combinations of side-chain rotamers using the A* algorithm. We suggest to use the two programs together, but in principle they could be used individually. The drawbacks of this method are that the designed pocket are of an artificial globular shape and that the binding sites are too smooth. Furthermore, by using only one GPS at a time the applicability of this approach is limited to binding sites consisting of only one pocket and not of several subpockets. However, for two out of the three systems, the PocketBuilder algorithm was able to induce pockets of suitable volumes and shapes so that the small-molecule inhibitors could bind in a native-like orientation. For the third system, the docking results improved significantly compared to docking into the apo structure. Thus, this chapter presented a pioneering work for approaches representing efficient alternatives to our MD-based pocket detection protocol introduced in Chapters 3 and 4 for cases when the location of the binding site is approximately known.

Chapter 6

Designing Binding Pockets on Protein Surfaces using an incremental Inflation Procedure

As binding pockets for SMPPIIs often consist of several subpockets, we suspected that taking this fact into account may significantly improve the performance of PocketScanner and PocketBuilder. We will now introduce an approach which implements the improvements suggested in the previous chapter. The method is still under development and the testing of various parameters is ongoing. Hence, the results should be considered as preliminary, but promising.

6.1 Introduction

In the previous chapter, a pioneering method was presented for the algorithmic design of ligand binding pockets on protein surfaces. We could show that scanning protein-protein interaction interfaces for inducible pockets is computationally feasible, but found that the scheme "one pocket - one GPS" is often inappropriate for designing native-like binding pockets. Notably, the examples of IL-2 and BCL-X_L illustrate that SMPPIIs often bind to several subpockets at the same time. Thus, docking these ligands into generated conformations with just a single cavity won't give a realistic estimate of the free binding energy. Furthermore, the native subpockets accommodating parts of the inhibitors are rather of irregular shape with a rough surface than globular with a smooth surface like the pockets whose openings are induced by relaxing the protein around a single GPS. In addition, the positions of the induced pockets depend on the orientation of the grid generated by PocketScanner. Although the protein is minimized, the presence of the GPS at this position may induce so much strain on the protein that this conformation won't be accessible *in vitro*. In such a case, only a minor adjustment of the pocket position may result in an energetically more favorable protein conformation. Moreover, although using the CHARMM EEF1 force field [158] seemed promising due to its implicit solvent term, we became skeptical about its appropriateness for the design of pockets on protein surfaces during the course of this work because all side-chains are modeled as non-charged. Thus, we later decided to switch to using the Amber 96 force field [159] to ensure that electrostatic contributions are adequately accounted for when optimizing side-chain orientations, especially as hot spots in protein-protein interaction interfaces are often represented by charged residues. We incorporated all these considerations into a new approach, termed *PocketInflator*, that can be considered as a combination of EPOS^{BP}, PocketScanner, and PocketBuilder.

In PocketInflator, the protein surface is scanned for initial pockets that are located next to user-defined protein residues using a modified version of EPOS^{BP}. Instead of placing the PASS probes at positions where they do not clash with protein atoms as before, we now introduce a *clash factor* that scales down the sum of the radii of the probe and the protein atoms during the filtering

step of the PASS algorithm and so allows for overlaps between probe patches and the protein. As in the original implementation of EPOS^{BP}, coherent PASS probes constitute a patch, and a patch represents the negative image of a pocket. Here, each probe is substituted by a GPS of the corresponding radius, and, thus, a (sub-) pocket is represented by a set of GPSs. This setup addresses the main problems of the previous pocket design approach: The position as well as the shape of the pockets are dictated by the protein structure and several (sub-) pockets may be induced at the same time. Analogously to PocketScanner, the protein is then relaxed in the presence of these patches of GPSs. After each energy minimization, the detection of the pockets is repeated with an increasing clash factor, resulting in an incremental reduction of the overlap between the PASS probes and the protein atoms. By doing so, similar to PocketBuilder, a set of partial solutions is generated that are scored by their potential energy, the ratio of the user-defined protein residues that line the current version of the pocket, and the deviation from the user-defined goal volume.

6.2 Methods and Materials

In this new approach, the algorithms of EPOS^{BP}, PocketScanner, and PockerBuilder were fused into one program, called PocketInflator. Like its precursors, it is implemented in C++ and uses the BALL library. All energies were calculated by the Amber 96 force field instead of CHARMM EEF1 and four different sizes of GPS (radii of 0.7, 1.8, 2.1, and 5.4 Å, all with a well depth of 0.2 kcal/mol) were considered.
The model systems and the used structures were the same as in the previous chapters. As already discussed in Chapter 4, the apo structure of MDM2 is inappropriate for energy minimizations in vacuo. Therefore, we used the same truncated structure as in Chapter 4.

6.2.1 The PocketInflator Algorithm

The flowchart of this approach aiming at inducing pockets of a predefined volume at defined positions is illustrated in Figure 6.1. In contrast to the PocketScanner/PocketBuilder method, an arbitrary number of (sub-) pockets can be induced at the same time. For this purpose the approximate location of each individual (sub-) pocket has to be defined by a set of residues that should line it. The definition of the (sub-) pocket's goal volume is optional. In addition, a set of starting structures and the number of solutions to be generated have to be defined. In order to generate energetically favorable conformations that possess accessible pockets at defined sites, intermediate solutions are calculated and stored in a priority queue. The score of such an intermediate solution is composed of three to four terms with values between 0 and 1:

- $score_{energy}$: the ratio of the energy of this conformation (E_{conf}) to the lowest energy ($E_{min} < 0$) of all starting structures after 1000 steps of L-BFGS energy minimization (if $\frac{E_{conf}}{E_{min}} < 0$, then $score_{energy} = 0$; if $\frac{E_{conf}}{E_{min}} > 1$, then $score_{energy} = 1$)

- $score_{hits}$: the ratio of the number of predefined residues that are found within 8 Å of the ASP of the induced pocket versus the total number of predefined residues

- $score_{cf}$ is set to the clash factor cf that allows for overlaps between the PASS probes of the patch and the protein atoms in the BALLPass algorithm by defining clashes as

$$distance(probe, atom) < cf \cdot (radius_{probe} + radius_{atom}) \tag{6.1}$$

  The maximum value of cf is 0.95 as this is the default value in EPOS^{BP} that best reproduced the results of the original PASS program.

- $score_{volume}$ (optional): the deviation of the pocket volume vol from the goal volume vol_{goal} calculated by

$$score_{volume} = e^{-\left(\frac{vol-vol_{goal}}{0.3 \cdot vol_{goal}}\right)^2} \tag{6.2}$$

  Note that this value is incorporated in the total score only if the corresponding cf is already at its maximum value of 0.95.

The total score is then calculated by

$$score = score_{energy} \cdot score_{cf} \cdot \frac{1}{n} \sum_{i=1}^{n} (score_{hits}(i)[\cdot score_{volume}(i)]) \tag{6.3}$$

where i runs over the individual subpockets. The pseudocode of the PocketInflator program is listed in Algorithm 5.
The program starts with energetically minimizing all input structures using 500 steps of L-BFGS (if afterwards, $E_{conf} > 0$, the minimization is repeated) and detecting the initial pocket patches in this conformation. The subsequent conformation is further minimized in order to determine E_{min}. All conformations are then scored and added to the priority queue if $score > 0$. (Note that if no pocket was detectable in the starting structure, the program terminates without a solution.) After this initialization stage, the algorithm iteratively extracts the best scored intermediate solution from the priority queue and further inflates the existing pocket until a predefined number (default: 50) of final solutions (with $cf = 0.95$ and $score \geq 0.3$) is generated. The minimum score was used to ensure that the resulting conformations are of low-energy, contain (sub-) pockets at the predefined locations, and, if defined, are of the desired volume.

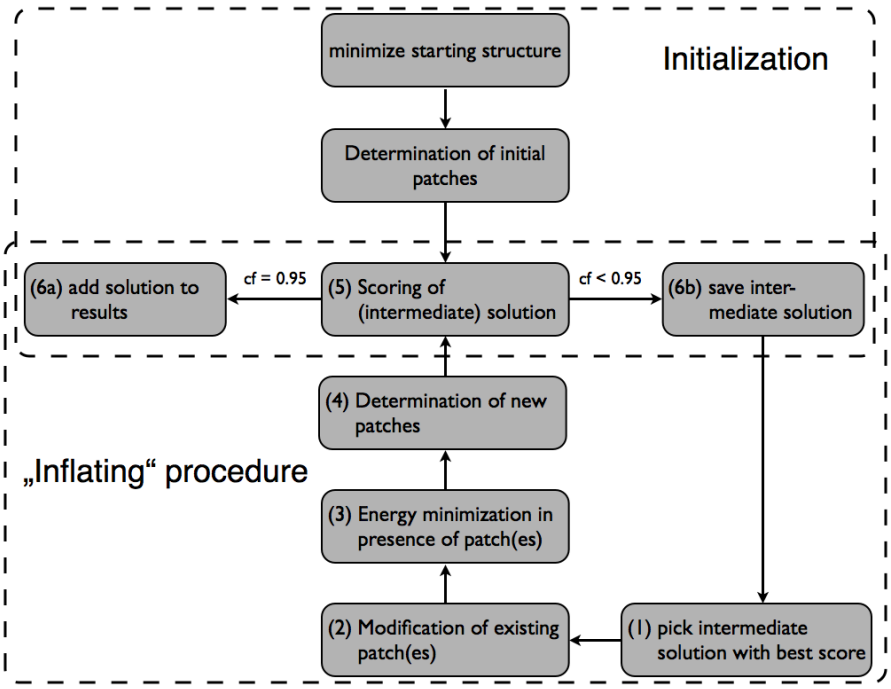
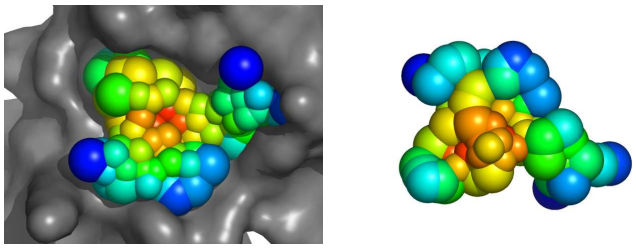


Figure 6.1: Flowchart of the PocketInflator approach. PocketInflator consists of an initialization and an "inflating" procedure. For inflating pockets, the algorithm (1) selects the best scored intermediate solution, (2) enlarges the existing patch(es), converts the probes to GPSs and (3) minimizes the protein in their presence, (4) determines new patch(es) that overlap less with the protein atoms, and (5) scores and (6) stores this (intermediate) solution. This procedure comprising steps 1 - 6 is repeated until a predefined number of solutions is generated or until no intermediate solutions are left.



(a) (b)

Figure 6.2: The distribution of the probe weights within a patch shown at the example of the native binding pocket of MDM2. All probes are represented by spheres and colored by their weight (normalized by the weight of the ASP) ranging from red (high probe weight) to blue (low probe weight). The top-view of the patch (a) including the protein surface (shown in grey) indicates that probes in the center of the patch have a higher weight than those located at the border. Subfigure (b) shows the same patch rotated by 180° displaying the moderate weights of the probes at the bottom of the patch.

The inflating procedure consists of three main steps that are described in the following: modification of the existing patch (step 2 in Figure 6.1), energy minimization to adopt the protein conformation to it (step 3), and determination of (a) new patch(es) (step 4).

Modification of the existing patch The PASS algorithm calculates a real-valued probe weight pw for each probe (see page 43) that reflects its burial count and the number of surrounding probes. As Figure 6.2 demonstrates, probes at the border of the patch have lower weights, probes located at the bottom have moderate weights, and those found in the center of the patch (like the ASP) have the highest weights. In order to inflate the pocket, it is reasonable to enlarge those probes having a high weight. Thus, for a given threshold th, the radius of each probe i with is tripled, if

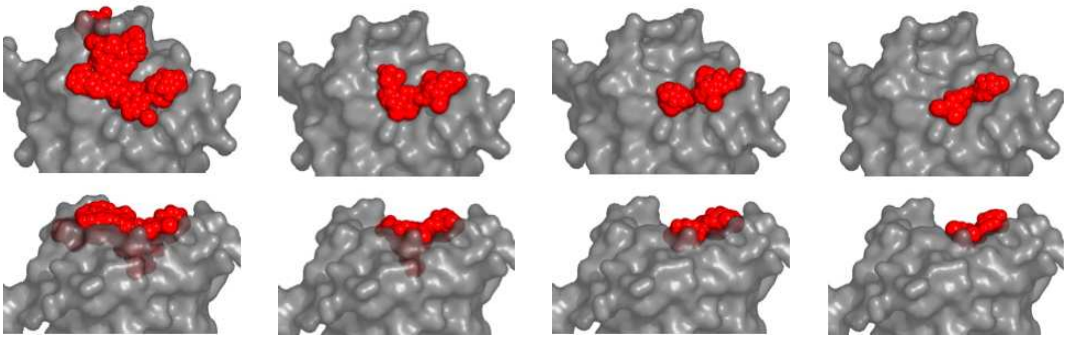
$$pw_i \geq th \cdot pw_{ASP} \tag{6.4}$$

All three steps described here are repeated for different thresholds. In this setup, we use four different thresholds (0, 0.3, 0.6, 0.9). Initial tests using a smaller step size showed that two similar threshold often resulted in the same modified patch and this unnecessarily increased the run time of PocketInflator. The threefold enlargement of the GPS radius may sound quite drastic but was necessary in order to achieve a sufficiently large (further) opening of the pockets. The reason for this is that we used a soft-core potential that will be discussed in the next paragraph. Doubling the radius had only minor effects on the protein conformations as initial tests indicated.

Energy minimization in presence of the patch All probes are translated to GPSs of the same radius. The protein conformation and the patch containing the enlarged probes are then subjected to 500 steps of L-BFGS energy minimization. The implementation of the force field was modified such that a GPS can only interact with protein atoms, i.e. different GPS spheres do not interact with each other. In contrast to PocketScanner, the van der Waals interaction energy between a protein atom i and a GPS j of radius r separated by a distance of d_{ij} is calculated by a soft-core potential

$$U_{vdW} = 4\epsilon_{ij} \cdot \left(\frac{\sigma_{ij}^{12}}{(d_{ij}^2 + r)^6} - \frac{\sigma_{ij}^{6}}{(d_{ij}^2 + r)^3} \right) \tag{6.5}$$

according to [160] and the positions of the GPSs are not fixed. This modification was necessary in order to avoid strained protein conformations and articifially smooth binding pockets. The conformations are stored every 100 steps, their internal protein energy is calculated, and new patches are determined.



(a) clash factor: 0.75, volume: 1,257 Å^3 (b) clash factor: 0.80, volume: 505 Å^3 (c) clash factor: 0.85, volume: 318 Å^3 (d) clash factor: 0.90, volume: 194 Å^3

Figure 6.3: The influence of the clash factor on the placement of the probes in the PASS algorithm. The protein is shown in grey surface representation and the probes are displayed as red spheres. In the lower figures the protein is rotated by 90° and the surface is made transparent to illustrate the depth of the patch. Note that the smaller the clash factor, the more overlap is allowed between the probes and the protein atoms and, thus, the more probes are kept and the larger the patch.

Determination of new patches The selection of a new patch is the trickiest part, especially when inducing multiple subpockets at once. The clash factor must be larger than the previous one and

$$score' = score_{cf} \cdot \frac{1}{n} \sum_{i=1}^{n} score_{hits}(i) \tag{6.6}$$

should be maximal. Thus, a set of different patches is calculated by running EPOS^{BP} with iteratively increasing clash factors. As illustrated in Figure 6.3, this clash factor affects the placement of the probes, i.e. more probes are kept when a smaller value is used that allows the probes to penetrate the protein atoms more deeply. After the placement of the probes, the ASPs are determined and the probes are assigned to them in order to form contiguous patches as described on page 57. Note that each subpocket should be represented by one patch. If they are vicinal, it may happen that a large patch is detected that covers multiple subpockets. In this case their individual properties cannot be controlled anymore and thus, the algorithm tries to avoid this situation. The number of ASPs and so of the patches can be controlled by modifying the minimum probe weight of an ASP, pw_{min}, and the minimal distance to an already existing one (see page 43). Therefore, the determination of the ASPs was modified. We reduced the minimal distance between two ASPs from 8 Å to 5 Å and for the purpose of obtaining varying numbers of patches for the same set of probes, pw_{min} was increased until no ASPs could be detected anymore. (Note that using a small pw_{min}, the set of probes is divided into multiple small patches, while the usage of a large pw_{min} results in a single large patch.) From all sets of patches resulting from the different runs of EPOS^{BP} with increasing clash factor and pw_{min}, the run for which $score'$ is maximal is determined. The definition of $score'$ ensures that all patches used to inflate the subpockets were derived from the same EPOS^{BP} run. The corresponding patch(es) are included in the protein conformation, the total score is calculated, and this intermediate solution is added to the priority queue.

6.2.2 Derivation of the Input Parameters

The input parameters that define where the (sub-) pockets should be induced and their volumes were derived from the holo structures. As the pockets induced by PocketInflator are based on EPOS^{BP}, we applied this program to the holo structures and identified those patches that overlapped with the bound inhibitors. These patches were then reduced by only keeping those probes

Algorithm 5 The PocketInflator algorithm

Input: $start\_confs \leftarrow$ input structures
Input: $subpockets\_regions \leftarrow$ set of residue IDs defining the approximate site where the pockets should be located
Input: $subpockets\_volumes \leftarrow$ the goal volume per subpocket that should be generated
{initialization}
$E_{min} = \infty$
$tmp = \emptyset$
for each $conf \in start\_confs$ **do**
  $conf \leftarrow runEM(conf)$ {minimize input structure by 500 steps of L-BFGS}
  if $E_{conf} \geq 0 kcal/mol$ **then**
    $conf \leftarrow runEM(conf)$ {repeat the energy minimization once}
  end if
  if $E_{conf} < 0\ kcal/mol$ **then**
    $conf' \leftarrow runEM(conf)$ {minimize it again to define E_{min}}
    if $E_{conf'} < E_{min}$ **then**
      $E_{min} \leftarrow E_{conf'}$
    end if
    $patches, score' \leftarrow getBestPatches(conf, subpockets\_regions, subpockets\_volumes, 0.75)$ {run EPOS^{BP} with different pw_{min} and cf to get a patch per subpocket with a maximal $score'$}
    $tmp\_sol \leftarrow conf, patches, E_{conf}, score'$
    $tmp \leftarrow tmp \cup tmp\_sol$ {add this intermediate solution tmp_sol to tmp}
  end if
end for
$priority\_queue = \emptyset$
$solutions = \emptyset$
for each $entry \in tmp$ **do**
  $tmp\_sol \leftarrow entry.tmp\_sol$
  $tmp\_sol.score_{energy} \leftarrow tmp\_sol.score_{energy}/E_{min}$ {normalize the energy by E_{min}}
  $score \leftarrow tmp\_sol.score' \cdot tmp\_sol.score_{energy}$ {update the score of this intermediate solution}
  if $(tmp\_sol.patches.cf = 0.95)$ AND $(score \geq 0.3)$ **then**
    $solutions.push(score, tmp\_sol)$ {add this result to solutions}
  else if $tmp\_sol.patches.cf < 0.95$ **then**
    $priority\_queue.push(score, tmp\_sol)$ {add this intermediate solution to the priority queue}
  end if
end for
{inflate existing patches}
while $priority\_queue \neq \emptyset$ AND $|solutions| < N$ **do**
  $tmp\_sol \leftarrow priority\_queue.pop$ {get intermediate solution with highest score} {modify the existing patches by iterating over different thresholds th}
  for $th = 0; th < 1; th = th + 0.3$ **do**
    $mod\_patches \leftarrow enlarge(tmp\_sol.patches, th$ {enlarge each probe i with $pw_i \geq th \cdot pw_{ASP}$}
    $conf\_with\_patches \leftarrow tmp\_sol.conf \cup mod\_patches$ {modify the existing patches by iterating over different thresholds th}
    for $step = 0; step < 5; step = step + 1$ **do**
      $conf\_with\_patches \leftarrow runEM(conf\_with\_patches)$ {minimize input structure using 100 steps of L-BFGS}
      $conf \leftarrow conf\_with\_patches \setminus conf\_with\_patches.patches$
      $patches, score' \leftarrow getBestPatches(conf, subpockets\_regions, subpockets\_volumes, tmp\_sol.patches.cf + 0.05)$ {run EPOS^{BP} with a cf greater than in the previous run}
      $score_{energy} \leftarrow \frac{E_{conf}}{E_{min}}$
      $score \leftarrow score_{energy} \cdot score'$
      $tmp\_sol' \leftarrow conf, patches, score_{energy}, score'$ {create a new intermediate solution tmp_sol'}
      if $(tmp\_sol'.patches.cf = 0.95)$ AND $(score \geq 0.3)$ **then**
        $solutions.push(score, tmp\_sol')$ {add this result to solutions}
      else if $tmp\_sol'.patches.cf < 0.95$ **then**
        $priority\_queue.push(score, tmp\_sol')$ {add this intermediate solution to the priority queue}
      end if
    end for
  end for
end while
return $solutions$

that overlapped with inhibitor atoms. Thereby, the native (sub-) pockets that may be larger in volume than the ligand itself are restricted to the relevant regions and so putative noise is excluded. For each refined patch, its volume and the residues lining the (sub-) pockets were extracted. (Note that in contrast to Chapter 3 and 4 where the overlap volumes were calculated using all patches of

the given conformation, we refined here each patch individually to obtain information about the different subpockets involved in inhibitor binding.)

