Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Antonia Stank

born in: Braunschweig, Germany
Oral examination date: 20.06.2017

Computational Studies on the Relation Between Macromolecular Dynamics and Protein Binding and Function

Referees:
Prof. Dr. Rebecca C. Wade
Prof. Dr. Ursula Kummer

Acknowledgement

Many people contributed to the success of this thesis, however I would like to thank some of them in particular.

The biggest thanks go to my supervisor, Prof. Rebecca C. Wade, who gave me the opportunity to gain so many insights into scientific work, and who gave me her trust during the whole of my PhD. I appreciate her continuous support, the time she invested in me, and her promotion of my scientific work.

Secondly, thanks go to my co–supervisor Prof. Ursula Kummer. She assisted me during my PhD and provided feedback that helped to guide and finish my thesis.

The Klaus Tschira Stiftung and Heidelberg Institute for Theoretical Studies (HITS) deserve big thanks, including all the administrative people who helped me with any questions I had.

I would like to thank the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp) at the University of Heidelberg for providing a great network of people, and additional helpful training.

Financial support from the German Federal Ministry of Education and Research (BMBF) for the Virtual Liver Network must also be thankfully mentioned.

In recent years I could always depend on my two mentors, Stefan and Daria, who encouraged me in my work, helped me in any situation, and discussed scientific and also less scientific topics with me. I would like to thank Stefan especially for his patience with me, and for answering so many questions.

Nowadays, scientific work is nearly impossible to do alone, therefore I would like to thank all my collaboration partners, especially Dr. Nadinath B. Nillegoda and Prof. Bernd Bukau for the very fruitful work, and the many discussions that pushed the quality of our joint work.

I met so many great people and friends during my time in the Molecular and Cellular Modeling (MCM) group who all influenced my work. Special thanks go to the previous members Jon, Xiaofeng, Mykhaylo and Musa, with whom I could laugh, discuss any topic, drink coffee or beer and successfully lose in Volleyball matches. Current group members, especially Neil, Mehmet and Mustafa also deserve big thanks for all the fun, help, discussions and insights into different cultures. I enjoyed my time with all of you – thanks a lot!

I would not be who I am today without my parents, family, and friends. So many people influenced my education and interest in natural sciences – thank you all. My parents always encouraged me and enabled my studies, they assisted and believed in me. There is no formula for success, but I would say your love and patience were fundamental to me during my studies. My sisters and paragons, Cornelia and Katharina, I would like to thank for listening, and being there whenever I needed them. Veit, it is impossible to summarize what you have done for me in the last few years. Thanks for being there and loving me.

Abstract

Computational methods can help to better understand and analyze the interaction of proteins and their binding partners. This interaction is influenced by many factors, including specific sequence variants, the dynamics and electrostatics of the proteins, as well as further physicochemical properties of the corresponding binding partners. A detailed investigation of these different, and often complicated, properties helps to better understand the functionality of proteins, for which the interaction with other molecules plays a crucial role.

The work presented here provides new methodologies, implemented in webservers and software, which assist during the analysis of proteins. Furthermore, in an application case, computational methods and analyses in combination with experimental results were used to detect a specific interaction network of proteins.

The new ProSAT$^+$ webserver enables the visualization of protein sequence annotations in the context of the three–dimensional protein structure and contains additional options for visualizing and sharing protein annotations. The sequence information allows an easy, but extensive analysis of proteins. The functionality of the ProSAT$^+$ webserver can be integrated into other webservers, which was done in the case of the two other webservers for the analysis of protein binding pockets described here. A tool for the LigDig webserver was developed that provides the comparison of protein binding pockets by the alignment and visualization of the binding pockets based on an existing algorithm. The new TRAPP webserver assists in the analysis of protein binding pocket dynamics. The existing TRAPP software was used, and a user web interface was implemented to simplify the usability. Additional new functionalities were also developed, such as the visualization of protein sequence conservation in context of all other TRAPP results in the three–dimensional structure. This allows the detection of conserved or non–conserved regions inside the binding pocket, which might influence the dynamics

of the pocket. This newly gained information can be used during the process of designing selective inhibitors.