6.2.3 Docking into Designed Pockets

Docking experiments were performed as described in Chapter 3. As the pockets consisted in most cases of two nearby subpockets, the center of the first one was used to define the grid center as the used grid dimensions were large enough for completely covering the second subpocket as well.

6.3 Results

While for MDM2 the inhibitor binds into a single pocket, the analysis of the patches in the holo structures overlapping with inhibitor atoms correctly predicted that the inhibitors of BCL-X_L and IL-2 bind into two vicinal subpockets. The volumes of these pockets and the residues lining them are compiled in Table 6.1. These data were employed for defining the positions and the volumes of the pockets that should be induced. For all systems, the apo structure was used as starting conformation. An example on how pockets are inflated on the native binding site of apo MDM2 is shown in Figure 6.4.
In order to test the impact of the goal volume, we repeated all runs of PocketInflator without these values. (We will refer to these two different runs as *Vol-run* and *noVol-run*.) The resulting conformations are only scored by the relative deviation of the internal protein energy from its minimum value, the clash factor, and the percentage of predefined residues that effectively neighbor the induced (sub-) pockets. The run time of the Vol-runs ranged between 42 minutes and 37 hours, while the noVol-runs took between 16 and 160 minutes on one 2.8 GHz Xeon CPU. Although we tried to generate 50 protein conformations per run, PocketInflator terminated for both runs of IL-2 and MDM2 before this number was reached, indicating that an insufficient number of intermediate solutions with a score greater than 0.3 could be found. For MDM2, both PocketInflator runs even terminated without any solution. In this case, we released the strict condition for the minimum score of an intermediate solutions and set this threshold to 0. But even with this change the program was only capable of generating 9 solutions in both runs. For IL-2, we kept the threshold of 0.3. But here, only 5 solutions were returned in the Vol-run and 14 in the noVol-run.

system	**subpocket 1**		**subpocket 2**	
	vol. [Å^3]	residues	vol. [Å^3]	residues
BCL-X_L	445	Ala^{89}, Leu^{90}, Ala^{93}, Glu^{96}, Phe^{97}, Arg^{100}, Tyr^{101}, Asn^{136}, Trp^{137}, Gly^{138}, Arg^{139}, Val^{141}, Ala^{142}, Phe^{191}, Tyr^{195}	265	Phe^{97}, Tyr^{101}, Arg^{103}, Ala^{104}, Phe^{105}, Leu^{108}, Leu^{130}, Gly^{138}, Arg^{139}, Ala^{142}
IL-2	225	Lys^{43}, Phe^{44}, Tyr^{45}, Glu^{62}, Pro^{65}, Thr^{111}	450	Lys^{35}, Arg^{38}, Met^{39}, Thr^{41}, Phe^{42}, Val^{69}, Leu^{72}, Ala^{73}
MDM2	520	Ser^{17}, Leu^{54}, Phe^{55}, Leu^{57}, Gly^{58}, Gln^{59}, Ile^{61}, Met^{62}, Tyr^{67}, Gln^{72}, His^{73}, Val^{75}, Val^{93}, His^{96}, Ile^{99}	-	-

Table 6.1: The input parameters used to test PocketInflator as derived from the EPOS^{BP} analysis of the holo structures.

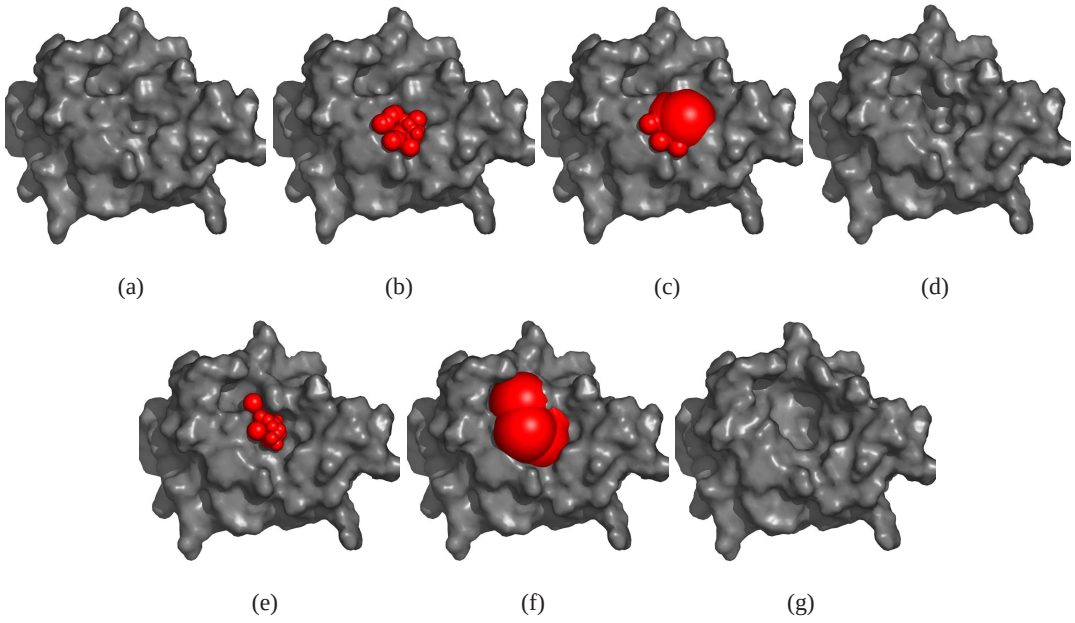
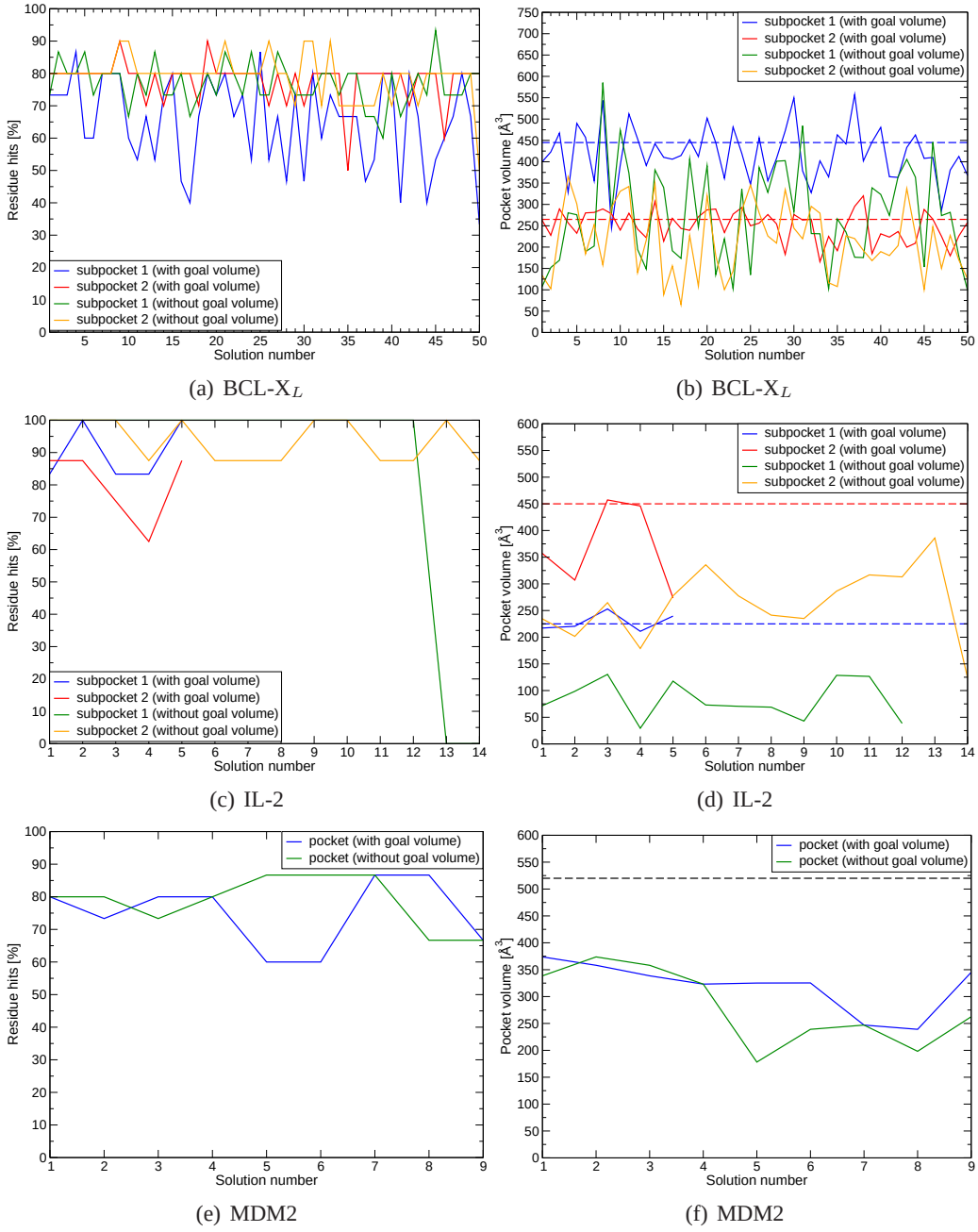


Figure 6.4: The pocket inflating procedure shown at the example of MDM2. The probes representing the pocket are shown as red spheres and the protein surface is colored grey. After minimizing the apo structure (a), an initial pocket is detected (b) and several probes are enlarged (c). The protein structure is then energetically minimized in the presence of these probes resulting in an intermediate solution (d) for which the calculation of new pockets (e), the probe enlargement (f), and the subsequent energy minimization (g) is repeated.

6.3.1 Properties of the Pockets Designed by PocketInflator

The magnitude to which the predefined properties were met in the resulting conformations are depicted in Figure 6.5. In most cases, the program was able to induce the subpockets at the desired locations. Only for IL-2, the first subpocket gets lost in the noVol-run. However, the volume of this pocket in the previous conformations points out that PocketInflator failed to make it fully accessible. In the Vol-run, a subpocket of this volume could be found, but here, the second subpocket is too small in most conformations. Interestingly, the second subpocket is always larger than the first one, even if no goal volume was defined. This finding suggests that an opening of a larger pocket is energetically more favorable at this second site. When focussing on the volume of the pocket induced on the surface of MDM2, it is not surprising that PocketInflator terminated without any solutions using the default value for the minimal score of an intermediate solution. The small size of the pockets induced in the noVol-run reveals that the enlargement of the pocket was energetically unfavorable. (Note that the $score_{hits}$ was sufficiently large as Figure 6.5 (f) demonstrates.) BCL-X_L was the only system for which the predefined number of conformations was generated. In the Vol-run, the volumes of the subpockets remained quite close to the goal values, but the low hit rates for the first subpocket shown in Figure 6.5 (a) suggest that this volume could be only obtained by moving the pocket a bit away from the desired location. When the subpocket volumes are not restricted to certain values, the resulting pockets are on the one hand smaller, but on the other hand they are really located at the desired location.



(a) BCL-X$_L$ (b) BCL-X$_L$

(c) IL-2 (d) IL-2

(e) MDM2 (f) MDM2

Figure 6.5: The compliance of the pockets designed by PocketInflator with the predefined properties plotted per generated conformation. Subfigures (a), (c), and (e) show the percentage of predefined residues effectively neighboring the induced (sub-) pockets and subfigures (b), (d), and (f) show the effective volume of the induced (sub-) pockets.

6.3.2 Docking into Pockets Designed by PocketInflator

As before, we docked the known inhibitors into the designed pockets to validate their appropriateness for virtual screening experiments. The results listed in Table 6.2 are very promising. For BCL-X$_L$ and IL-2, the docking scores are even better than those obtained when re-docking into the holo structures listed in Table 3.5 on page 65. While for IL-2 the results were very similar for the conformations generated in the two PocketInflator runs, the quality of the docking results improved for BCL-X$_L$ when the goal volume was not considered. Here, the slightly higher RMSD of

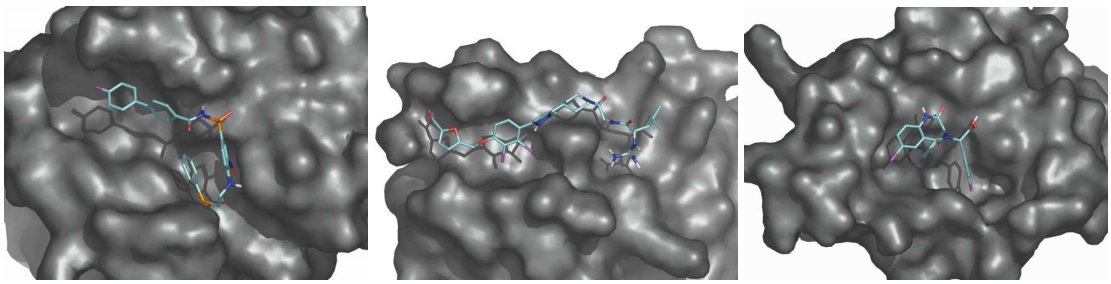
system	with goal volume				without goal volume			
	score [kcal/mol]	RMSD [Å]	rank [%]	solution no.	score [kcal/mol]	RMSD [Å]	rank [%]	solution no.
BCL-X_L - N3B	-11.4	1.3	2.4	24	-11.4	1.4	1.9	2
IL-2 - FRH	-12.4	2.2	20.0	5	-13.0	2.0	4.8	14
MDM2 - DIZ	-8.9	3.2	82.2	9	-8.6	2.0	61.2	7

Table 6.2: Docking results for conformations generated by PocketInflator. Shown are the best scored docking results with RMSD $\leq$ 2 Å or the docking result with lowest RMSD.

2.2 Å may be explained by the already mentioned observation that this subpocket was not exactly located at its native site. The native binding pocket of MDM2 was only designed successfully in the noVol-run. But here the score is significantly higher (by about 4.5 kcal/mol) than in the re-docking experiments. Furthermore, the high ranking of this docking result demonstrates that other docking poses not resembling the native binding mode were predicted to be more favorable. This suggests that the physicochemical properties of the designed pocket, especially the pocket volume, are not similar enough to those of the native binding pocket. One may speculate that the residues defining the desired pocket location are not really appropriate to represent the location of the native binding pocket. In contrast, the near-native docking solutions are very low ranked for the noVol-runs of the other two systems suggesting that in these cases, PocketInflator designed a pocket that is suitable for accommodating the known inhibitors. The PocketInflator conformations for which the best docking poses could be predicted are shown in Figure 6.6.

6.4 Discussion

We were quite surprised to learn from the docking results that the solutions generated in the noVol-runs were more appropriate for structure-based drug design than the solutions that contained pockets of the same magnitude than the native binding pocket. In fact, those conformations for which the best docking results were obtained possessed only small pockets. In the example of IL-2, the first subpocket was even missing. But when focussing on the residues lining the designed pockets it becomes evident that the best solutions with respect to the docking results correspond to those in which the pocket is lined by most protein residues used to define the pocket location. This observation suggests that the location of the pocket induced by PocketInflator is more important than its volume. Furthermore, AutoDock3 predicted binding to a pocket which was not detected by $EPOS^{BP}$. This raises the question whether only those pockets are druggable that were calculated using a clash factor of 0.95, or whether pockets whose detection required a reduction of the clash



(a) BCL-X_L (b) IL-2 (c) MDM2

Figure 6.6: The best scored docking poses with RMSD $\leq$ 2 Å obtained when docking into conformations generated by the noVol-run of PocketInflator corresponding to those listed in Table 6.2.

factor should also be taken into account. But here one should keep in mind that lowering the clash factor results in the detection of more pockets and, thus, in a longer run times.

6.4.1 Comparison to the PocketScanner/PocketBuilder Approach

The main idea - representing pockets by GPSs that interact with protein atoms and inducing pocket openings by energetically minimizing the protein structure in their presence - is identical to that of the PocketScanner/PocketBuilder approach. However, based on our previous experiences several important modifications were made in the implementation of PocketInflator: a pocket is no longer represented by a single GPS and the van der Waals interactions with the protein atoms are now calculated by a soft-core potential. Moreover, the pocket positions are not static anymore, they are chosen in dependence of the protein conformation. These improvements led to the generation of rather native-like, irregular shaped pockets. In addition, the conformations containing these pockets are less strained because the GPSs were treated flexible during the energy minimization and the energy penalty for clashes was damped by the soft-core potential. For the moment, we abdicated the local refinement using the A* algorithm. As the initial results were very promising, we presumed that the longer run times caused by additional A* searches (As multiple subpockets are considered in the PocketInflator approach, the number of flexible residues would be considerably higher than before.) would not compensate the expected slight improvement of the docking results. But this, of course, remains to be validated.
These algorithmic improvements implemented in PocketInflator have an impact on the docking results as well. When comparing those listed in Table 6.2 to the best docking results per system obtained using any settings of PocketBuilder (shown in Table 5.3 on page 94), it becomes evident that PocketInflator is better suited to design a druggable binding pocket than the PocketScanner/PocketBuilder approach. While the docking score for BCL-X_L is of the same order of magnitude, the RMSD is lower when using PocketInflator. For IL-2, although the RMSD is a bit higher, the docking score is almost by 6 kcal/mol more favorable than in the PocketScanner/PocketBuilder approach. Only for MDM2, the previous method achieved a slightly better docking score (by about 0.5 kcal/mol) although the RMSD was slightly worse. However, for all systems the rank of the best scored near-native docking pose was much lower when using PocketInflator and, in addition, fewer conformations were needed. But this may also be due to the fact that the location of the native binding sites was defined more precisely in PocketInflator.
Finally, it should be mentioned here that the two approaches were tested with different force fields. In order to exclude that the better performance of PocketInflator arises solely from this difference, the PocketScanner/PocketBuilder method has to be tested again using the Amber 96 force field.

6.4.2 Critical Assessment of the Approach

This approach requires a lot of *a priori* knowledge about the native binding site. Although we could show that the volumes of the (sub-) pockets that are to be designed are not essential, the definition of the residues that should line the (sub-) pocket is definitely crucial. So the more interacting residues are known, the more exactly is the position of the (sub-) pockets determined. However, if only a few hot spot residues are known, e.g. from mutagenesis studies, this may impose a problem. A further drawback of this approach is that (sub-) pockets are only defined to exist if they are detected by EPOS^{BP}. The example of IL-2 discussed before indicates that the existence of a pocket identified by our program may be not necessary for successful ligand docking. This raises the question how important the calculated pocket volume is. Obviously, the formula we use to calculate the score for the deviation of the goal volume is too strict, especially when taking into account that it is hard to define this value when no small-molecule binders are known. Moreover, $score_{volume}$ is only used to calculate the total score if the clash factor is 0.95 because the effective

volume of the pocket is hard to estimate using a smaller clash factor. When more overlaps with protein atoms are allowed, the pocket is larger. The selection of the intermediate solutions may be improved if the calculated volume would be considered even if the clash factor is small. In this case it would be recommendable to scale this volume down.
The advantage of this approach is that the minimum score ensures that only energetically favorable protein conformations with pockets fulfilling the predefined criteria are returned. When using a reasonable value, the formation of a pocket at a certain position or with a certain volume will not be enforced if it is energetically unfavorable. In such a case lowering this threshold loosens the strict criteria and allows for the opening of pockets which may deviate from the predefined properties, but are more native-like. This was shown at the example of MDM2, for which no solutions were found using the default minimal score of 0.3.
By using multiple GPSs that are flexible during the energy minimization and a soft-core potential for calculating their van der Waals interactions with the protein atoms, the shapes of the resulting (sub-) pockets are more native-like. However, the protein conformation still differs from the holo structure. As the apo and holo structures were resolved in different labs, often under different conditions, and using different methods, one cannot expect to obtain identical structures.

6.5 Summary and Conclusion

We presented a new method that combines pocket detection with EPOS^{BP} with the idea of PocketScanner and PocketBuilder of representing pockets by GPSs that interact with protein atoms via van der Waals interactions. This representation allows for inducing pocket openings by an energy minimization of the protein structure in the presence of these GPSs. But in contrast to the previous approach, multiple subpockets can now be induced at the same time. Each (sub-) pocket is modeled by multiple GPSs of varying size and their positions were calculated by the PASS algorithm as implemented in EPOS^{BP}. In order to form pockets at positions where no cavities were detectable in the starting structure, the approach starts by placing probes that overlap with protein atoms. Some of these probes are enlarged so that even more overlaps occur. These clashes are subsequently reduced by energy minimizations of the protein structure. Afterwards, new pockets are determined for the adopted structure, but now fewer clashes are tolerated. This procedure is repeated until protein conformations of low energy are generated that possess (sub-) pockets at predefined positions and are detectable even if no overlaps are tolerated. The resulting (sub-) pockets are more native-like than those designed by the PocketScanner/PocketBuilder method. By testing the approach with and without defining the goal volume, we observed that the pocket volume is not essentiel. In contrary, the definition of the residues that should line the designed (sub-) pockets is crucial. By docking the known inhibitors into these pockets we could show that they are indeed appropriate for accommodating ligands and may, thus, represent an efficient alternative method for the structure-based design of inhibitors binding to known sites for which no putative pockets could be detected in available crystal structures. At the moment, we are testing different parameters to further enhance the performance of this approach.

Chapter 7

Application of the Pocket Detection Protocol

After introducing the newly developed approaches for the detection, design, and analysis of transient pockets, we will now show the application of the MD-based protocol to two test systems for which the binding modes of small-molecule inhibitors are unknown. The first case study using the mitochondrial CYP11A1 electron transfer system is submitted for publication. The second case study addresses the XIAP protein that is involved in apoptosis.