During the protein disaggregation process, members from different classes of the so-called J–protein (HSP40) co–chaperones play a crucial role. The synergetic application of different computational methods and experiments enabled the detection of an interclass specific J–protein interaction and indicated that the interaction evolved to enable a high efficiency in the disaggregation process. The resulting data of performed protein domain docking simulations required an update of the standard clustering workflow. This new methodology can be applied for protein docking in cases that have problems with multiple, weakly specific interaction sites.

The work presented here facilitates in many ways the analysis of proteins, including their structure and sequence features, as well as, their dynamics and interactions with their binding partners. The new methods are provided as webservers and therefore are accessible, and easy to use for all researchers. This can assist in many research projects and provide relevant information. The analyses of the J–proteins improved the knowledge about their biological role and functionality, and therefore provide an important contribution for a better understanding of the overall protein disaggregation process.

Zusammenfassung

Computer gestützte Methoden können dabei helfen die Interaktionen von Proteinen und deren Bindepartner besser zu verstehen und zu analysieren. Diese Interaktion wird von vielen Faktoren beeinflusst, unter anderem von spezifischen Sequenzvariationen, der Dynamik und Elektrostatik der Proteine, sowie weiteren physikochemischen Eigenschaften der jeweiligen Bindepartner. Eine detaillierte Untersuchung dieser verschiedenen, oft komplizierten Eigenschaften, hilft die Funktionsweisen von Proteinen, für welche die Interaktion mit anderen Molekülen eine entscheidende Rolle spielt, zu entschlüsseln.

Die hier präsentierte Arbeit liefert neue Methoden, unter anderem in Form von Webservern, die die Analyse von Proteinen unterstützen. Außerdem wird an einem Anwendungsbeispiel gezeigt, dass das Zusammenspiel von computergestützten Simulationen und Analysen, sowie experimentellen Ergebnissen zur Entschlüsselung eines spezifischen Interaktionsnetzwerkes von Proteinen genutzt werden kann.

Der neu entwickelte ProSAT$^+$ Webserver ermöglicht die Visualisierung von Proteinsequenzannotationen innerhalb der dreidimensionalen Proteinstruktur und verfügt über weitere Optionen zur Visualisierung und Weiterleitung von Proteinannotationen. Die verschiedenen Sequenzinformationen erlauben eine einfache, aber umfangreiche Analyse von Proteinen. Die Funktionalität des ProSAT$^+$ Webservers kann auf Grund seiner Architektur in andere Webserver integriert werden, was bei den zwei weiteren, mitentwickelten Webservern zur Untersuchung von Proteinbindetaschen getan wurde. Für den LigDig Webserver wurde ein zusätzliches Modul für den Vergleich von Proteinbindetaschen durch eine Superpositionierung und Visualisierung der Bindetaschen, basierend auf einem bereits vorhandenen Algorithmus, entwickelt. Der neue TRAPP Webserver unterstützt bei der Untersuchung von Proteinbindetaschendynamiken. Hierfür wurde die bereits vorhan-

dene TRAPP Software verwendet und die Anwendung mit Hilfe einer Weboberfläche vereinfacht. Weitere Funktionalitäten, wie die Visualisierung der Konserviertheit der Proteinsequenz wurden neu entwickelt. Dies ermöglicht die Untersuchung von konservierten, beziehungsweise nicht konservierten Bereichen innerhalb der Proteinbindetasche, welche möglicherweise die Dynamik der Tasche beeinflussen und dessen Information bei der Entwicklung von selektiven Inhibitoren zur Anwendung kommen kann.

Innerhalb des Proteindisaggregationsprozesses spielen verschiedene Klassen der sogenannten J–Protein (HSP40) co–Chaperone eine wichtige Rolle. Die Zusammenarbeit von verschiedenen computergestützten Methoden und Experimenten ermöglichte die Entschlüsselung einer Interaktion von J–Proteinen aus verschiedenen Klassen und gab den Hinweis auf eine evolutionär bedingte Interaktion, welche entscheidend für die Effizienz des Disaggregationsprozesses ist. Die hierbei durchgeführten Docking Simulationen mit Proteindomänen lieferten Ergebnisdaten, welche eine Überarbeitung des standard Clusteringverfahrens notwendig machten. Das neue Verfahren kann im Fall von Protein Docking Simulationen mit mehreren, weniger spezifischen Interaktionsstellen zur Anwendung kommen.