7.1 Introduction

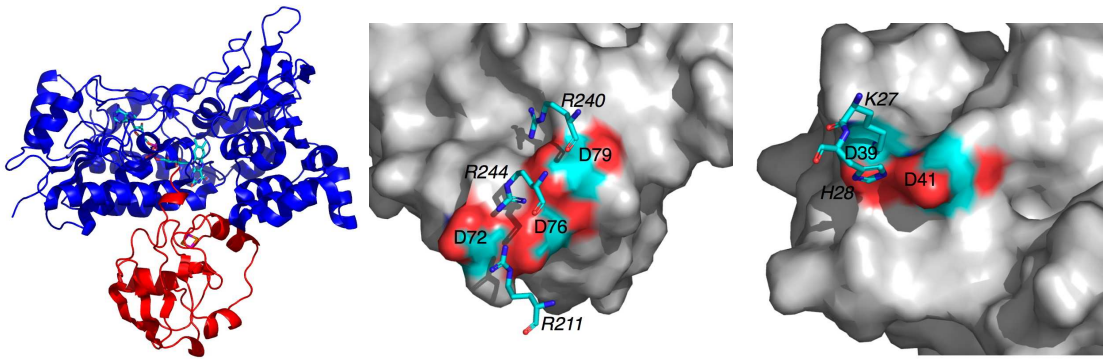
How *in silico* methods may assist the discovery of new hits and the design of new leads or drugs was discussed in Chapter 2. As most existing approaches rely on an *a priori* known binding region, structure-based design cannot be applied to those proteins for which the location of the binding site is unknown. Unfortunately, this is usually the case when the target protein is involved in protein-protein interactions. In Chapters 3 and 4, we presented a pocket detection protocol that is not only able to predict putative binding sites at which transient pockets open but also may suggest different protein conformations that may be appropriate for docking experiments. This protocol was validated using three model systems with known small-molecule inhibitors and for which crystal structures revealing their binding modes were resolved. In this Chapter, we present the application of the MD-based protocol to two proteins, Adrenodoxin (Adx) and the BIR2-domain of the X chromosome-linked inhibitor of apoptosis protein (XIAP), whose interactions with another protein are promising drug targets. For both proteins, hit compounds have been identified experimentally. Therefore, instead of asking "What binds and where?", we are capable of using the "what" to answer the "where". Note that knowing the approximate location of the binding site, e.g. the protein-protein interaction interface, may be not sufficient for assisting drug design as these surface regions may be huge compared to the size of the inhibitor (up to 1,500-3,000 $Å^2$ [25]) and non-contiguous. Furthermore, if this region does not contain accessible pockets in the known protein structure, structure-based drug design attempts will be limited.
In contrast to Adx, where the location of the inhibitor binding site is completely unknown and it could not be excluded that the compounds bind to a partner protein, the protein region to which the inhibitors identified for XIAP-BIR2 bind is surmised. Hence, the procedure for the two proteins differs. For Adx, all detected pockets, not only on the surface of Adx itself but also on its partner protein Adrenodoxin reductase, were considered in the docking experiments. Whereas for XIAP-BIR2, the pocket detection protocol was used to identify all pockets opening on the protein surface, but only those located at the presumed binding site were used for docking.

7.2 The Test Systems

Our computational study of the two test systems occurred in the framework of collaborations with the Biochemistry group of Prof. Dr. Rita Bernhardt at the Saarland University (Adx) and with Dr. Jose Luis Medina-Franco from the Computer Aided Drug Design devision of the Torrey Pines Institute for Molecular Studies in Florida (XIAP-BIR2). In the following, the two test systems are briefly introduced and characterized in terms of their physiological importance, structural details of the targeted interaction, and state of knowledge concerning the inhibitory mechanisms of the experimentally identified modulators.

7.2.1 Test System 1: Adrenodoxin

Adx is an important component of the mitochondrial CYP11A1 electron transfer system that catalyses the key step of steroid hormone biosynthesis, namely the oxidative side-chain cleavage of cholesterol to pregnenolone. Adrenodoxin reductase (AdR), a NADPH-dependent FAD containing reductase, transfers the electrons derived from NADPH to the iron-sulfur cluster of Adx that, in turn, reduces and so activates the molecular oxygen bound to the cytochrome P450, CYP11A1 [161]. In numerous subsequent hydroxylation steps pregnenolone is then converted to aldosterone. As increased concentrations of this steroid hormone cause hypertension and heart diseases, this step of the steroid hormone biosynthesis represents an interesting drug target [162]. The system has been studied in detail. Particularly, the three-dimensional atomic structures of bovine Adx, AdR, and the cross-linked Adx-AdR complex (shown in Figure 7.1) have been solved [163–166] and homology 3D-models of CYP11A1 and the CYP11A1 - Adx complex [167] are available. These structures in conjunction with site-directed mutagenesis studies [168] reveal that the binding sites on Adx for AdR and CYP11A1 overlap. While residues Asp^{72}, Glu^{73}, Asp^{76}, and Asp^{79} are most important in binding CYP11A1, the binding interface for AdR consists of two regions. In the primary region, Arg^{211}, Arg^{240}, and Arg^{244} of the NADPH-domain of AdR form numerous salt bridges to Asp^{72}, Asp^{76}, and Asp^{79} of Adx (Figure 7.1 (b)). The secondary interaction region is located on Adx around Asp^{39} and Asp^{41} that are in contact with Lys^{27} and His^{28} of AdR (Figure 7.1 (c)) [166]. All these studies indicate that the complex formations occurring between components of the CYP11A1 electron transport chain are mainly driven by electrostatic



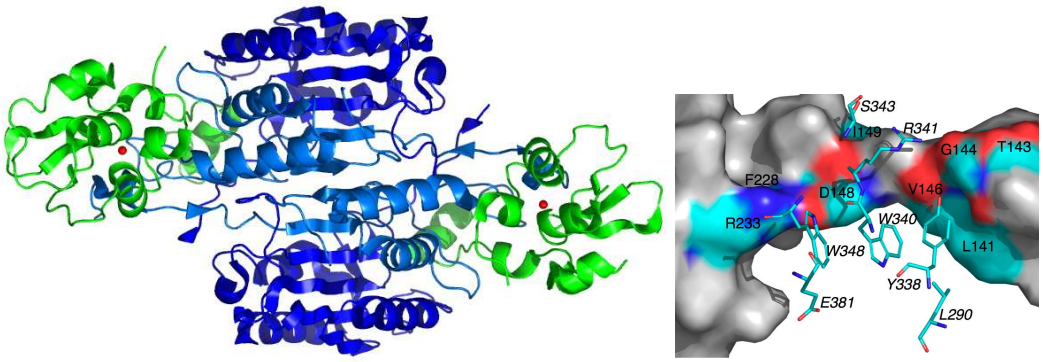
(a) Adx - AdR complex (b) primary interaction site (c) secondary interaction site

Figure 7.1: The Adx - AdR complex structure (PDB entry 1e6e). (a) Cartoon representation of Adx (red) and AdR (blue). The co-factors FAD and the Fe_2S_2 cluster are shown in licorized representation and are colored by element. (b) + (c) The molecular surface of Adx in the primary (b) and secondary (c) interaction region. The main interacting residues are colored by element and shown in licorized representation for AdR (italic residue labels) and in surface representation for Adx.

interactions. Especially the negatively charged residues on Adx seem to play a crucial role in the recognition of positively charged residues on AdR and CYP11A1 [169]. This suggests that these interactions may be modulated by positively charged molecules. Indeed, Berwanger et al. used a broad set of experimental techniques to show that the small, polycationic, and highly abundant natural polyamines putrescine (Put), spermidine (Spd), and spermine (Spn) modulate the interactions between Adx, AdR, and CYP11A1 (unpublished data). Interestingly, optical biosensor analysis of the binding affinities revealed that while the polyamines enhance the assembly of the Adx - AdR complex, the interaction between Adx and CYP11A1 is weakened. Although it was assumed that the polyamines bind to the negatively charged interface regions on Adx, their accurate binding site was unknown. As previous MD simulations revealed an increased flexibility in the binding regions of Adx [170], we applied our pocket detection protocol and docked the three polyamines into transient pockets that were observed during MD simulations of oxidized Adx in water. As it could not be ruled out that they bind to the surface of the other two proteins as well, we additionally docked the ligands into cavities found in the crystal structure of AdR and the Adx - AdR complex.

7.2.2 Test System 2: XIAP-BIR2

XIAP is the best characterized member of the Inhibitor of Apoptosis Proteins (IAPs) family. IAPs are endogenous caspase inhibitors [171, 172] that share a conserved structure, the BIR domain [173]. As caspases are responsible for apoptosis, their inhibition leads to the survival of damaged cells and, thus, to tumor proliferation [174, 175]. Not surprisingly, some IAP family proteins are commonly overexpressed in human cancers [176] and therefore important drug targets. The activity of XIAP is regulated by inhibitory proteins like Smac that disrupts XIAP-caspase complexes [177]. XIAP is composed of three BIR domains (called BIR1 to BIR3) and a RING zinc-finger motif. BIR2 and the linker region connecting BIR2 to BIR1 bind and inhibit caspase-3 and -7, while BIR3 suppresses caspase-9 [178, 179]. While the molecular details of the interactions with caspase-3 [180], -7 [181], and -9 [182], as well as the interaction of the BIR3 domain with Smac have been resolved [183, 184], it is still unclear whether Smac also binds to the BIR2 domain. The X-ray structure of the BIR2 - caspase-3 complex [180] shown in Figure 7.2 and site-directed mutagenesis studies [185] reveal that the interaction interface involves mainly the linker region



(a) XIAP-BIR2 - caspase-3 complex (b) linker region

Figure 7.2: The XIAP-BIR2 - caspase-3 complex structure (PDB entry 1i3o). (a) Cartoon representation of the homodimer complex between XIAP-BIR2 (green; zinc ions shown in red) and caspase-3 (P12 subunit: light blue; P17 subunit: dark blue). (b) The molecular surface of the linker region of XIAP-BIR2. The main interacting residues are colored by element and shown in licorized representation for caspase-3 (italic residue labels) and in surface representation for XIAP-BIR2.

(residues 124-168) of BIR2. Residues Leu^{140}, Leu^{141}, and Val^{146} and the caspase-3 residues Leu^{290}, Tyr^{338}, Trp^{340}, and Phe^{381} form a first hydrophobic cluster and Ile^{149} and Ile^{153} and the caspase-3 residue Phe^{381} a second one. Asp^{148} forms a salt bridge to Arg^{233} at the C-terminus of the BIR domain and hydrogen bonds to the caspase residues Arg^{341}, Ser^{343}, Trp^{348}, and Phe^{381}. Furthermore, a network of hydrogen bonds from the XIAP residues Thr^{143}, Gly^{144}, and Val^{146} contributes to the interaction (Figure 7.2 (b)). A second suggested interaction site is located on the BIR2 domain and contains a pocket that is topologically and chemically very similar to the Smac binding pocket of the BIR3 domain. But the site-directed mutagenesis studies suggest that this second site may be more important for the binding of caspase-7 than for caspase-3.
Researchers at the Torrey-Pines Institute identified a series of polyphenylurea-based compounds that selectively target the BIR2 domain of XIAP and stimulate an increased caspase-3 activity [186–188]. The mechanism of action of these compounds has been studied using biochemical, molecular biological, and genetic methods. Since the inhibitors did not compete with SMAC, it was assumed that they bind to the flexible linker region. However, the exact binding site and mode was unknown. This also impeded structure-based drug design attempts using the X-ray structure of the XIAP-BIR2 - caspase-3 complex. Within a collaboration with the Torrey-Pines Institute in Florida, we identified and analyzed transient pockets that open in this region during MD simulations in water or methanol and predicted potential binding modes by docking the three most potent inhibitors 1540-14, 1396-34, and 1396-11 into these pockets.

7.3 Methods

Transient pockets for both systems were identified with the MD-based pocket detection protocol. As the details of the MD simulations, pocket detection, and docking procedure differed slightly they will be separately described in the following.

7.3.1 Preparation of the Experimental Structures

The following X-ray structures of the CYP11A1 electron transfer system were used: oxidized Adx with a resolution of 1.85 Å (PDB entry 1ayf [164], chain A), AdR with a resolution of 1.7 Å (PDB entry 1cjc [165]), and the Adx-AdR complex with a resolution of 2.3 Å (1e6e [166], chains A and C). The partial charges, bond lengths, angles, and dihedrals for the oxidized Fe_2S_2 cluster of Adx were taken from [170]. The oxidized FAD was parametrized using the partial charges listed in [189].
The computational study of XIAP-BIR2 was based on two different experimental structures, the average NMR solution structure of the apo protein (PDB entry 1c9q [190]) and the X-ray structure of the complex between XIAP-BIR2 and caspase-3 (PDB entry 1i3o [180], chain E). As no parameterization of the Cys_3His-Zinc finger for the OPLS-AA force field was available, the parameters were derived computationally as described in the Appendix, Chapter D.

7.3.2 Molecular Dynamics Simulations

The dynamics of both proteins in water were simulated for 10 ns as described in Chapter 3. For Adx, the apo structure (including the co-factor) was used as starting structure. XIAP-BIR2 was simulated twice: one simulation started from the complex X-ray structure (after caspase-3 was manually removed) and one from the apo NMR structure. In addition, XIAP-BIR2 was simulated in methanol following the setup listed in Chapter 4. Here again, we conducted two simulations starting from the two different structures.
During all simulations, the proteins were fully flexible and 4,001 equally spaced snapshots were stored.

7.3.3 Pocket Detection

The MD simulations yielded one conformational ensemble for Adx and four for XIAP-BIR2. The pockets were identified, clustered, and analyzed using EPOS^{BP} as described in Chapter 4. For the CYP11A1 electron transfer system, the crystal structures of Adx, AdR, and the Adx-AdR complex were additionally scanned for pockets.

7.3.4 Docking Setup

For the CYP11A1 electron transfer system, all docking experiments were performed with AutoDock4. In this newer version, the scoring of the putative protein-ligand complexes is based on a semi-empirical free energy force field that incorporates intramolecular energies as well as a charge-based method for evaluating desolvation energies [100] and seemed therefore more suitable for predicting binding poses that are expected to be dominated by electrostatic interactions. The polyamines were prepared manually in their fully protonated forms (total charge of +2e for putrescine, +3e for spermidine, and +4e for spermine) and optimized with the MM+ force field as implemented in HyperChem [191]. AutoDockTools 1.4.6 was used for adding hydrogen atoms to the crystal structures (the MD snapshots already contained hydrogens), calculating Gasteiger atomic charges for the ligands and the receptors, and for assigning AutoDock4 atom types. For putrescine, 5 rotatable bonds were assigned with AutoTors, 9 for spermidine, and 13 for spermine. As before, the centers of the grid maps generated by AutoGrid4 were defined by the centers of mass of the pocket patches. The grid dimensions were chosen to be 26.25 Å x 26.25 Å x 26.25 Å to obtain an adequate coverage of the protein surface and the grid spacing was set to the default value of 0.375 Å. The docking procedure followed the standard LGA protocol with default parameters. 50 independent docking runs were carried out for each pocket detected in a crystal structure and 25 for each pocket detected in a MD snapshot.
For XIAP-BIR2, the docking experiments were performed with AutoDock3 because this version is much faster than the AutoDock4 version and the ligands are quite flexible with 12 rotatable bonds for 1540-14, 13 for 1396-34, and 16 for 1396-11. The ligands were set as neutral, Gasteiger atom charges were assigned, and AutoTors was used for defining the flexible torsions. The MD snapshots were prepared and the grid maps with a dimension of 26.25 Å x 26.25 Å x 26.25 Å were calculated as described in Chapter 3 for the PID-docking. The docking procedure also followed the standard LGA protocol, but with one exception: the initial population was increased to 150 randomly placed individuals to obtain a broader sampling of the docking poses. Here, 20 independent docking runs were carried out for each pocket.

7.3.5 Post-Processing of the Docking Poses

As the number of calculated docking poses was too high for visual inspection, only those that obtained a docking score better than a given threshold were retained and clustered using an agglomerative single-linkage approach based on the match of the protein atoms within 5 Å. For docking into the MD snapshots of Adx, the thresholds were -5 kcal/mol for putrescine, -8 kcal/mol for spermidine, and -9 kcal/mol for spermine. For XIAP-BIR2, only docking scores lower than -12 kcal/mol for 1540-14, -14 kcal/mol for 1396-34, and -16 kcal/mol for 1396-11 were considered.

7.4 Results

In this section, we report the application of the MD-based pocket detection protocols to the two test systems. By docking into the identified pockets we were able to suggest several energetically favorable binding sites that will be discussed in the following.

7.4.1 Putative Binding Sites Detected on the Surface of Adx and AdR

The particular challenge in the computational study of the CYP11A1 electron transfer system was that although it was assumed that the polyamines bind to the Adx protein, binding to CYP11A1 or AdR could not be excluded. We confined our search for putative binding sites to Adx and AdR as no crystal structures for CYP11A1 are available and using homology models for docking is quite error-prone due to the intrinsic uncertainties of the modeled structures.

Pockets Detected in the MD Snapshots and the Crystal Structures The crystal structures of oxidized Adx, AdR, and the Adx-AdR complex were subjected to a pocket detection run. 9 cavities were found on the surface of apo Adx (with volumes in the range of 306 to 826 Å^3) and 50 on apo AdR (with volumes in the range of 247 to 1,050 Å^3). From the 42 detected cavities on the complex structure, only 2 were located on the surface of Adx, 28 on AdR, and 12 were identified at the binding interface with volumes ranging from 282 to 954 Å^3. As it was assumed that the binding site is located on Adx, this protein was submitted to a careful examination with our MD-based pocket detection protocol. As Adx was stable during the MD simulation in water (see section B.4), the extracted snapshots were scanned for pockets. In total, 23 different transient pockets were identified with volumes up to 1,513 Å^3.

Detecting Favorable Binding Sites by Docking The polyamines were docked into the transient pockets of the MD snapshots of Adx, as well as into the pockets detected in the three crystal structures. Clustering the best scored docking poses suggested five putatively favorable binding sites that are illustrated in Figure 7.3. Hereof two binding sites are located on Adx, two on AdR, and one at the binding interface (corresponding to neither the first nor the second interaction site). Note that docking suggested favorable binding sites at positions where no pockets were detected by EPOS^{BP}.

The corresponding docking scores listed in Table 7.1 suggest that the three polyamines prefer binding to Adx although binding to negatively charged patches on AdR and the Adx-AdR complex is also possible (compare to Table 7.2). Interestingly, the two binding sites located on Adx are the most favorable ones and correspond to the primary (binding site 5) and secondary AdR interaction

binding site	max. score in Adx [kcal/mol]			max. score in AdR [kcal/mol]			max. score in Adx - AdR [kcal/mol]			max. score in Adx MD snapshots [kcal/mol]		
	Put	Spd	Spn	Put	Spd	Spn	Put	Spd	Spn	Put	Spd	Spn
1	-7.6	-8.9	-9.4	-	-	-	-4.7	-5.4	-5.4	-7.1	-8.8	-10.4
2	-	-	-	-	-3.1	-3.3	-5.7	-7.0	-7.9	-	-	-
3	-	-	-	-5.5	-6.7	-5.9	-5.7	-7.7	-8.5	-	-	-
4	-	-	-	-4.6	-5.1	-5.1	-5.9	-7.1	-7.6	-	-	-
5	-6.5	-7.4	-7.6	-	-	-	-4.8	-6.6	-6.5	-9.0	-10.4	-11.6

Table 7.1: Overview of the binding sites on Adx and AdR that are predicted to be most favorable for polyamine binding by flexible ligand docking. For each binding site, the best docking scores per polyamine and protein are reported. The numbering of binding sites corresponds to that used in Figure 7.3.

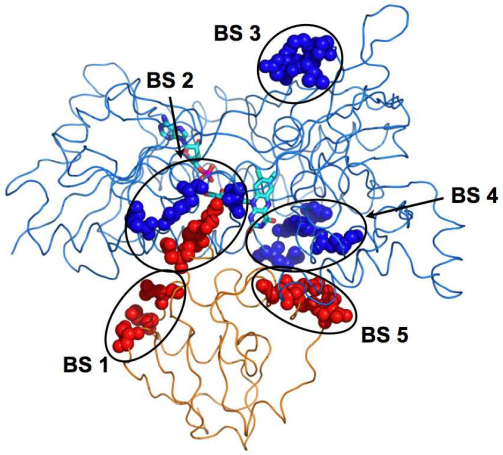


Figure 7.3: Location of the identified binding sites (BS) mapped on the Adx-AdR complex. Adx is shown in orange, AdR in light blue. For both proteins, a ribbon representation is used except for the residues lining the binding sites (blue: AdR residues, red: Adx residues) that are shown as ball-and-stick models and FAD that is shown in licorized representation (colorized by element).

sites (binding site 1). The docking scores suggest that although the affinity is reduced at these sites upon Adx-AdR assembly, binding of the polyamines is still possible. Furthermore, docking scores for binding site 5 improve when the results for the Adx crystal structure are compared to the results for the Adx MD snapshots, whereas docking scores for binding site 1 are similar for MD snapshots and X-ray structure. For these two binding sites, the best scored docking complexes per polyamine among the MD snapshots of Adx are shown in Figure 7.4. This finding suggests that the intrinsic dynamics of apo Adx favors the binding of the polyamines to binding site 5. (Note that the best results for the MD snapshots were selected from dockings to 4,001 different Adx conformers, whereas the best results for the crystal structures were only selected from dockings to different sites of the same conformer.) The affinities of the two putative binding sites located on AdR are markedly increased in the AdR-Adx complex. As binding site 3 is located distant from the binding interface, it may appear surprising that the affinity of this binding sites changes upon complex formation. However, the two X-ray structures show AdR in two different conformational states [164]. These differences may well reflect the spread among different conformational substates that are assigned in the two classes. Especially the docking scores for binding site 2 (complex interface) are significantly reduced with apo AdR as docking receptor and no appropriate binding sites were predicted for apo Adx.

To evaluate the docking approach and to test how favorable the docking scores are, we re-docked the polyamines into six selected co-crystal structures of polyamine binding proteins taken from the PDB. The re-docking scores listed in Table 7.3 are of comparable magnitude as the scores for docking into binding site 1 (when using either the crystal structure or the MD snapshots of

binding site	residues on Adx	residues on AdR
1	Asp^{15}, Asp^{39}, Asp^{41}	-
2	Asp^{113}, Glu^{116}, Ser^{117}	Glu^{353}, Arg^{370}, Thr^{373}
3	-	Ala^{109}, Asp^{111}, Glu^{115}, Glu^{116}
4	-	Asp^{54}, His^{55}, Glu^{57}, Glu^{212}
5	Asp^{72}, Glu^{73}, Asp^{76}, Asp^{79}	-

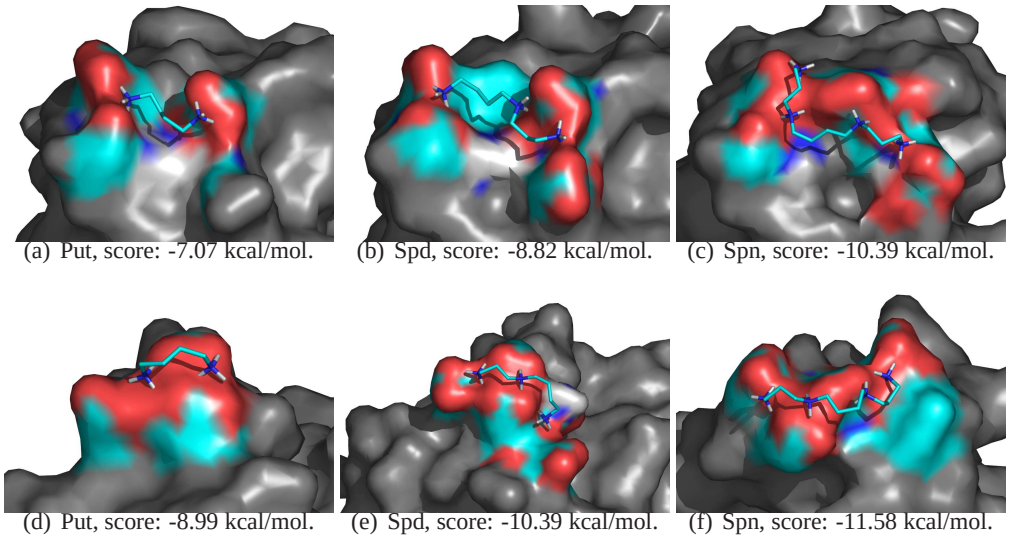
Table 7.2: Protein residues of Adx and AdR interacting with the polyamines in the most favorable binding sites listed in Table 7.1 and illustrated in Figure 7.3.

PDB entry	description	re-docking score [kcal/mol]	RMSD [Å]
1a99	Putrescine bound to E. coli Putrescine Binding Protein [192]	-7.9	0.6
2o06	Putrescine bound to Human Spermidine Synthase [193]	-8.0	1.5
1pot	Spermidine bound to E. coli Spermidine Binding Protein [194]	-11.7	1.1
3c6k	Spermidine bound to Human Spermidine Synthase [195]	-10.3	0.9
3b7p	Spermine bound to Plasmodium Falciparum Spermidine Synthase	-13.5	1.1
3c6m	Spermine bound to Human Spermine Synthase [195]	-14.0	1.8

Table 7.3: The selected co-crystal structures of the polyamines and their binding proteins and the corresponding re-docking results with AutoDock4. (The docking was performed as described in section 7.3.4.)

Adx) and the scores for docking into binding site 5 (when using MD snapshots). This suggests that induced-fit effects provide an additional stabilization of 2-4 kcal/mol, but also that the conformational flexibility sampled in molecular dynamics simulations at room temperature is sufficient to generate binding pockets of comparable binding affinities as those of proteins known to bind polyamines. In contrast, the scores for docking into the crystal structures of AdR and the Adx-AdR complex are less favorable than the re-docking results suggesting that there is a clear preference for polyamine binding to apo Adx.

Experimental Validation of the Binding Sites by Site-Directed Mutagenesis To verify these docking results experimentally, two Adx mutants (D15K and D15N) were created by Anja Berwanger in the Biochemistry department. As Asp^{15} is located in the secondary interaction domain that is of crucial importance for polyamine binding to Adx as suggested by the docking experiments, the exchange of the negative charge with a neutral (D15N) and a positive one (D15K) should lead to remarkable differences in the affinity of Adx to its redox partner AdR and CYP11A1. As shown in Table 7.4, the introduction of the neutral and even more of the positive charge resulted in a significantly increased affinity of Adx to AdR (decreased K_d). In contrast, the polyamines weakened



(a) Put, score: -7.07 kcal/mol. (b) Spd, score: -8.82 kcal/mol. (c) Spn, score: -10.39 kcal/mol.

(d) Put, score: -8.99 kcal/mol. (e) Spd, score: -10.39 kcal/mol. (f) Spn, score: -11.58 kcal/mol.

Figure 7.4: The most favorable docking poses per ligand in binding site 1 ((a) - (c)) and 5 ((d) - (f)) on MD snapshots of Adx. The protein is shown in grey surface representation and the atoms within 5 Å of the ligand as well as the ligand itself (shown as sticks) are colored by element. The corresponding schematic representations of the binding modes are shown in the Appendix, Fig. A.3.

system	Adx - AdR assembly			Adx - CYP11A1 assembly		
	k_{on} [$M^{-1}s^{-1}$]	k_{off} [s^{-1}]	K_d [M]	k_{on} [$M^{-1}s^{-1}$]	k_{off} [s^{-1}]	K_d [M]
WT (control)	$3.81 \cdot 10^3$	$2.86 \cdot 10^{-3}$	$7.50 \cdot 10^{-7}$	$1.83 \cdot 10^5$	$6.76 \cdot 10^{-3}$	$3.69 \cdot 10^{-8}$
WT + Put	$3.12 \cdot 10^4$	$5.03 \cdot 10^{-3}$	$1.61 \cdot 10^{-7}$	$8.64 \cdot 10^3$	$3.11 \cdot 10^{-3}$	$3.60 \cdot 10^{-7}$
WT + Spd	$2.74 \cdot 10^4$	$4.71 \cdot 10^{-3}$	$1.71 \cdot 10^{-7}$	$1.86 \cdot 10^4$	$3.53 \cdot 10^{-3}$	$1.90 \cdot 10^{-7}$
WT + Spn	$1.76 \cdot 10^4$	$5.77 \cdot 10^{-3}$	$3.28 \cdot 10^{-7}$	$2.71 \cdot 10^4$	$2.79 \cdot 10^{-3}$	$1.03 \cdot 10^{-7}$
D15N	$1.97 \cdot 10^4$	$1.83 \cdot 10^{-3}$	$9.29 \cdot 10^{-8}$	$3.76 \cdot 10^3$	$4.14 \cdot 10^{-3}$	$1.01 \cdot 10^{-6}$
D15K	$1.10 \cdot 10^4$	$3.46 \cdot 10^{-3}$	$3.15 \cdot 10^{-7}$	$2.08 \cdot 10^3$	$4.41 \cdot 10^{-3}$	$2.12 \cdot 10^{-6}$

Table 7.4: Experimental data obtained by Anja Berwanger: Optical biosensor analysis of the interactions of oxidized AdxWT, and the mutants AdxD15K and AdxD15N with AdR_{ox} and $CYP11A1_{ox}$ in presence and absence of polyamines (all at ionic strength I = 1 mM). Binding of AdR or CYP11A1 (both analytes 100 - 500 nM) to Adx immobilized on a CM5 chip ($\sim$300 RU) was studied in Biacore HBS-EP buffer at 25°C. Binding curves of the interaction partners were analyzed using the Biacore evaluation software 4.1 with a 1:1 binding model. Standard deviations (n$\geq$4) were within $\pm$ 10% of the displayed values. The K_d values were calculated by k_{off}/k_{on}.

the stronger Adx-CYP11A1 binding.
The polyamine binding sites on the Adx protein that we suggested are well suited to explain the experimental Adx - AdR binding data. As all three polyamines seem to preferably bind to the primary (Asp^{72}, Glu^{73}, Asp^{76}, binding site 5) and secondary (Asp^{15}, Asp^{39}, Asp^{41}, binding site 1) AdR binding regions of apo Adx, they may also promote the complex formation by overcoming repulsive charges between the two proteins, resulting in a faster reduction of Adx [166, 196, 197]. This is also consistent with the kinetic constants measured for the assembly of the oxidized Adx with AdR and CYP11A1 (Table 7.4) that indicate that the molecular recognition and thus the association of Adx and AdR is enhanced in the presence of polyamines, while it is decreased for CYP11A1. As a result, the Adx - AdR complex tightens and the Adx - CYP11A1 complex weakens. In both complexes, modulation took place either in the order putrescine < spermidine < spermine, or in the order neutral charge < positive charge. This data supports our hypothesis that the secondary binding region around Asp^{15} is not only important for the protein-protein recognition but it is also a specific interaction site of the polyamines with the Adx.

7.4.2 Putative Binding Sites Detected in the Linker Region of XIAP-BIR2

For this system, the region of the binding site was only tentatively known. In this case, conformational sampling by MD and detection of transient pockets on the entire protein surface seemed the most reliable protocol. Additionally to the simulation in water, this protein was also simulated in methanol and both simulations were run twice, either starting from the apo NMR structure or from the holo X-ray structure (after removal of caspase-3). XIAP-BIR2 was stable in all simulations. The secondary structure remained most conserved during the MD simulation of the X-ray structure in methanol (see Section B.5, Appendix).

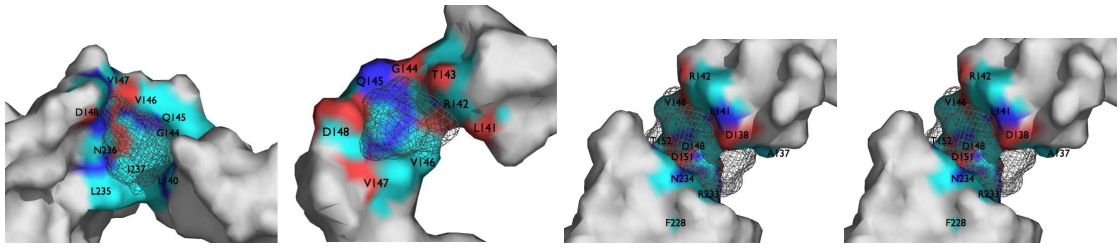
Pockets Detected in the MD Snapshots of XIAP-BIR2 EPOS^{BP} was applied to the snapshots extracted from all MD simulations to identify transient pockets. After removal of all pockets appearing only once, a set of 41 different transient pockets was obtained that are spread all over the protein surface. Surprisingly, of these 41 pockets, 9 were located in the linker region. The properties of these pockets are reported in Table 7.5. They are all overlapping but were not assigned to the same cluster because their lining protein residues vary too much depending on the MD simulation setup. The properties of these distinct transient pockets also strongly differ, suggesting that these pockets are highly mobile and adaptable.

PID	MD setup	residues	freq. [%]	mean vol. [Å^3]	polarity
18	NMR in methanol	124-126, 129-131, 133-135, 137, 140, 141, 144-148, 235-237	42.5	360.9	0.36
19	NMR in methanol	145-149, 151-153, 228, 231- 239	10.0	476.8	0.36
25	NMR in water	145-149, 151-157	7.7	143.5	0.36
28	NMR in water	140-148	7.0	175.1	0.36
29	X-ray in methanol	137, 141, 146, 148, 150, 225-228, 231-235, 237	62.8	342.8	0.35
31	X-ray in methanol	141, 146, 148, 150, 154, 203, 204, 224-228, 233, 234	46.7	282.3	0.34
36	X-ray in methanol	141, 142, 144, 146- 149, 151, 152, 233	22.7	290.0	0.36
38	X-ray in water	141, 146-151, 154, 228, 233-236	13.9	183.2	0.38
41	X-ray in water	148, 150, 154, 202, 203, 224, 226-228, 231, 233	13.3	189.8	0.32

Table 7.5: The transient pockets detected in the linker region and their most frequently occurring pocket lining residues, frequency, mean volume, and polarity. Note that although all pockets overlap, they were assigned to different clusters (PIDs) and so each transient pocket shown here was only observed in one MD simulation.

Detecting Favorable Binding Sites by Docking into Promising Transient Pockets From our previous experience on the BCL-X_L, IL-2, and MDM2 systems, individual snapshots with pocket volumes larger than 200 Å^3 appear to be promising candidates for docking studies. Examples of promising transient pockets are illustrated in Figure 7.5. Thus, all transient pockets located in the linker region having a pocket volume $\geq$ 200 Å^3 were selected as putative binding sites for the ligands 1540-14, 1396-34, and 1396-11. This resulted in the selection of 6,662 pockets from the four different MD simulations (1,624 from the NMR structure in water, 137 from the NMR structure in methanol, 418 from the X-ray structure in water, and 4,483 from the X-ray structure in methanol).

As expected, ligand 1396-11 is the most potent inhibitor, followed by ligand 1396-34. The clustering of the docking poses revealed 61 different favorable binding sites. When considering only those clusters having at least 100 members for at least one ligand, the selection can be reduced to the 14 most favorable binding sites compiled in Table 7.6. For all ligands, the best docking score is predicted for binding site 2. The best docking poses of each inhibitor are shown in Figure 7.6. Interestingly, for binding sites 11 to 14, only ligand 1540-14 achieved a docking score smaller



(a) PID 18 (b) PID 28 (c) PID 29 (d) PID 38

Figure 7.5: Examples for transient pockets opening in the flexible linker region of XIAP-BIR2. The pocket patch representing the negative image of the pockets is shown as mesh, the pocket lining atoms are colored by element and the residues lining the pocket are labeled.

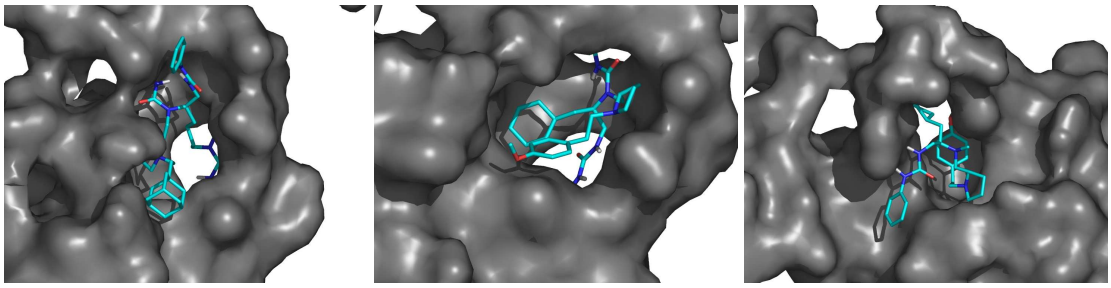
than the used cut-off, suggesting that the local conformations of these sites favor binding of the smallest and most rigid ligand. Note that a transient pocket ID does not correspond to a binding site cluster, because the transient pockets as well as the binding sites are overlapping as shown in Table 7.7 and the size of the grid maps allow the ligands to bind to vicinal cavities.
Structural studies suggested that the inhibition of caspase-3 and by XIAP-BIR2 is achieved by a two-site interaction. Besides the N-terminal linker of BIR2 (residues 124-168), it was hypothesized that the putative Smac binding pocket is also involved in caspase binding [180]. The hot spots identified in this study are Asp^{148}, Glu^{219}, and His^{223}. However, as binding assays suggested that the inhibitors do not compete with Smac for a common binding site, this region was

binding site	**ligand 1396-11**			**ligand 1396-34**			**ligand 1540-14**		
	best score [kcal/mol]	mean score [kcal/mol]	no. poses	best score [kcal/mol]	mean score [kcal/mol]	no. poses	best score [kcal/mol]	mean score [kcal/mol]	no. poses
1	-19.4	-16.7	327	-17.2	-14.7	254	-15.1	-12.9	522
2	-19.6	-16.9	265	-17.9	-14.8	253	-16.3	-12.8	1542
3	-19.5	-16.7	348	-17.1	-14.5	803	-15.2	-12.6	522
4	-19.2	-16.7	267	-17.6	-14.7	199	-15.1	-12.6	839
5	-19.0	-16.6	254	-16.5	-14.5	514	-15.6	-12.5	651
6	-18.0	-16.6	67	-16.1	-14.6	153	-14.5	-12.6	434
7	-19.2	-16.8	164	-16.7	-14.7	263	-14.9	-12.7	216
8	-18.7	-16.6	306	-16.0	-14.5	135	-14.1	-12.5	167
9	-17.6	-16.4	70	-16.4	-14.5	53	-14.7	-12.6	251
10	-17.3	-16.5	33	-15.7	-14.4	172	-13.4	-12.4	75
11	-	-	-	-	-	-	-15.6	-12.6	2331
12	-	-	-	-	-	-	-14.4	-12.5	240
13	-	-	-	-	-	-	-14.2	-12.5	237
14	-	-	-	-	-	-	-14.5	-12.5	104

Table 7.6: The most favorable binding sites on XIAP-BIR2 as identified by docking with AutoDock3. Here, only those binding sites are shown that have at least 100 members (no. of poses) for at least one of the three ligands.

binding site	**residues**	**PIDs**
1	124-126, 128-131, 134, 137, 140, 141, 145, 146, 235, 236	18, 19, 28
2	146-153, 226-228, 231, 232, 234-238	18, 19, 25, 29, 31, 36, 38
3	137, 141, 146-148, 150, 151, 226-228, 231-237	18, 19, 29, 31, 36, 38, 41
4	146-153, 228, 231, 234-238	18, 19
5	137, 141, 146-148, 150, 226-228, 231-237	18, 19, 28, 29, 31, 36
6	125, 126, 128-131, 134, 140, 141, 145-148, 236, 237	18, 19, 28
7	124, 146, 148-153, 161, 228, 231, 234-237	18, 19, 25, 29
8	141, 147, 148, 150, 226-228, 231-235	18, 29, 31, 36, 38
9	125, 126, 128-131, 140, 141, 145-147, 236	18, 19, 28
10	146-149, 151-153, 228, 231, 234-238	18, 19
11	141, 147, 148, 150, 226-228, 231-234	29, 31
12	141, 148, 150, 226-228, 231-236	29, 31, 36, 38
13	125, 126, 128, 129, 131, 140, 141, 145, 146, 148, 236	18, 19, 28
14	124-131, 134-138	18, 19, 28

Table 7.7: The protein residues lining the most favorable binding sites and the IDs of the transient pockets that led to the calculation of these binding poses. Note that the same binding site was predicted although the ligands were docked into different transient pockets and also into snapshots extracted from different MD simulations.



(a) 1396-11, score: -19.6 kcal/mol. (b) 1396-34, score: -17.9 kcal/mol. (c) 1540-14, score: -16.3 kcal/mol.

Figure 7.6: The most favorable docking poses per ligand. The protein is shown in grey surface representation and the ligand is shown as sticks colored by element. The corresponding schematic representation of the binding modes are shown in the Appendix, Figure A.4. Note that all plots show the ligands bound to binding site 2 on a MD snapshot of the NMR structure in water.

not considered in our docking experiments. Interestingly, almost all favorable docking complexes involve interactions between the inhibitors and Asp^{148} (see Table 7.7 and Fig. A.4). These results support the hypothesis that the polyphenylurea-based inhibitors bind to the flexible linker region of XIAP-BIR2 and so impede the assembly of the XIAP - caspase-3 complex.

7.5 Discussion

In this chapter, we presented the application of our pocket detection protocol to two test systems, for which binders have been identified experimentally. Structure-based design approaches aiming at optimizing these hits to lead compounds were hampered as their binding mode was unknown. However, in contrast to Adx, where the entire protein surface (as well as the surface of its partner proteins) had to be considered in the docking experiments, the approximate binding site of the inhibitors identified for XIAP-BIR2 was known and the docking experiments could be limited to the transient pockets located within this region. A further difference between the two applications is that Adx was only simulated in water using only one starting structure while XIAP-BIR2 was additionally simulated in methanol and each simulation was repeated with a different starting structure showing the protein in another conformation (either in its apo state or complexed to caspase-3). The challenge with the XIAP protein was that no OPLS-AA force field parameters were available for the Cys_3His-Zinc finger. These values were derived by density functional theory-based quantum mechanical calculations and had to be tested thoroughly in MD simulations. Moreover, unlike Adx, for which we expected the polyamines to bind into charged pockets that are more favorable to open in water, the pockets accommodating the inhibitors of XIAP-BIR2 were expected to be rather nonpolar and, thus, better sampled during a simulation in methanol.

We decided to confine the conformational sampling to MD simulations because as demonstrated in Chapter 3 and 4, MD was the only method that yielded reliable results for all three model systems. As the binding region was approximately known for XIAP-BIR2, the PocketScanner/PocketBuilder or the PocketInflator approach could have also been applied. However, as the linker region is extremely flexible as indicated by the MD simulations, it is questionable whether the energy minimization and side-chain rearrangement would have been effective enough to induce the openings of pockets comparable to those that were observed using MD.

Moreover, the application of this protocol to XIAP-BIR2 stressed an interesting characteristic of protein-protein interaction interfaces. The fact that the same binding site was predicted although the ligands were docked into different transient pockets indicates that these binding sites are strewn with small pockets and that small molecules binding to protein-protein interaction interfaces often

occupy several (sub-) pockets at the same time. This observation suggests that when trying to identify binding sites for SMPPIIs on protein surfaces, one should rather focus on regions where (small) transient pockets accumulate than on isolated pockets.

7.6 Summary and Conclusion

Structure-based drug design does not only assist the identification of hits, but also the selection or refinement of hit compounds to lead structures, or of lead structures to drug candidates. Especially when binders were identified by experimental methods, the elucidation of their binding site and mode is a precondition for the successful application of computational methods assisting the identification of lead compounds or drug candidates. As case study, we discussed two proteins involved in protein-protein interactions for which small-molecule binders were identified by *in vitro* experiments but their binding site and, thus, the molecular details of their interaction were unknown. For one protein, the BIR2 domain of XIAP, the binding site was assumed to be located in a very flexible region. For the other protein, Adx, the location of the binding site was totally unknown and it could even not be excluded that the modulators (also) bind to its partner proteins. We applied our pocket detection protocol to both proteins to identify transient pockets that open on the protein surface. The experimentally identified ligands were then docked into all pockets that were accessible in the MD snapshots or in the crystal structures and the most favorable docking poses could be clustered into five putative binding sites. Of these sites, the most favorable ones were located on the Adx protein and corresponded to known sites of interaction with AdR. The plausibility of these binding sites was supported by site-directed mutagenesis studies.

As the binding site of the molecules targeting XIAP-BIR2 was approximately known, the docking experiment could be restricted to those pockets that opened within this region. The differing properties of the detected transient pocket emphasized the flexibility of this region. By clustering the most favorable docking poses, 14 putative binding sites were identified. Interestingly, Asp^{148}, a hot spot for the interaction of XIAP with caspase-3 that is targeted by the studied ligands, is involved in almost all favorable docking poses suggesting that the predicted binding modes are reasonable.

In summary, the application of our pocket detection protocol to two test proteins indicated that it is capable of suggesting plausible binding sites and ligand binding modes regardless of the *a priori* knowledge about the location of the binding region. However, whether this protocol can also be used to predict binders and non-binders remains to be evaluated in the future.

Chapter 8

Conclusion and Outlook

Protein-protein interaction interfaces are a real challenge for structure-based drug design. Due to their intrinsic properties they often contain no deep, accessible pockets that may be targeted by small-molecule ligands. As a consequence, *in silico* drug design approaches were so far limited to those systems for which the interface either contained a druggable pocket in the unbound structure or for which a crystal structure with a small molecule bound was available.

8.1 Summary and Conclusion

The goal of this work was the development of protocols that assist the structure-based design of small-molecule protein-protein interaction inhibitors (SMPPIIs) by providing protein conformations that contain transient pockets which may be targeted in virtual screening experiments. We tested all developed methods using three model systems, BCL-X_L, IL-2, and MDM2. All these proteins are involved in protein-protein interactions, and the native binding pockets of SMPPIIs (known from the crystal structures of the protein-inhibitor complex) are not, or not fully, accessible in the apo structures.

In our initial study presented in Chapter 3, we showed that the binding pockets of these SMPPIIs are, like many other pockets on the protein surface, only accessible in some conformations: They are transient binding pockets. When no druggable pockets are detectable in any available structure of the protein, this initial pocket detection protocol may thus be an interesting starting point. Molecular dynamics simulations of the apo structure in water were conducted and all trajectory snapshots were scanned for cavities using the PASS algorithm. All detected pockets were subsequently clustered to determine the distinct transient pockets. We found that they all opened within 2.5 ps, and most of them appeared multiple times. They were even reproducible by a second MD simulation. The general impression was quite similar for all three systems. At the native binding site, transient pockets could be identified that were of similar size than the native binding pocket. To validate the appropriateness of this protocol for virtual screening, we docked the known inhibitors with AutoDock3 into these identified transient pockets. For all systems we obtained docking poses that were within 2 Å RMSD of the native binding mode.

In the follow-up study described in Chapter 4, we investigated which aspects of the natural conformational dynamics of proteins induce the formation of these transient pockets. The same pocket detection protocol was applied to three different conformational ensembles that were extracted from three different MD simulations; (a) of the inhibitor bound structure (after removal of this ligand) in water, (b) of the apo structure in water (that was used in Chapter 3), and (c) of the apo structure in methanol. For MDM2, we additionally studied the impact of backbone mobility by MD simulations in which all backbone atoms were harmonically restrained. The results emphasized the influence of solvent polarity and backbone rearrangements on the formation of transient pockets. Furthermore this study revealed that the native binding pocket is unstable in the absence of the ligand explaining why it is only partly accessible or even absent in the apo structure. More-

over, we tested whether the more efficient CONCOORD, tCONCOORD or normal mode analysis (NMA) techniques may substitute the time-consuming MD simulations. While the conformations generated by CONCOORD and NMA possessed significantly smaller pockets, only the tCONCOORD conformations contained pockets that were comparable to those observed in MD simulations for two of the three systems. This finding indicates that MD simulations are to date the most robust method to sample transient pockets if the binding site is unknown and the entire protein surface has to be taken into account.

In many structure-based approaches the binding site of the ligand is approximately known. In such cases, running the MD-based pocket detecion protocol appears quite time consuming as the sampling can be limited to this region of the surface. It is not clear whether this protocol is successful for all kinds of binding pockets. For example, at some binding sites the presence of the ligand may be required to induce the pocket opening, whereas the MD-based protocol is only capable of finding cavities that open spontaneously. For these reasons we developed two algorithmic approaches that design pockets of desired properties in a predefined region of the protein surface. Based on our findings presented in Chapter 4, these methods account for protein backbone and side chain flexibility. The main idea of both approaches is to represent the pocket by a "generic pocket sphere" (GPS) that interacts with the protein atoms. The first approach is discussed in Chapter 5. It starts by scanning the protein surface for potential pocket positions using a grid of predefined size. At each grid point, a GPS is placed and the protein is minimized energetically while the position of the GPS remains fixed. Subsequently the residues lining this pocket are then further refined by searching for the best combinations of side-chain rotamers using the A* algorithm. For two out of the three test systems, conformations could be generated with pockets into which the known inhibitor could be docked in a native-like orientation. However, due to their representation by a single GPS the designed pockets were of an artifical shape. Moreover, many SMPPIIs consist of multiple subpockets that cannot be induced at the same time using this method. All these considerations indicate that the applicability of this approach is limited.

In the second algorithmic approach presented in Chapter 6, we tried to solve these problems that emerged in Chapter 5. Here, multiple subpockets of predefined volume and location are designed simultaneously. Rather than representing them by a single GPS, they are now represented by patches of coherent probes that were placed by the PASS algorithm. As the initial structure usually contains no detectable pockets, we modified the PASS algorithm in such a way that probes that overlap with protein atoms up to a certain degree are kept. This degree can be controlled by our program. After selecting such a precursor-pocket, some probes are enlarged, and the protein conformation is energetically minimized. By doing so, the protein adopts its conformation to this pocket and the number of clashes is reduced. Subsequently, new precursor-pockets are calculated for this relaxed protein conformation. Thereby it is taken care that the degree of allowed overlap is always smaller or equal to that in the previous cavity detection. These steps are repeated until a predefined number of low-energy protein conformations are designed that contain pockets detectable without tolerating overlaps larger than the default value. One can pictorially describe this procedure as inflating pockets in the protein surface. Surprisingly, the results indicated that the target volumes of the designed pockets are not crucial. We found that when considering only the locations in the design process, the resulting pockets are more native-like than those designed by the first approach. This became also evident from the scores obtained in the docking experiment. For all systems, docking poses within 2 Å from the native binding mode were suggested. For two systems, these poses even obtained better docking scores than those observed in the inhibitor-bound crystal structure. Therefore, we suggest using this protocol for cases, in which the binding site is known, but contains no druggable pockets in the available crystal structures.

After validating the developed approaches, we show in Chapter 7 their application to two systems. Although small-molecule modulators were identified experimentally for both systems, their bind-

ing mode was unknown. In case of the first system, it was even unclear whether the molecules bind to the targeted protein, Adrenodoxin, or its interacting protein, Adrenodoxin Reductase. We used the MD-based pocket detection protocol for identifying transient pockets opening anywhere on the surface of Adrenodoxin and additionally considered all other pockets found in the crystal structures of Adrenodoxin Reductase and the protein-protein complex. In the second system, the BIR2 domain of XIAP, the region in which the binding site is located could be delineated. Even though this would enable the application of the pocket inflating protocol presented in Chapter 6, we decided to conduct MD simulations in water and in methanol because this protein region was highly flexible and it was not clear whether the algorithmic approach could handle this extreme mobility correctly in vacuo. For both test systems, we were capable to suggest favorable binding sites by docking the ligands into transient pockets that were identified by our pocket detection protocol. These real-world examples emphasize the applicability and usefulness of our presented protocols.
In summary, we think that our findings will be helpful in future generation of transient pockets as putative ligand binding sites at protein-protein interfaces or even for the identification of new allosteric pockets for any kind of proteins.

8.2 Outlook

The design of small-molecule protein-protein interaction inhibitors is a relatively new and very interesting research field. However, our studies suffered from the low number of model systems currently available. We hope that owing to the continuous progress made in this field more and more high quality crystal structures of proteins in complex with their SMPPIIs will become available. This would enable us to verify our protocols on a larger number of test systems. An important question is, for example, whether the native binding pocket may be identified from a set of transient pockets. More generally, it would be of advantage if one could narrow down the number of conformations in which the transient pocket under consideration is available and extract those pocket states that are most druggable. By doing so, the time needed to dock the putative ligands may be significantly reduced. Furthermore, it would be interesting to investigate whether one can predict those regions on the protein surface where transient pockets will open. If this was possible, one could systematically induce pockets using the algorithmic approach at these sites. In addition, one could use GPS of different properties (e.g. with charges) to influence the chemical properties of the designed pockets. Finally, it is worth testing whether our approach is only applicable to detect pockets for competitive inhibitors, i.e. pockets opening at protein-protein interaction interfaces, or whether it can be generally applied to identify new allosteric pockets that are not accessible in the absence of a ligand.

Bibliography

[1] H Kitano. Systems biology: a brief overview. *Science*, 295:1662–1664, 2002.

[2] SK Sharma, TM Ramsey, and KW Bair. Protein-protein interactions: lessons learned. *Curr Med Chem Anti-cancer Agents*, 2:311–330, 2002.

[3] H Lodish, D Baltimore, and A Berk. *Molecular Cell Biology*. W H Freeman & Co (Sd), 5rd edition, 2003.

[4] E Alm and AP Arkin. Biological networks. *Curr Opin Struc Biol*, 13:193–202, 2003.

[5] S Boccaletti, V Latora, Y Moreno, M Chavez, and DU Hwang. Complex networks: Structure and dynamics. *Phys Rep*, 424:175 – 308, 2006.

[6] HJ Böhm, G Klebe, and H Kubinyi. *Wirkstoffdesign*. Spektrum Akad. Verl., 1996.

[7] AS Reddy, SP Pati, PP Kumar, HN Pradeep, and GN Sastry. Virtual screening in drug discovery – a computational perspective. *Curr Protein Pept Sci*, 8:329–351, 2007.

[8] BO Villoutreix, K Bastard, O Sperandio, R Fahraeus, JL Poyet, et al. In silico-in vitro screening of protein-protein interactions: towards the next generation of therapeutics. *Curr Pharm Biotechnol*, 9:103–122, 2008.

[9] AL Hopkins and CR Groom. The druggable genome. *Nat Rev Drug Discov*, 1:727–730, 2002.

[10] X Barril, R E Hubbard, and SD Morley. Virtual screening in structure-based drug discovery. *Mini Rev Med Chem*, 4:779–791, 2004.

[11] PD Lyne. Structure-based virtual screening: an overview. *Drug Discov Today*, 7:1047–1055, 2002.

[12] FL Stahura and J Bajorath. New methodologies for ligand-based virtual screening. *Curr Pharm Des*, 11:1189–1202, 2005.

[13] AC Anderson. The process of structure-based drug design. *Chem Biol*, 10:787–797, 2003.

[14] E Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber Dtsch Chem Ges*, 27, 1894.

[15] DE Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A*, 44:98–104, 1958.

[16] HR Bosshard. Molecular recognition by induced fit: how fit is the concept? *News Physiol Sci*, 16:171–173, 2001.

[17] H Wieman, K Tondel, E Anderssen, and F Drablos. Homology-based modelling of targets for rational drug design. *Mini Rev Med Chem*, 4:793–804, 2004.

[18] J Liang, H Edelsbrunner, and C Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*, 7:1884–1897, 1998.

[19] DC Fry. Protein-protein interactions as targets for small molecule drug discovery. *Biopolymers*, 84:535–552, 2006.

[20] AI Archakov, VM Govorun, AV Dubanov, YD Ivanov, AV Veselovsky, P Lewi, and P Janssen. Protein-protein interactions as a target for drugs in proteomics. *Proteomics*, 3:380–391, 2003.

[21] IMA Nooren and JM Thornton. Diversity of protein–protein interactions. *EMBO J*, 22:3486–3492, 2003.

[22] S Zhong, AT Macias, and AD Jr MacKerell. Computational identification of inhibitors of protein-protein interactions. *Curr Top Med Chem*, 7:63–82, 2007.

[23] MR Arkin and JA Wells. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov*, 3:301–317, 2004.

[24] AV Veselovsky, Yu D Ivanov, AS Ivanov, AI Archakov, P Lewi, and P Janssen. Protein-protein interactions: mechanisms and modification by drugs. *J Mol Recognit*, 15:405–422, 2002.

[25] JA Wells and CL McClendon. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450:1001–1009, 2007.

[26] D Gonzalez-Ruiz and H Gohlke. Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. *Curr Med Chem*, 13:2607–2625, 2006.

[27] T Berg. Modulation of protein-protein interactions with small organic molecules. *Angew Chem Int Ed Engl*, 42:2462–2481, 2003.

[28] L Pagliaro, J Felding, K Audouze, SJ Nielsen, RB Terry, C Krog-Jensen, and S Butcher. Emerging classes of protein-protein interaction inhibitors and new tools for their development. *Curr Opin Chem Biol*, 8:442–449, 2004.

[29] H Yin and AD Hamilton. Strategies for targeting protein-protein interactions with synthetic agents. *Angew Chem Int Ed Engl*, 44:4130–4163, 2005.

[30] PL Toogood. Inhibition of protein-protein association by small molecules: approaches and progress. *J Med Chem*, 45:1543–1558, 2002.

[31] T Clackson and JA Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267:383–386, 1995.

[32] SB Shuker, PJ Hajduk, RP Meadows, and SW Fesik. Discovering high-affinity ligands for proteins: Sar by nmr. *Science*, 274:1531–1534, 1996.

[33] DA Erlanson, AC Braisted, DR Raphael, M Randal, RM Stroud, EM Gordon, and JA Wells. Site-directed ligand discovery. *Proc Natl Acad Sci U S A*, 97:9367–9372, 2000.

[34] JY Trosset, C Dalvit, S Knapp, M Fasolini, M Veronesi, S Mantegani, et al. Inhibition of protein-protein interactions: the discovery of druglike beta-catenin inhibitors by combining virtual and biophysical screening. *Proteins*, 64:60–67, 2006.

[35] M Mallya, RL Phillips, SA Saldanha, B Gooptu, SCL Brown, DJ Termine, et al. Small molecules block the polymerization of z alpha1-antitrypsin and increase the clearance of intracellular aggregates. *J Med Chem*, 50:5357–5363, 2007.

[36] AL Bowman, Z Nikolovska-Coleska, H Zhong, S Wang, and HA Carlson. Small molecule inhibitors of the mdm2-p53 interaction discovered by ensemble-based receptor models. *J Am Chem Soc*, 129:12809–12814, 2007.

[37] P Block, N Weskamp, A Wolf, and G Klebe. Strategies to search and design stabilizers of protein-protein interactions: a feasibility study. *Proteins*, 68:170–186, 2007.

[38] R Najmanovich, J Kuttner, V Sobolev, and M Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins*, 39:261–268, 2000.

[39] MJ Betts and MJ Sternberg. An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng*, 12:271–283, 1999.

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

[41] F Jensen. *Introduction to Computational Chemistry*. John Wiley & Sons, 2006.

[42] A Leach. *Molecular Modelling: Principles and Applications (2nd Edition)*. Prentice Hall, 2001.

[43] WD Cornell, P Cieplak, CI Bayly, IR Gould, KM Merz, DM Ferguson, DC Spellmeyer, T Fox, JW Caldwell, and PA Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc*, 117:5179–5197, 1995.

[44] A D MacKerell, D Bashford, Bellott, R L Dunbrack, J D Evanseck, M J Field, S Fischer, J Gao, H Guo, S Ha, D Joseph-McCarthy, L Kuchnir, K Kuczera, F T K Lau, C Mattos, S Michnick, T Ngo, D T Nguyen, B Prodhom, W E Reiher, B Roux, M Schlenkrich, J C Smith, R Stote, J Straub, M Watanabe, J Wiorkiewicz-Kuczera, D Yin, and M Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102:3586–3616, 1998/04/01.

[45] L Schuler, X Daura, and W van Gunsteren. An improved gromos96 force field for aliphatic hydrocarbons in the condensed phase. *J Comput Chem*, 22:1205–1218, 2001.

[46] WL Jorgensen, DS Maxwell, and J Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc*, 118:11225–11236, 1996.

[47] T Darden, D York, and L Pedersen. Particle mesh Ewald: An N. log (N) method for Ewald sums in large systems. *J Chem Phys*, 98:10089–10092, 1993.

[48] P Cozzini, GE Kellogg, F Spyrakis, DJ Abraham, G Costantino, et al. Target flexibility: an emerging consideration in drug discovery and design. *J Med Chem*, 51:6237–6255, 2008.

[49] M Christen and WF van Gunsteren. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *J Comput Chem*, 29:157–166, 2008.

[50] A Ahmed, S Kazemi, and H Gohlke. Protein Flexibility and Mobility in Structure-Based Drug Design. *Front Drug Des Discov*, 3:455–476, 2007.

[51] WF Van Gunsteren and HJC Berendsen. Algorithms for macromolecular dynamics and constraint dynamics. *Mol Phys*, 34:1311–1327, 1977.

[52] PL Freddolino, F Liu, M Gruebele, and K Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding ww domain. *Biophys J*, 94:L75–7, 2008.

[53] N Go, T Noguti, and T Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci U S A*, 80:3696–3700, 1983.

[54] B Brooks and M Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A*, 80:6571–6575, 1983.

[55] S Hayward and N Go. Collective Variable Description Of Native Protein Dynamics. *Annu Rev Phys Chem*, 223:50, 1995.

[56] F Tama and YH Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Eng*, 14:1–6, 2001.

[57] J Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13:373–380, 2005.

[58] BL de Groot, DM van Aalten, RM Scheek, A Amadei, G Vriend, and HJ Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29:240–251, 1997.

[59] D Seeliger, J Haas, and BL de Groot. Geometry-based sampling of conformational transitions in proteins. *Structure*, 15:1482–1492, 2007.

[60] R Chandrasekaran and GN Ramachandran. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformation in proteins. *Int J Protein Res*, 2:223, 1970.

[61] J Janin and S Wodak. Conformation of amino acid side-chains in proteins. *J Mol Biol*, 125:357, 1978.

[62] JW Ponder and FM Richards. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol*, 193:775, 1987.

[63] RL Jr Dunbrack and M Karplus. Backbone-dependent rotamer library for proteins. application to side-chain prediction. *J Mol Biol*, 230:543–574, 1993.

[64] RL Jr Dunbrack. Rotamer libraries in the 21st century. *Curr Opin Struct Biol*, 12:431–440, 2002.

[65] J Desmet, MD Maeyer, B Hazes, and I Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[66] AA Canutescu, AA Shelenkov, and RL Jr Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12:2001–2014, 2003.

[67] C Hartmann, I Antes, and T Lengauer. Irecs: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci*, 16:1294–1307, 2007.

[68] M Nayal and B Honig. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, 63:892–906, 2006.

[69] B Huang and M Schroeder. Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. *BMC Struct Biol*, 6:19, 2006.

[70] GP Jr Brady and PF Stouten. Fast prediction and visualization of protein binding pockets with pass. *J Comput Aided Mol Des*, 14:383–401, 2000.

[71] RA Laskowski. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph*, 13:323–330, 1995.

[72] DG Levitt and LJ Banaszak. Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph*, 10:229–234, 1992.

[73] M Hendlich, F Rippmann, and G Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*, 15:359–363, 1997.

[74] ATR Laurie and RM Jackson. Q-sitefinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21:1908–1916, 2005.

[75] MR Landon, DR Jr Lancia, J Yu, SC Thiel, and S Vajda. Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. *J Med Chem*, 50:1231–1240, 2007.

[76] CM Stultz and M Karplus. Mcss functionality maps for a flexible protein. *Proteins*, 37:512–529, 1999.

[77] RA Laskowski, NM Luscombe, MB Swindells, and JM Thornton. Protein clefts in molecular recognition and function. *Prot Sci*, 5:2438–2452, 1996.

[78] M Morita, S Nakamura, and K Shimizu. Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins*, 73:468–479, 2008.

[79] PJ Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem*, 28:849–857, 1985.

[80] DB Kitchen, H Decornez, JR Furr, and J Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov*, 3:935–949, 2004.

[81] SF Sousa, PA Fernandes, and MJ Ramos. Protein-ligand docking: current status and future challenges. *Proteins*, 65:15–26, 2006.

[82] ID Kuntz, JM Blaney, SJ Oatley, R Langridge, and TE Ferrin. A geometric approach to macromolecule-ligand interactions. *J Mol Biol*, 161:269–288, 1982.

[83] K Raha and KM Jr Merz. 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*, 48:4558–4575, 2005.

[84] TI Oprea. Property distribution of drug-related chemical databases. *J Comput Aided Mol Des*, 14:251–264, 2000.

[85] TJ Ewing, S Makino, AG Skillman, and ID Kuntz. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des*, 15:411–428, 2001.

[86] M Rarey, B Kramer, T Lengauer, and G Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261:470–489, 1996.

[87] RA Friesner, JL Banks, RB Murphy, TA Halgren, JJ Klicic, DT Mainz, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J Med Chem*, 47:1739–1749, 2004.

[88] GM Morris, DS Goodsell, RS Halliday, R Huey, WE Hart, RK Belew, and AJ Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem*, 19, 1998.

[89] PFW Stouten, C Frömmel, H Nakamura, and C Sander. An effective solvation term based on atomic occupancies for use in protein simulations. *Mol Simulat*, 10:97–120, 1993.

[90] K Gunasekaran and R Nussinov. How different are structurally flexible and rigid binding sites? sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J Mol Biol*, 365:257–273, 2007.

[91] ML Verdonk, PN Mortenson, RJ Hall, MJ Hartshorn, and CW Murray. Protein-ligand docking against non-native protein conformers. *J Chem Inf Model*, 48:2214–2225, 2008.

[92] RE Amaro, R Baron, and JA McCammon. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J Comput Aided Mol Des*, 22:693–705, 2008.

[93] JH Lin, AL Perryman, JR Schames, and JA McCammon. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J Am Chem Soc*, 124:5632–5633, 2002.

[94] JH Lin, AL Perryman, JR Schames, and JA McCammon. The relaxed complex method: Accommodating receptor flexibility for drug design with an improved scoring scheme. *Biopolymers*, 68:47–62, 2003.

[95] H Claussen, C Buning, M Rarey, and T Lengauer. Flexe: efficient molecular docking considering protein structure variations. *J Mol Biol*, 308:377–395, 2001.

[96] MI Zavodszky and LA Kuhn. Side-chain flexibility in protein-ligand binding: The minimal rotation hypothesis. *Protein Sci*, 14, 2005.

[97] AR Leach. Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol*, 235:345–356, 1994.

[98] G Jones, P Willett, and RC Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol*, 245:43–53, 1995.

[99] G Jones, P Willett, RC Glen, AR Leach, and R Taylor. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267:727–748, 1997.

[100] R Huey, GM Morris, AJ Olson, and DS Goodsell. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem*, 28:1145–1152, 2007.

[101] J Meiler and D Baker. Rosettaligand: protein-small molecule docking with full side-chain flexibility. *Proteins*, 65:538–548, 2006.

[102] M Zacharias. Rapid protein-ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: binding of fk506 to fkbp. *Proteins*, 54:759–767, 2004.

[103] CN Cavasotto, JA Kovacs, and RA Abagyan. Representing receptor flexibility in ligand docking through relevant normal modes. *J Am Chem Soc*, 127:9632–9640, 2005.

[104] Y Zhao, D Stoffler, and M Sanner. Hierarchical and multi-resolution representation of protein flexibility. *Bioinformatics*, 22:2768–2774, 2006.

[105] Y Zhao and MF Sanner. Flipdock: docking flexible ligands into flexible receptors. *Proteins*, 68:726–737, 2007.

[106] A Cavalli, G Bottegoni, C Raco, M De Vivo, and M Recanatini. A computational study of the binding of propidium to the peripheral anionic site of human acetylcholinesterase. *J Med Chem*, 47:3991–3999, 2004.

[107] J Apostolakis, A Pluckthun, and A Caflisch. Docking small ligands in flexible binding sites. *J Comput Chem*, 19, 1998.

[108] SB Nabuurs, M Wagener, and J de Vlieg. A flexible approach to induced fit docking. *J Med Chem*, 50:6507–6518, 2007.

[109] S Eyrisch and V Helms. Transient pockets on protein surfaces involved in protein-protein interaction. *J Med Chem*, 50:3457–3464, 2007.

[110] T Frembgen-Kesner and AH Elcock. Computational sampling of a cryptic drug binding site in a protein receptor: explicit solvent molecular dynamics and inhibitor docking to p38 map kinase. *J Mol Biol*, 359:202–214, 2006.

[111] V Helms. Protein dynamics tightly connected to the dynamics of surrounding and internal water molecules. *ChemPhysChem*, 8:23–33, 2007.

[112] HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, H Weissig, IN Shindyalov, and PE Bourne. The protein data bank. *Nucleic Acids Res*, 28:235–242, 2000.

[113] XM Yin, ZN Oltvai, and SJ Korsmeyer. Bh1 and bh2 domains of bcl-2 are required for inhibition of apoptosis and heterodimerization with bax. *Nature*, 369:321–323, 1994.

[114] V Kirkin, S Joos, and M Zornig. The role of bcl-2 family members in tumorigenesis. *Biochim Biophys Acta*, 1644:229–249, 2004.

[115] T Oltersdorf, SW Elmore, AR Shoemaker, RC Armstrong, DJ Augeri, BA Belli, et al. An inhibitor of bcl-2 family proteins induces regression of solid tumours. *Nature*, 435:677–681, 2005.

[116] SW Muchmore, M Sattler, H Liang, RP Meadows, JE Harlan, HS Yoon, D Nettesheim, et al. X-ray and nmr structure of human bcl-xl, an inhibitor of programmed cell death. *Nature*, 381:335–341, 1996.

[117] T Chittenden, C Flemington, AB Houghton, RG Ebb, GJ Gallo, B Elangovan, G Chinnadurai, and RJ Lutz. A conserved domain in bak, distinct from bh1 and bh2, mediates cell death and protein binding functions. *EMBO J*, 14:5589–5596, 1995.

[118] M Sattler, H Liang, D Nettesheim, RP Meadows, JE Harlan, M Eberstadt, et al. Structure of bcl-xl-bak peptide complex: recognition between regulators of apoptosis. *Science*, 275:983–986, 1997.

[119] PJ Hajduk, JR Huth, and SW Fesik. Druggability indices for protein targets derived from nmr-based screening data. *J Med Chem*, 48:2518–2525, 2005.

[120] SP Brown and PJ Hajduk. Effects of conformational dynamics on predicted protein druggability. *ChemMedChem*, 1:70–72, 2006.

[121] W Novak, Hg Wang, and G Krilov. Role of protein flexibility in the design of bcl-x(l) targeting agents: insight from molecular dynamics. *J Comput Aided Mol Des*, 23:49–61, 2009.

[122] KA Smith. Interleukin-2: inception, impact, and implications. *Science*, 240:1169–1176, 1988.

[123] M Rickert, X Wang, MJ Boulanger, N Goriatcheva, and KC Garcia. The structure of interleukin-2 complexed with its alpha receptor. *Science*, 308:1477–1480, 2005.

[124] SF Liparoto and TL Ciardelli. Biosensor analysis of the interleukin-2 receptor complex. *J Mol Recognit*, 12:316–321, 1999.

[125] AC Church. Clinical advances in therapies targeting the interleukin-2 receptor. *QJM-Int J Med*, 96:91–102, 2003.

[126] J Theze, PM Alzari, and J Bertoglio. Interleukin 2 and its receptors: recent advances and new immunological functions. *Immunol Today*, 17:481–486, 1996.

[127] MR Arkin, M Randal, WL DeLano, J Hyde, TN Luong, JD Oslob, et al. Binding of small molecules to an adaptive protein-protein interface. *Proc Natl Acad Sci U S A*, 100:1603–1608, 2003.

[128] CD Thanos, M Randal, and JA Wells. Potent small-molecule binding to a dynamic hot spot on il-2. *J Am Chem Soc*, 125:15280–15281, 2003.

[129] KM Ryan, AC Phillips, and KH Vousden. Regulation and function of the p53 tumor suppressor protein. *Curr Opin Cell Biol*, 13:332–337, 2001.

[130] D Michael and M Oren. The p53-mdm2 module and the ubiquitin system. *Semin Cancer Biol*, 13:49–58, 2003.

[131] M Hollstein, D Sidransky, B Vogelstein, and CC Harris. p53 mutations in human cancers. *Science*, 253:49–53, 1991.

[132] DI Zheleva, DP Lane, and PM Fischer. The p53-mdm2 pathway: targets for the development of new anticancer therapeutics. *Mini Rev Med Chem*, 3:257–270, 2003.

[133] S Uhrinova, D Uhrin, H Powers, K Watt, D Zheleva, P Fischer, C McInnes, and PN Barlow. Structure of free mdm2 n-terminal domain reveals conformational adjustments that accompany p53-binding. *J Mol Biol*, 350:587–598, 2005.

[134] PH Kussie, S Gorina, V Marechal, B Elenbaas, J Moreau, AJ Levine, and NP Pavletich. Structure of the mdm2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science*, 274:948–953, 1996.

[135] BL Grasberger, T Lu, C Schubert, DJ Parks, TE Carver, HK Koblish, , et al. Discovery and cocrystal structure of benzodiazepinedione hdm2 antagonists that activate p53 in cells. *J Med Chem*, 48:909–912, 2005.

[136] CP Barrett, BA Hall, and MEM Noble. Dynamite: a simple way to gain insight into protein motions. *Acta Cryst D*, 60:2280–2287, 2004.

[137] LM Espinoza-Fonseca and JG Trujillo-Ferrara. Conformational changes of the p53-binding cleft of mdm2 revealed by molecular dynamics simulations. *Biopolymers*, 83:365–373, 2006.

[138] SG Dastidar, DP Lane, and CS Verma. Multiple peptide conformations give rise to similar binding affinities: molecular simulations of p53-mdm2. *J Am Chem Soc*, 130:13514–13515, 2008.

[139] W Humphrey, A Dalke, and K Schulten. VMD: visual molecular dynamics. *J Mol Graph*, 14:33–38, 1996.

[140] PIW de Bakker, MA DePristo, DF Burke, and TL Blundell. Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the amber force field with the generalized born solvation model. *Proteins*, 51:21–40, 2003.

[141] E Lindahl, B Hess, and D van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model*, 7:306–317, 2001.

[142] WL Jorgensen, J Chandrasekhar, JD Madura, RW Impey, and ML Klein. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*, 79:926, 1983.

[143] HJC Berendsen, JPM Postma, WF Van Gunsteren, A DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *J Chem Phys*, 81:3684, 1984.

[144] B Hess, H Bekker, HJC Berendsen, and JGEM Fraaije. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem*, 18, 1997.

[145] MF Sanner. Python: a programming language for software integration and development. *J Mol Graph Model*, 17:57–61, 1999.

[146] J Gasteiger and M Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36:3219–3228, 1980.

[147] S Eyrisch and V Helms. What induces pocket openings on protein surface patches involved in protein-protein interactions? *J Comput Aided Mol Des*, 23:73–86, 2009.

[148] DR Lide. *CRC handbook of chemistry and physics: a ready-reference book of chemical and physical data*. CRC press, 2004.

[149] DO Alonso and V Daggett. Molecular dynamics simulations of protein unfolding and limited refolding: characterization of partially unfolded states of ubiquitin in 60methanol and in water. *J Mol Biol*, 247:501–520, 1995.

[150] H Kovacs, AE Mark, J Johansson, and WF van Gunsteren. The effect of environment on the stability of an integral membrane helix: molecular dynamics simulations of surfactant protein C in chloroform, methanol and water. *J Mol Biol*, 247:808–822, 1995.

[151] RA Engh and R Huber. Accurate bond and angle parameters. *Acta Crystallogr A*, 47:392–400, 1991.

[152] O Kohlbacher and HP Lenhof. Ball–rapid software prototyping in computational molecular biology. biochemicals algorithms library. *Bioinformatics*, 16:815–824, 2000.

[153] MA McCoy, JJ Gesell, MM Senior, and DF Wyss. Flexible lid to the p53-binding domain of human mdm2: implications for p53 regulation. *Proc Natl Acad Sci U S A*, 100:1645–1648, 2003.

[154] S Eyrisch and V Helms. Designing binding pockets on protein surfaces using the a* algorithm. *Lecture Notes in Informatics (LNI)*, P-136:64–74, 2008.

[155] PE Hart, NJ Nilsson, and B Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Trans. on SSC*, 4:100–107, 1968.

[156] G Bottegoni, I Kufareva, M Totrov, and R Abagyan. A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (scare). *J Comput Aided Mol Des*, 22:311–325, 2008.

[157] IM Withers, MP Mazanetz, H Wang, PM Fischer, and CA Laughton. Active site pressurization: a new tool for structure-guided drug design and other studies of protein flexibility. *J Chem Inf Model*, 48:1448–1454, 2008.

[158] T Lazaridis and M Karplus. Effective energy function for proteins in solution. *Proteins*, 35:133–152, 1999.

[159] PA Kollman. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc Chem Res*, 29:461–469, 1996.

[160] M Zacharias, TP Straatsma, and JA McCammon. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J Chem Phys*, 100:9025, 1994.

[161] R Bernhardt. Cytochrome p450: structure, function, and generation of reactive oxygen species. *Rev Physiol Biochem Pharmacol*, 127:137–221, 1996.

[162] B Pitt, F Zannad, W J Remme, R Cody, A Castaigne, A Perez, J Palensky, and J Wittes. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. randomized aldactone evaluation study investigators. *N Engl J Med*, 341:709–717, 1999.

[163] D Beilke, R Weiss, F Lohr, P Pristovsek, F Hannemann, R Bernhardt, and H Ruterjans. A new electron transport mechanism in mitochondrial steroid hydroxylase systems based on structural changes upon the reduction of adrenodoxin. *Biochemistry*, 41:7969–7978, 2002.

[164] A Muller, JJ Muller, YA Muller, H Uhlmann, R Bernhardt, and U Heinemann. New aspects of electron transfer revealed by the crystal structure of a truncated bovine adrenodoxin, adx(4-108). *Structure*, 6:269–280, 1998.

[165] GA Ziegler, C Vonrhein, I Hanukoglu, and GE Schulz. The structure of adrenodoxin reductase of mitochondrial p450 systems: electron transfer for steroid biosynthesis. *J Mol Biol*, 289:981–990, 1999.

[166] JJ Muller, A Lapko, G Bourenkov, K Ruckpaul, and U Heinemann. Adrenodoxin reductase-adrenodoxin complex structure suggests electron transfer path in steroid biosynthesis. *J Biol Chem*, 276:2786–2789, 2001.

[167] SA Usanov, SE Graham, GI Lepesheva, TN Azeva, NV Strushkevich, AA Gilep, RW Estabrook, and JA Peterson. Probing the interaction of bovine cytochrome p450scc (cyp11a1) with adrenodoxin: evaluating site-directed mutations by molecular modeling. *Biochemistry*, 41:8310–8320, 2002.

[168] VM Coghlan and LE Vickery. Site-specific mutations in human ferredoxin that affect binding to ferredoxin reductase and cytochrome p450scc. *J Biol Chem*, 266:18606–18612, 1991.

[169] AV Grinberg, F Hannemann, B Schiffler, J Muller, U Heinemann, and R Bernhardt. Adrenodoxin: structure, stability, and electron transfer properties. *Proteins*, 40:590–612, 2000.

[170] SK Shakya, W Gu, and V Helms. Molecular dynamics simulation of truncated bovine adrenodoxin. *Biopolymers*, 78:9–20, 2005.

[171] N Roy, QL Deveraux, R Takahashi, GS Salvesen, and JC Reed. The c-IAP-1 and c-IAP-2 proteins are direct inhibitors of specific caspases. *EMBO J*, 16:6914–6925, 1997.

[172] QL Deveraux, N Roy, HR Stennicke, T Van Arsdale, Q Zhou, SM Srinivasula, ES Alnemri, GS Salvesen, and JC Reed. IAPs block apoptotic events induced by caspase-8 and cytochrome c by direct inhibition of distinct caspases. *EMBO J*, 17:2215–2223, 1998.

[173] JC Reed. The survivin saga goes in vivo. *J Clin Invest*, 108:965–969, 2001.

[174] GM Cohen et al. Caspases: the executioners of apoptosis. *Biochem J*, 326:1–16, 1997.

[175] NA Thornberry and Y Lazebnik. Caspases: enemies within. *Science*, 281:1312–1316, 1998.

[176] I Tamm, SM Kornblau, H Segall, S Krajewski, K Welsh, S Kitada, DA Scudiero, G Tudor, YH Qui, A Monks, et al. Expression and Prognostic Significance of IAP-Family Genes in Human Cancers and Myeloid Leukemias 1, 2000.

[177] C Du, M Fang, Y Li, L Li, and X Wang. Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating iap inhibition. *Cell*, 102:33–42, 2000.

[178] R Takahashi, Q Deveraux, I Tamm, K Welsh, N Assa-Munt, GS Salvesen, and JC Reed. A single BIR domain of XIAP sufficient for inhibiting caspases, 1998.

[179] QL Deveraux, E Leo, HR Stennicke, K Welsh, GS Salvesen, and JC Reed. Cleavage of human inhibitor of apoptosis protein XIAP results in fragments with distinct specificities for caspases. *EMBO J*, 18:5242–5251, 1999.

[180] SJ Riedl, M Renatus, R Schwarzenbacher, Q Zhou, C Sun, SW Fesik, RC Liddington, and GS Salvesen. Structural basis for the inhibition of caspase-3 by xiap. *Cell*, 104:791–800, 2001.

[181] J Chai, E Shiozaki, S M Srinivasula, Q Wu, P Datta, E S Alnemri, and Y Shi. Structural basis of caspase-7 inhibition by xiap. *Cell*, 104:769–780, 2001.

[182] EN Shiozaki, J Chai, DJ Rigotti, SJ Riedl, P Li, SM Srinivasula, ES Alnemri, R Fairman, and Y Shi. Mechanism of XIAP-mediated inhibition of caspase-9. *Mol Cell*, 11:519–527, 2003.

[183] Z Liu, C Sun, E T Olejniczak, R P Meadows, S F Betz, T Oost, J Herrmann, J C Wu, and S W Fesik. Structural basis for binding of smac/diablo to the xiap bir3 domain. *Nature*, 408:1004–1008, 2000.

[184] G Wu, J Chai, T L Suber, J W Wu, C Du, X Wang, and Y Shi. Structural basis of iap recognition by smac/diablo. *Nature*, 408:1008–1012, 2000.

[185] FL Scott, JB Denault, SJ Riedl, H Shin, M Renatus, and GS Salvesen. Xiap inhibits caspase-3 and -7 using two binding sites: evolutionarily conserved mechanism of iaps. *EMBO J*, 24:645–655, 2005.

[186] AD Schimmer, K Welsh, C Pinilla, Z Wang, M Krajewska, MJ Bonneau, IM Pedersen, S Kitada, FL Scott, B Bailly-Maitre, et al. Small-molecule antagonists of apoptosis suppressor XIAP exhibit broad antitumor activity. *Cancer Cell*, 5:25–35, 2004.

[187] Z Wang, M Cuddy, T Samuel, K Welsh, A Schimmer, F Hanaii, R Houghten, C Pinilla, and JC Reed. Cellular, biochemical, and genetic analysis of mechanism of small molecule IAP inhibitors. *J Biol Chem*, 279:48168, 2004.

[188] AP Kater, F Dicker, M Mangiola, K Welsh, R Houghten, J Ostresh, A Nefzi, JC Reed, C Pinilla, and TJ Kipps. Inhibitors of XIAP sensitize CD40-activated chronic lymphocytic leukemia cells to CD95-mediated apoptosis. *Blood*, 106:1742–1748, 2005.

[189] PAW van den Berg, KA Feenstra, AE Mark, HJC Berendsen, and A Visser. Dynamic conformations of flavin adenine dinucleotide: simulated molecular dynamics of the flavin cofactor related to the time-resolved fluorescence characteristics. *J Phys Chem B*, 106:8858–8869, 2002.

[190] C Sun, M Cai, AH Gunasekera, RP Meadows, H Wang, J Chen, H Zhang, W Wu, N Xu, SC Ng, et al. NMR structure and mutagenesis of the inhibitor-of-apoptosis protein XIAP. *Nature*, 401:818–822, 1999.

[191] Hyperchem 6.0.2 Hypercube ig, Florida 32601.

[192] DG Vassylyev, H Tomitori, K Kashiwagi, K Morikawa, and K Igarashi. Crystal structure and mutational analysis of the escherichia coli putrescine receptor. structural basis for substrate specificity. *J Biol Chem*, 273:17604–17609, 1998.

[193] H Wu, J Min, Y Ikeguchi, H Zeng, A Dong, P Loppnau, AE Pegg, and AN Plotnikov. Structure and mechanism of spermidine synthases. *Biochemistry*, 46:8331–8339, 2007.

[194] S Sugiyama, Y Matsuo, K Maenaka, D G Vassylyev, M Matsushima, K Kashiwagi, K Igarashi, and K Morikawa. The 1.8-a x-ray structure of the escherichia coli potd protein complexed with spermidine and the mechanism of polyamine binding. *Protein Sci*, 5:1984–1990, 1996.

[195] H Wu, J Min, H Zeng, DE McCloskey, Y Ikeguchi, P Loppnau, AJ Michael, AE Pegg, and AN Plotnikov. Crystal structure of human spermine synthase: implications of substrate binding and catalytic mechanism. *J Biol Chem*, 283:16135–16146, 2008.

[196] T Hara and T Kimura. Purification and catalytic properties of a cross-linked complex between adrenodoxin reductase and adrenodoxin. *J Biochem*, 105:594–600, 1989.

[197] T Hara and T Miyata. Identification of a cross-linked peptide of a covalent complex between adrenodoxin reductase and adrenodoxin. *J Biochem*, 110:261–266, 1991.

[198] AC Wallace, RA Laskowski, and JM Thornton. Ligplot: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*, 8:127–134, 1995.

[199] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22, 1983.

[200] EJ Bylaska, WA de Jong, K Kowalski, TP Straatsma, M Valievand, et al. Nwchem, a computational chemistry package for parallel computers, version 5.0, Pacific Northwest National Laboratory, Richland, Washington 99352-0999, usa, 2006.

[201] AD Becke. Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. *J Chem Phys*, 107:8554, 1997.

[202] KM Merz. Carbon dioxide binding to human carbonic anhydrase ii. *J Am Chem Soc*, 113:406–411, 1991.

Appendix A

Ligand Binding Modes

All plots were generated by LigPlot [198]. The legend is shown in Figure A.1.

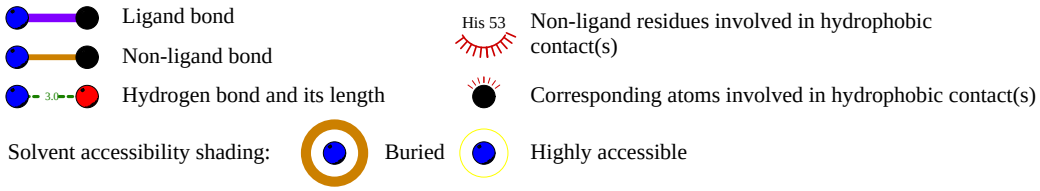
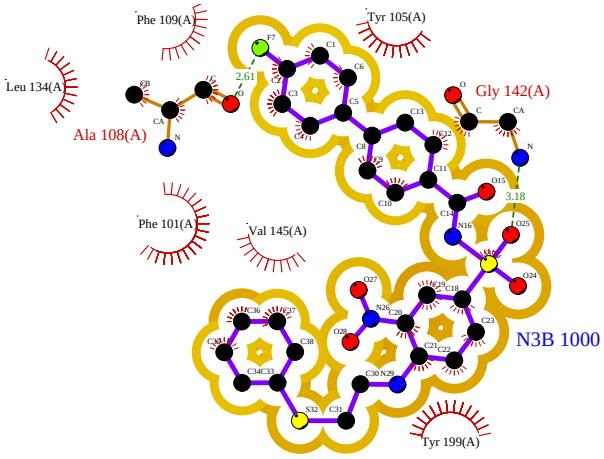
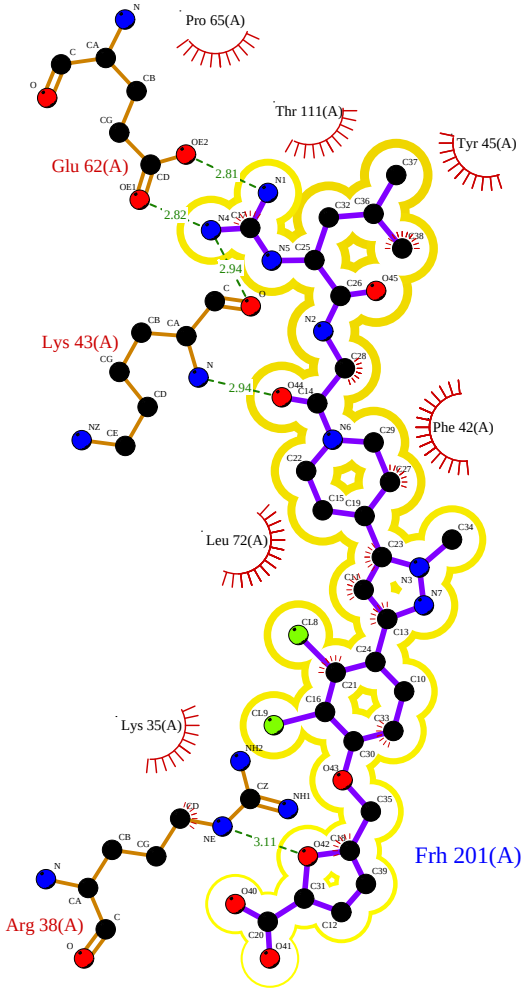


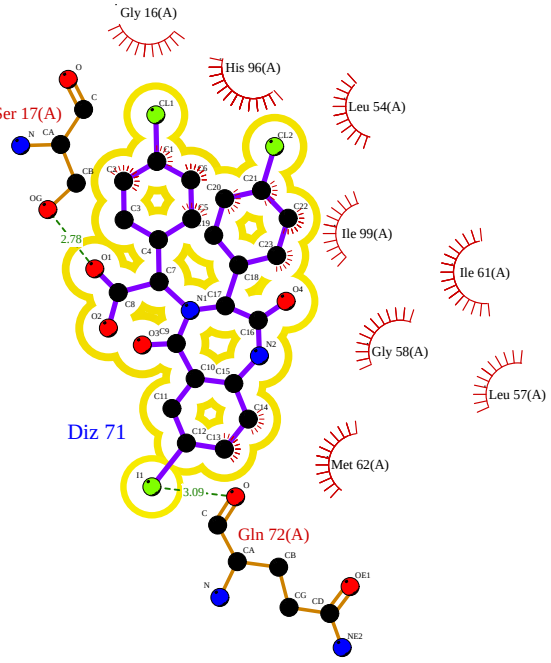
Figure A.1: Legend for the plots showing the ligand binding modes.



(a) BCL-X_L - N3B

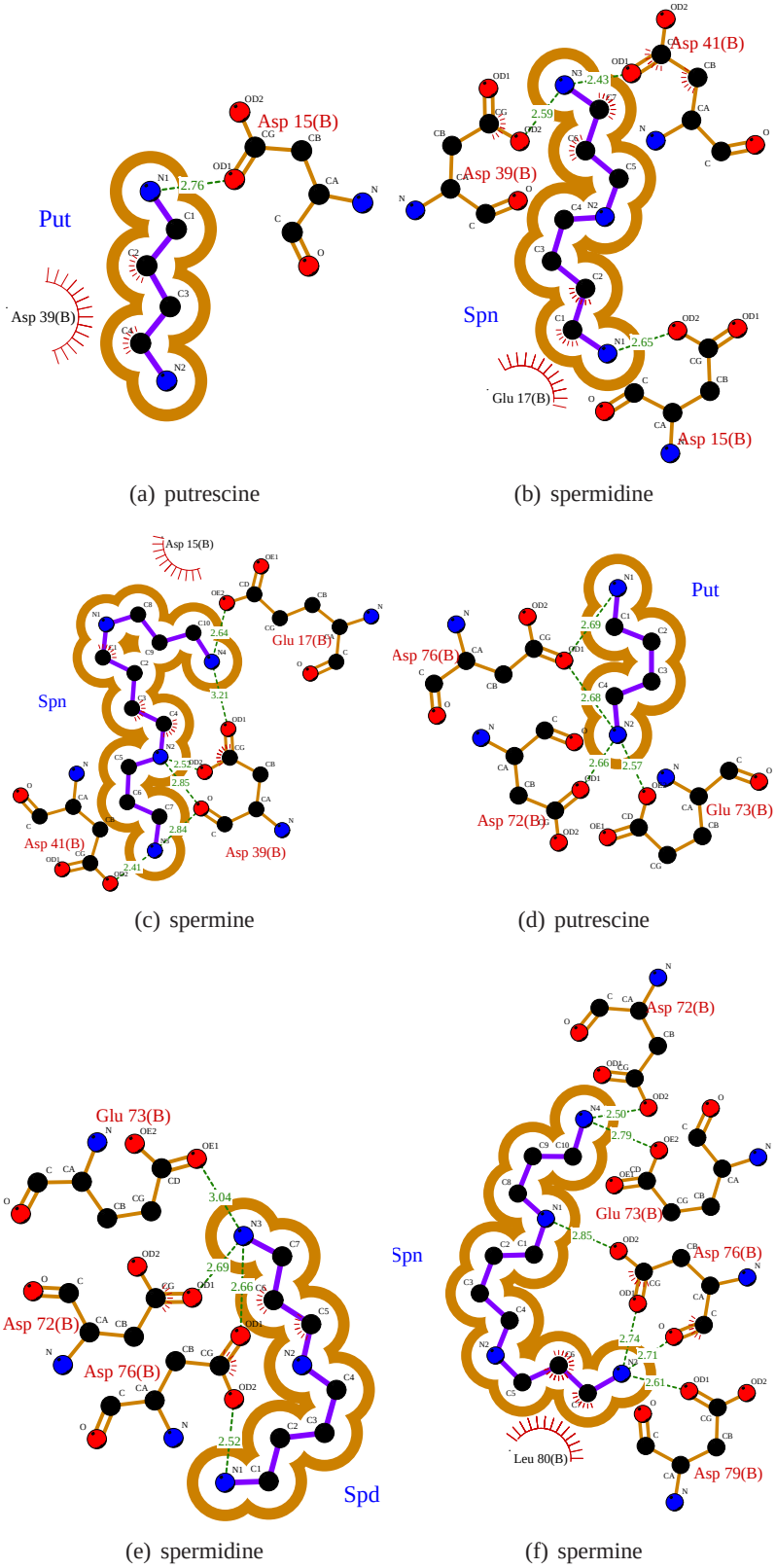


(b) IL-2 - FRH



(c) MDM2 - DIZ

Figure A.2: The native ligand binding modes of the BCL-X_L, IL-2, and MDM2 proteins as resolved in the inhibitor-bound complex structures (a) 1YSI.pdb, (b) 1PY2.pdb, and 1T4E.pdb.

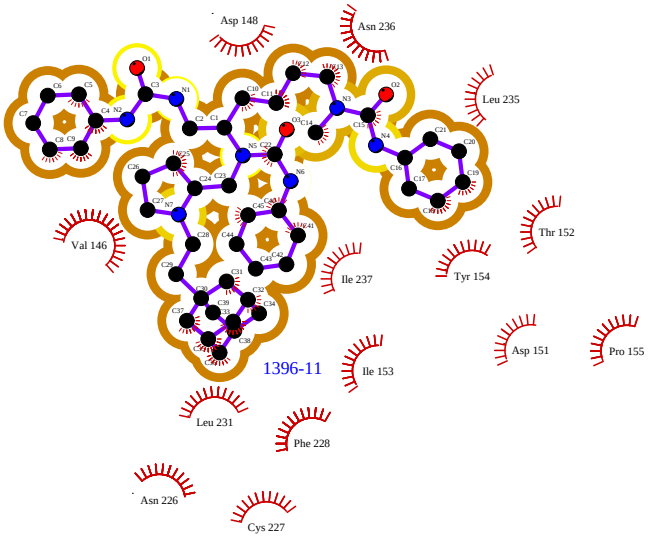


(a) putrescine (b) spermidine

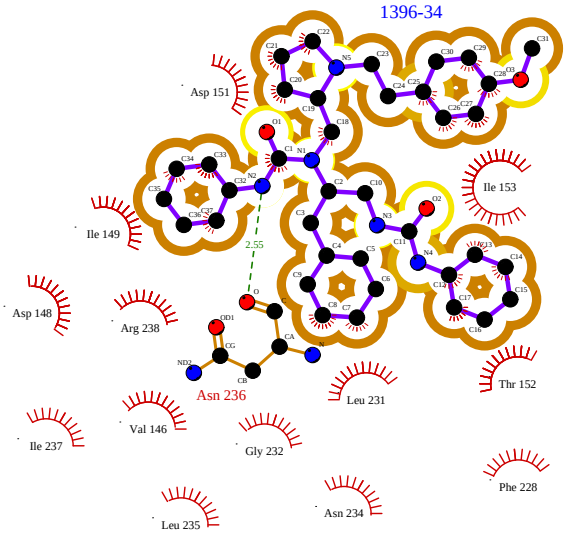
(c) spermine (d) putrescine

(e) spermidine (f) spermine

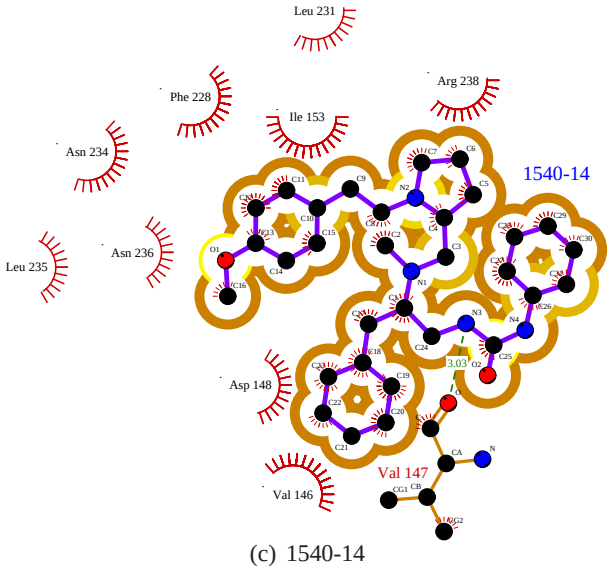
Figure A.3: The best scored docking complex between MD snapshots of Adx and the three polyamines for binding site 1 ((a) - (c)) and 5 ((d) - (f)).



(a) 1396-11



(b) 1396-34



(c) 1540-14

Figure A.4: The best scored docking complex between MD snapshots of XIAP-BIR2 (all from the simulation of the NMR structure in water) and the three inhibitors.

Appendix B

Stability of the Proteins During the Molecular Dynamics Simulations

All secondary structure plots were generated by the program *do_dssp* of the GROMACS 3.3.1 package [141] that reads a MD trajectory and computes the secondary structure for each time frame by calling the DSSP program [199]. The description of the secondary structure elements is taken from the corresponding PBD files. Note that the length of the individual elements may vary in different crystal structures of the same protein. Here, the residues that form that secondary structure element in the apo structure are listed.

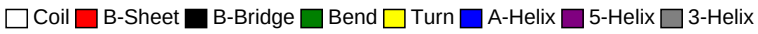
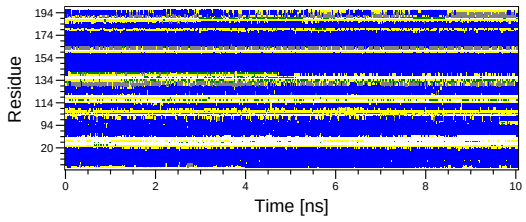


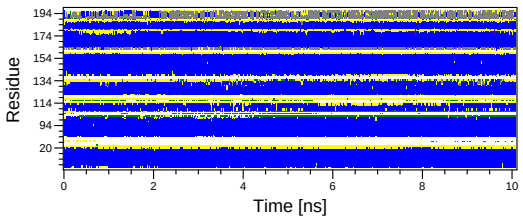
Figure B.1: Legend for the DSSP plots showing the stability of the secondary structures.

B.1 Stability of the Secondary Structure of BCL-X_L

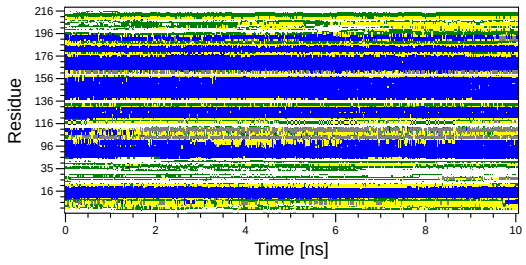
The apo X-ray structure of BCL-X_L consists of eight α-helices (helix 1: residues 1-20, helix 2: residues 82-101, helix 3: residues 105-113, helix 4: residues 119-128, helix 5: residues 136-157,



(a) apo structure in water



(b) apo structure in methanol



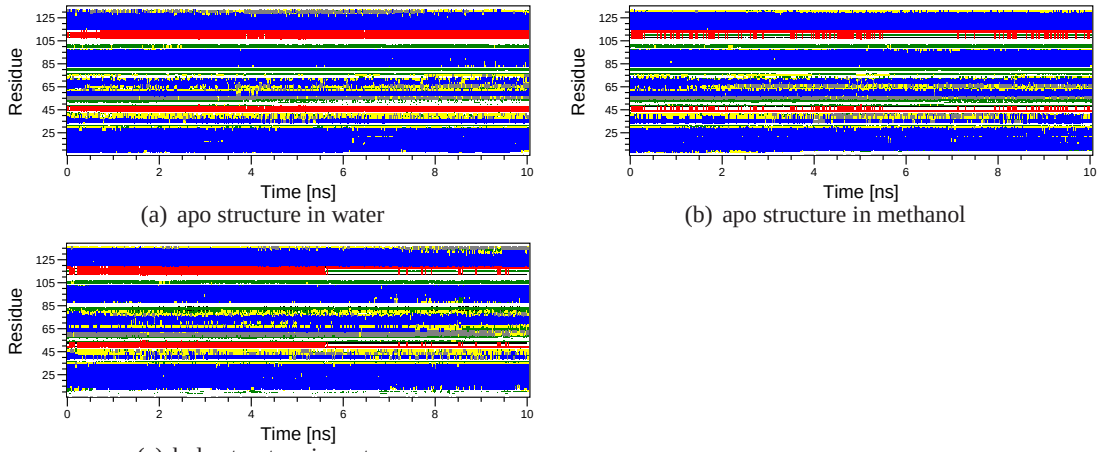
(c) holo structure in water

Figure B.2: The stability of the secondary structure during the MD simulations of (a) apo BCL-X_L in water (run 1) and (b) methanol, and for the simulation of the (c) holo structure in water.

helix 6: residues 161-178, helix 7: residues 178-185, helix 8: residues 187-196) and a 3_{10}-helix (residues 129-132). Figures B.2 (a) and (b) reveal that except for the terminal helix 8 and the 3_{10}-helix that partly unfolded, the secondary structure remained stable during the MD simulations in water and in methanol. The 3_{10}-helix is missing in the holo structure, here residues 129-131 are part of α-helix 4. The secondary structure of this conformation appeared less stable during the simulation in water (see Fig. B.2 (c)). Helix 3, for example, was converted to a π-helix and helix 1 comprised only residues 10-20. However, one should keep in mind that the holo structure contains loop regions (residues 28-44 and 197-217) that were not resolved in the apo X-ray structure and, thus, more structural transitions were observable during the simulation of this structure.

B.2 Stability of the Secondary Structure of IL-2

The apo and holo crystal structures of IL-2 contain both six α-helices (helix 1: residues 6-30, helix 2: residues 32-40, helix 3: residues 56-61, helix 4: residues 62-74, helix 5: residues 81-98, helix 6: residues 113-130) and one 3_{10}-helix (residues 52-55) that remained stable during all simulations as Figures B.3 (a) - (c) indicate. Furthermore, in all simulations two β-sheets comprising residues 44-49 and 107-114 were formed that were not observed in the crystal structures. These sheets were most stable during the simulation of the apo structure in water.



(a) apo structure in water

(b) apo structure in methanol

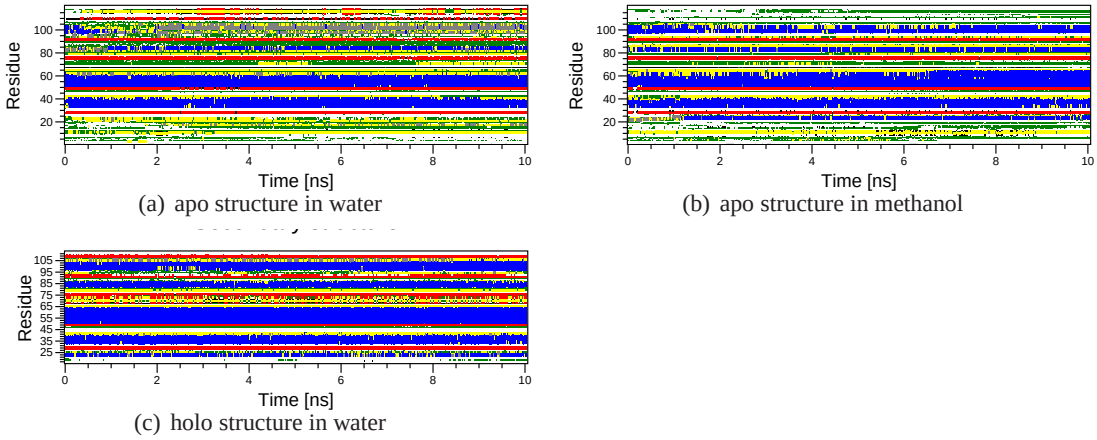
(c) holo structure in water

Figure B.3: The stability of the secondary structure during the MD simulations of (a) apo IL-2 in water (run 1) and (b) methanol, and for the simulation of the (c) holo structure in water.

B.3 Stability of the Secondary Structure of MDM2

MDM2 was the protein for which most changes in the secondary structure took place during the MD simulations. As already indicated by the crystal structures, MDM2 was more stable in the holo form than in the apo form (see also Fig.B.4 (c)). In this conformation, residues 20-25 that are disordered in the NMR models of apo MDM2 form an α-helix. Interestingly, when simulating the apo structure in methanol, this helix is also formed after about 1 ns simulation time (see Fig.B.4 (b)). The protein contains four other α-helices (helix 2: residues 34-42, helix 3: residues 51-63, helix 4: residues 80-86, helix 5: residues 95-104) that were observed in both conformations. However, as Figure B.4 (a) demonstrates, helix 5 formed a π- instead of a α-helix during the simulation of the apo structure in water, and helix 4 was also less stable than during the simulation in methanol. Moreover, the apo structure contains two short β-sheets (sheet 1a: residues 74-75,

sheet 2a: residues 91-92) not contained in the holo structure, where three short β-sheets are formed (sheet 1h: residues 27-30, sheet 2h: residues 48-49, sheet 3h: residues 107-109). β-sheet 3h was only stable in the MD simulation of the holo structure. This sheet fell apart during the simulation of the apo structure in water, while it unfolded completely during the simulation in methanol. In contrast, the DSSP analysis identified the other four β-sheets in all three MD simulations.



(a) apo structure in water (b) apo structure in methanol

(c) holo structure in water

Figure B.4: The stability of the secondary structure during the MD simulations of (a) apo MDM2 in water (run 1) and (b) methanol, and for the simulation of the (c) holo structure in water.

B.4 Stability of Adx

In the X-ray structure of oxidized Adx, the protein consists of five β-sheets (sheet 1: residues 7-12, sheet 2: residues 18-23, sheet 3: residues 56-58, sheet 4: residues 88-90, sheet 5: residues 103-106), three α-helices (helix 1: residues 29-35, helix 2: residues 61-64, helix 3: residues 72-78), and two 3_{10}-helices (helix 4: residues 91-93, helix 5: residues 98-100). Most secondary structure remained conserved during the MD simulation in water, only β-sheets 3 and 4, and the 3_{10}-helix 4 temporarily unfolded. However, the RMS deviation from the crystal structure remained continuously below 2 Å suggesting that the overall protein structure was not distorted.

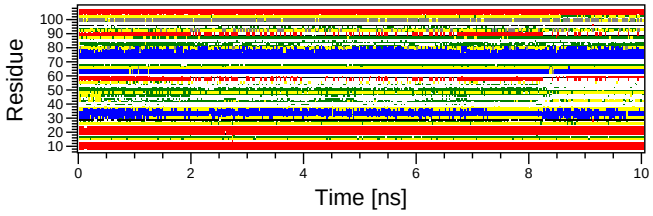


Figure B.5: The stability of the secondary structure during the MD simulation of oxidized Adx in water.

B.5 Stability of the BIR2 Domain of XIAP

The BIR2 domain of XIAP contains in its apo NMR as well as in its complexed X-ray structure three β-sheets (sheet 1: residues 189-194, sheet 2: residues 197-200, sheet 3: residues 205-207), five α-helices (helix 1: residues 136-141, helix 2: residues 162-170, helix 3: residues 180-187,

helix 4: residues 217-224, helix 5: residues 227-233), and a 3_{10}-helix (helix 6: residues 157-161). In the apo structure, residues 125-129 form an additional 3_{10}-helix that is not available in the complex structure because residues 124 to 126 are missing there. Instead, in this structure a 3_{10}-helix is built up by residues 149-153.

The analysis of the MD simulations of XIAP-BIR2 reveals that the zinc finger motif remained very close to its optimized geometry (see section D) and, thus, did not distort the overall protein structure. The RMSD of the backbone atoms and the zinc finger motif is depicted in Figure B.6 (a). Figure B.6 (b) illustrates that, as expected, the N-terminal linker region as well as the C-terminus are highly flexible. The DSSP plots shown in Figure B.7 reveal that, overall, the secondary structure remained stable throughout the simulation. In general, the secondary structure elements were more conserved in the simulations that started from the X-ray structure of the complexed XIAP BIR2 domain. For example, the third β-sheet was only stable in the simulation of this structure in methanol. The five α-helices of this starting structure remained more stable in water. The first and the fifth α-helix was quite unstable in the simulations that started from the apo NMR structure, indicating the high mobility of the flexible linker region and the C-terminus.

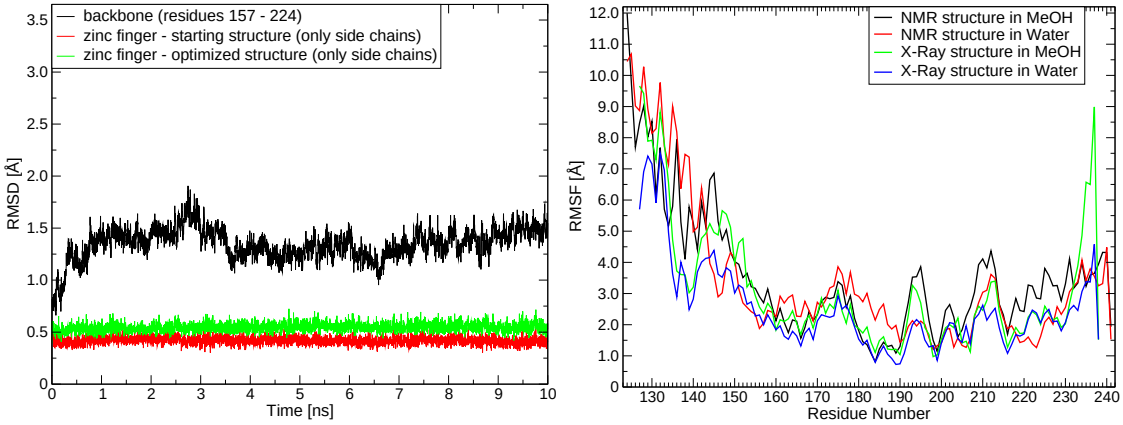
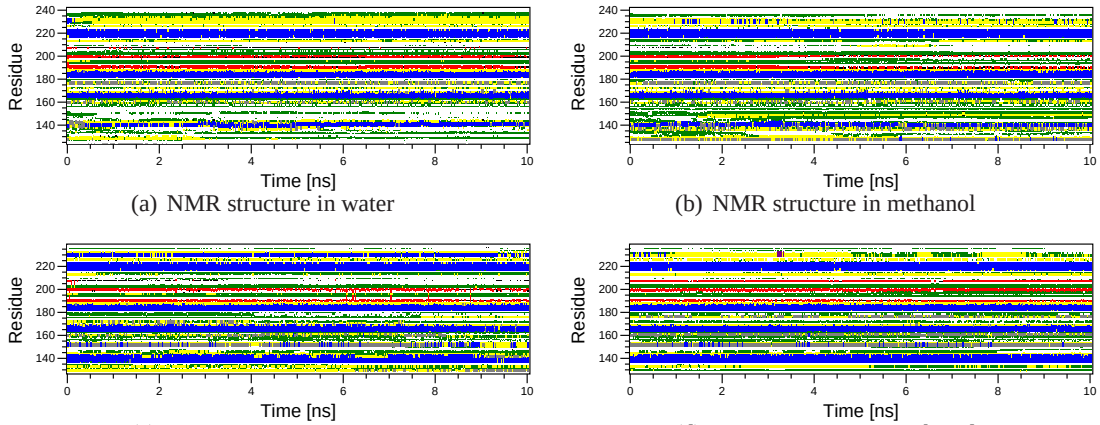


Figure B.6: The stability and the mobility of the protein during the MD simulations. Shown is (a) the RMS deviation of the backbone atoms from the crystal structure and of the zinc finger motif from the crystal structure and the optimized geometry and (b) the mean RMS fluctuations of the Cα-atoms from the crystal structure during the four MD simulations.



(a) NMR structure in water (b) NMR structure in methanol

(c) X-ray structure in water (d) X-ray structure in methanol

Figure B.7: The stability of the secondary structure during the MD simulations of XIAP-BIR2.

Appendix C

User Manuals for the Developed Programs

C.1 EPOS^{BP}: Detecting Ensembles of Pockets on Protein Surfaces

Preparation of the Input Files

All ligands, solvent molecules, and other hetero atoms have to be removed before running EPOS^{BP}. The protein file must be supplied in PDB or HIN format. If a PDB file contains several models, the PASS algorithm is applied to each of them. Ligand coordinate files can be given in MOL2, PDB, or HIN format.

Command Line Options

The following command line options are available:

option	description	required option
-file <PDB/HIN file>	apply the PASS algorithm to a single PDB or HIN file	
-list <file>	apply the PASS algorithm to the PDB or HIN files listed in <file>	
-read <file>	apply the clustering, analysis, or subpocket calculation to previously determined pockets listed in <file>	
-cluster <cutoff> <use index> <cluster file>	cluster pockets with similarities less than <cutoff> percent, write the clustering results and rename the patch and PLA files; only set <use index> to 1 if the atoms have the same index in all files	”-file”, ”-list”, or ”-read”
-readclust <cluster file>	read in a previously calculated cluster file and apply the clustering to the read-in patches	”-read”
-analyze <analysis file>	analyze the pocket properties of the different pocket clusters and write the results to an output file	”-cluster” or ”-readclust”
-subpocket <prefix> <sim cutoff>	write the subpocket (PLAs that are present in at least <sim cutoff> percent of all PLAs) of each pocket cluster to files with the given prefix	”-analyze”
-overlap <ligand file> <overlap file>	calculate the overlap volume between a given ligand (in pdb, hin, or mol2 format) and the patches	”-file”, ”-list”, or ”-read”
-compare <file1> <file2> <sim table>	write the pairwise similarities of the PLAs or subpocket files listed in file1 and file2 to the given output file (format: one file with path per line)	
-v	run program in verbose mode	

PASS Parameter File

If you want to use your own parameters instead of the default values make sure that a parameter file called “BALLPass.ini” is available in your current directory and that the path of the file containing the atom radii is correct.

entry	description	default value
HEAVY_ONLY	ignore hydrogens	1
PARSE_INI_FILE	use parameters defined in local parameter file instead of default values	1
RADIUS_HYDROGEN	radius of a hydrogen atom [Å]	1.2
RADIUS_OXYGEN	radius of an oxygen atom [Å]	1.52
RADIUS_NITROGEN	radius of a nitrogen atom [Å]	1.55
RADIUS_CARBON	radius of a carbon atom [Å]	1.7
RADIUS_SULFUR	radius of a sulfur atom [Å]	1.8
PROBE_SPHERE_RADIUS	radius of a probe in the 1. layer when hydrogens are considered [Å]	1.5
PROBE_SPHERE_RADIUS _HYDROGEN_FREE	radius of a probe in the 1. layer when hydrogens are ignored [Å]	1.8
PROBE_LAYER_RADIUS	radius of a probe in the accretion layers [Å]	0.7
MINIMUM_PROBE_SEPARATION	minimal distance between two probes [Å]	1.0
BURIAL_COUNT_THRESHOLD	minimal number of surrounding protein atoms for defining a probe as buried probe when hydrogens are considered	75
BURIAL_COUNT_THRESHOLD _HYDROGEN_FREE	minimal number of surrounding protein atoms for defining a probe as buried probe when hydrogens are ignored	45
BURIAL_COUNT_RADIUS	radius used for computing the burial counts of a probe [Å]	8.0
PW_SQUARE_WELL	parameter for defining the probe weight envelope function (see [70])	2.0
PW_GAUSSIAN_WIDTH	parameter for defining the probe weight envelope function (see [70])	1.0
ASP_SEPARATION	minimal distance between two ASPs [Å]	8.0
MINIMUM_PROBE_WEIGHT	minimal probe weight for an ASP	1150
CLASH_FACTOR	factor for reducing clashes between probes and protein atoms	0.95
RADII_FILE	file containing the radii of the protein atoms	PARSE.siz

File Formats

The input files have to be of the following format:

”-list <file>”:	(path of) one PDB/ HIN file with path per line
”-read <file>”:	(path of) one patch file with path per line
”-compare <file> ...”:	(path of) one PLAs/ subpocket file with path per line

The generated output files are of the format:

PLAs/ subpocket file:	atoms of the input protein that line the pocket in PDB format
patch file:	Probes are represented by a carbon (initial layer) or hydrogen atoms (accretion layer) in PDB format. The atom name is the atom symbol followed by the layer, the residue name is "PKT", and the residue number corresponds to the pocket ID. The ASP always corresponds to the atom with index 1.
cluster file:	one conversion per line in the format *<file prefix>: <old PID> -> <new PID>*
analysis file:	a header line, followed by one line per PID n the format *<PID> <freq.[%]> <mean vol.[Å^3]> <min. vol.[Å^3]> <max. vol.[Å^3]> <mean pol.> <min. pol.> <max. pol.> <mean depth[Å]> <min. depth[Å]> <max. depth[Å]>*
overlap file:	one line per structure in the format *<file prefix> <overlap vol. [Å^3]> <overlapped ligand atoms [%]> PIDs: <PID 1>...<PID n> <pol.>*. (Note that the overlap volumes and polarities are calculated per structure and not per patch.)
similarity table:	one line per entry in <file 1> of the format *<sim(f1:i,f2:1)> <sim(f1:i,f2:2)> ... <sim(f1:i,f2:m)>* where <sim(f1:i,f2:j)> is the percentage of common PLAs between the ith entry in <file 1> and the jth entry in <file 2>

C.2 PocketScanner and PocketBuilder

File Format

The starting structures have to be in PDB format. Hetero atoms should be removed unless they are correctly parametrized in the CHARMM EEF1 force field. Both programs write the different protein conformations together with the used GPS to files in PDB format. The GPSs are represented by atoms of the residue name "UNK" and their atom name is "Pr", where r is the radius (either 2, 3, or 4 Å). The residue number is arbitrary. As PocketBuilder extracts the GPS directly from the input PDB file, the GPSs can be added manually using this format.

PocketScanner Parameter File

PocketScanner is called by *PocketScanner <parameter file>*. All options are set in the parameter file in the following way:

entry	**description**	**default value**
Structure: <PDB file>	path and filename of the starting structure	
Grid: <x> <y> <z> <no. points> <spacing >	center, dimension, and edge length in Å of the grid placed on the protein surface	
GPS: <GPS radius>	radius of the GPS in Å	2
Minimal BC: <value>	minimal burial count (number of protein atoms with 8 Å) of a GPS before and after energy minimization of the protein	65
SE: <value>	maximal distance of a GPS from a surface exposed atom in Å	2
Outfile: <prefix>	prefix of the generated PDB files	
Force Field: <file>	path and filename of the CHARMM EEF1 force field containing the parameters for the GPSs	

PocketBuilder Parameter File

PocketBuilder is called by *PocketBuilder <parameter file>*. Like in PocketScanner, the parameter file is used to define all options of PocketBuilder. The tuneable parameters are:

entry	description	default value
Structures: <PDB file with GPS 1> ... <PDB file with GPS n>	path and filename of the input structures	
Radius: <value>	radius for defining the flexible residues in Å	8
Number: <value>	number of solutions that should be calculated	50
Pocket Weight: <value>	weighting factor for the interaction energy between protein atoms and the GPS	0.5
Energy Weight: <value>	weighting factor for the internal protein energy	0.5
Outfile: <prefix>	prefix of the generated PDB files	
Dir: <directory>	directory for writing temporary files (needed to save memory)	
Library: <rotamer library>	patch and filename of a rotamer library	
Force Field: <file>	path and filename of the CHARMM EEF1 force field containing the parameters for the GPSs	

C.3 PocketInflator

File Format

The starting structures have to be in PDB format. Hetero atoms should be removed unless they are correctly parametrized in the Amber96 force field. The program is called by *PocketInflator <parameter file>*. The different generated protein conformations are written to files in PDB format. The corresponding patches and plas files are written to files with the same prefix. As PocketInflator uses EPOS^{BP}, a BALLPASS parameter file called "BALLPass.ini" should be available in your current directory. It is recommended to set "ASP_SEPARATION" to 5 Å.

PocketInflator Parameter File

For defining multiple subpockets, one entry of "Resids: <res id 1> ... <resid n>" and "Goal Volume: <value>" has to be provided per subpocket. Note that the ith entry of "Goal Volume" is assigned to the ith entry of "Resids".

entry	description	default value
Structures: <PDB file 1> ... <PDB file n>	path and filename of the input structures	
Number: <value>	number of solutions that should be calculated	50
Outfile: <prefix>	prefix of the generated PDB files	
Force Field: <file>	path and filename of the Amber96 force field containing the parameters for the GPSs	
Stepsize: <value>	step size for increasing the clash factor	0.01
Resids: <res id 1> ... <resid n>	IDs of the residues that should have any atoms within 8 Å of the ASP	
Goal Volume: <value>	goal volume of the induced pocket whose location is defined by the "Resids" entry; use 0 to ignore the volume	0

Appendix D

Parameterization of the Cys_3His-Zinc finger

The parameterization of the Cys_3His-Zinc finger was based either on the energy minimized average NMR structure of the unbound XIAP-BIR2 or on the X-ray structure of XIAP-BIR2 bound to caspase-3. Geometry optimizations were performed using NWChem 4.7 [200]. The ligating cysteines were modeled as CH_3S- and the histidine as imidazole, thus the resulting system had a total charge of -1. The geometries were optimized without constraints by the density function theory (DFT) [201] module using the B3LYP exchange-correlation functional and the 6-31G* basis set. The number of iterations was set to 500 and the default convergence criteria were used for the optimization. The optimized geometry was then used for calculating the electrostatic potential fit (ESP) using the Hartree Fock method with the same basis set. Both input geometries converged to the same minimum energy with an RMSD of 0.8 Å on the heavy atoms and the calculated ESP charges were approximately the same (maximum deviation: 0.018 e). As the optimized geometry based on the X-ray structure was closer to the conformation in either experimental structures than the one based on the NMR structure (0.7 and 0.6 Å instead of 0.8 and 0.9 Å), the former was used for the parameterization of the Cys_3His-Zinc finger in the OPLS-AA force field. The ESP charges obtained from the HF calculation shown in Table D.1 were used for Coulombic interactions. The van der Waals parameters for the zinc ion were taken from [202]. The interactions between the Cys:Sγ or the His:Nϵ_2 and the Zn^{2+} were modeled as bonded interactions and the equilibrium values for bond lengths (Table D.2), angles (Table D.3), and dihedrals (Table D.4) were taken

atom	charge [e]
Zn^{2+}	1.0497
Cys:Cβ	0.2430
Cys:Hβ_1	-0.0591
Cys:Hβ_2	-0.0709
Cys:Sγ	-0.8289
His:Cβ	0.0804
His:Hβ_1	-0.0085
His:Hβ_2	-0.0074
His:Cγ	-0.0339
His:Nδ_1	-0.2319
His:Hδ_1	0.2919
His:Cδ_2	-0.0051
His:Hδ_2	0.1083
His:Cϵ_1	0.0227
His:Hϵ_1	0.1467
His:Nϵ_2	-0.2653

Table D.1: ESP charges calculated for the atoms of the Cys_3His-Zinc finger

from the optimized geometry. The force constants were set in analogy to similar groups in the OPLS-AA force field.

atom 1	atom 2	bond length [Å]
Zn^{2+}	Cys^{200}:Sγ	2.35
Zn^{2+}	Cys^{203}:Sγ	2.34
Zn^{2+}	Cys^{227}:Sγ	2.32
Zn^{2+}	His^{220}:Nϵ_2	2.13

Table D.2: Optimal bond length determined for the Cys_3His-Zinc finger

atom 1	atom 2	atom 3	angle [°]
Cys^{200}:Sγ	Zn^{2+}	Cys^{203}:Sγ	121.4
Cys^{200}:Sγ	Zn^{2+}	Cys^{227}:Sγ	113.1
Cys^{203}:Sγ	Zn^{2+}	Cys^{227}:Sγ	114.8
Cys^{200}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	98.0
Cys^{203}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	96.5
Cys^{227}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	109.2
Zn^{2+}	Cys^{200}:Sγ	Cys^{200}:Cβ	100.8
Zn^{2+}	Cys^{203}:Sγ	Cys^{203}:Cβ	101.7
Zn^{2+}	Cys^{227}:Sγ	Cys^{227}:Cβ	101.4
Zn^{2+}	His^{220}:Nϵ_2	His^{220}:Cϵ_1	122.3
Zn^{2+}	His^{220}:Nϵ_2	His^{220}:Cδ_2	131.0

Table D.3: Optimal angles determined for the Cys_3His-Zinc finger

atom 1	atom 2	atom 3	atom 4	dihedral angle [°]
Zn^{2+}	His^{220}:Nϵ_2	His^{220}:Cδ_2	His^{220}:Cγ	177.6
Zn^{2+}	His^{220}:Nϵ_2	His^{220}:Cϵ_1	His^{220}:Nδ_1	-177.8
Cys^{200}:Cβ	Cys^{200}:Sγ	Zn^{2+}	Cys^{203}:Sγ	28.1
Cys^{200}:Cβ	Cys^{200}:Sγ	Zn^{2+}	Cys^{227}:Sγ	170.6
Cys^{200}:Cβ	Cys^{200}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	-74.6
Cys^{200}:Sγ	Zn^{2+}	Cys^{203}:Sγ	Cys^{203}:Cβ	91.0
Cys^{200}:Sγ	Zn^{2+}	Cys^{227}:Sγ	Cys^{227}:Cβ	166.6
Cys^{200}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	His^{220}:Cϵ_1	-21.2
Cys^{200}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	His^{220}:Cδ_2	161.5
Cys^{203}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	His^{220}:Cϵ_1	-144.2
Cys^{203}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	His^{220}:Cδ_2	38.4
Cys^{203}:Sγ	Zn^{2+}	Cys^{227}:Sγ	Cys^{227}:Cβ	-48.3
Cys^{203}:Cβ	Cys^{203}:Sγ	Zn^{2+}	Cys^{227}:Sγ	-42.5
Cys^{203}:Cβ	Cys^{203}:Sγ	Zn^{2+}	His^{220}:Nϵ_2	-157.1
His^{220}:Cδ_2	His^{220}:Nϵ_2	Zn^{2+}	Cys^{227}:Sγ	-80.7
His^{220}:Cϵ_1	His^{220}:Nϵ_2	Zn^{2+}	Cys^{227}:Sγ	96.7
His^{220}:Nϵ_2	Zn^{2+}	Cys^{227}:Sγ	Cys^{227}:Cβ	58.7

Table D.4: Optimal dihedral angles determined for the Cys_3His-Zinc finger