Die hier präsentierte Arbeit erleichtert in vielerlei Hinsicht die Analyse von Proteinen, deren Strukturen und Sequenzeigenschaften, sowie deren Dynamik und Interaktionen mit anderen Bindepartnern. Die neuen Funktionalitäten stehen auf Grund Ihrer Verfügbarkeit als Webserver und der vereinfachten Bedienung allen Wissenschaftlern zur Verfügung und kann die Arbeit vieler Forschungsprojekte vereinfachen und relevante Informationen liefern. Die durchgeführten Analysen der J–Proteine hat dazu beigetragen ihre Rolle und Funktionalität besser zu verstehen und lieferte daher einen wichtigen Beitrag zum besseren Verständnis des gesamten Proteindisaggregationsprozesses.

Publications

1. **Antonia Stank**, Stefan Richter, Rebecca C. Wade. ProSAT^{+}: visualizing sequence annotations on 3D structure. Protein Engineering, Design and Selection, (2016) 29:281-284

2. **Antonia Stank**, Daria B. Kokh, Jonathan C. Fuller and Rebecca C. Wade. Protein Binding Pocket Dynamics. Account of Chemical Research, (2016) 49:809-815

3. Nadinath B. Nillegoda, Janine Kirstein, Anna Szlachcic, Mykhaylo Berynskyy, **Antonia Stank**, Florian Stengel, Kristin Arnsburg, Xuechao Gao, Annika Scior, Ruedi Aebersold, D. Lys Guilbride, Rebecca C. Wade, Richard I. Morimoto, Matthias P. Mayer and Bernd Bukau. Crucial HSP70 co-chaperone complex unlocks metazoan protein disaggregation. Nature, (2015) 524:247-251

4. Jonathan C. Fuller, Michael Martinez, Stefan Henrich, **Antonia Stank**, Stefan Richter and Rebecca C. Wade. LigDig: a web server for querying ligand-protein interactions. Bioinformatics, (2015) 31:1147-1149

5. **Antonia Stank**, Daria B. Kokh, Max Horn, Elena Sizikova, Rebecca Neil, Joanna Panecka, Stefan Richter and Rebecca C. Wade. TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets. (2017) Submitted.

6. Nadinath B. Nillegoda, **Antonia Stank**, Duccio Malinverni, Niels Alberts, Anna Szlachcic, Alessandro Barducci, Paolo De Los Rios, Rebecca C. Wade and Bernd Bukau. Evolution of an intricate J-protein network driving protein disaggregation by the Hsp70 chaperone machinery. (2017) Submitted.

Contents

Abbreviations

AC	Accession Number
APBS	Adaptive Poisson–Boltzmann Solver
BCL–XL	B–cell lymphoma–extra–large
BD	Brownian Dynamics
BLAST	Basic Local Alignment Search Tool
CADD	Computer Aided Drug Design
CASP	Critical Assessment of protein Structure Prediction
CDD	Conserved Domain Database
COG	Center of Geometry
CSS	Cascading Style Sheets
CTD	C–terminal domain
DHFR	Dihydrofolate Reductase
FRET	Förster Resonance Energy Transfer
HSP	Heat Shock Protein
IL–2	Interleukin 2
JD	J–domain
JS	JavaScript
LPBE	linearized Poisson–Boltzmann equation
MD	Molecular Dynamics
MSA	Multiple Sequence Alignment
MVC	Model–View–Controller
NMA	Normal Mode Analysis
NMR	Nuclear Magnetic Resonance
ns–LTP	nonspecific Lipid Transfer Protein
PBE	Poisson–Boltzmann equation
PDB	Protein Data Bank

PDBe	Protein Data Bank in Europe
PIPSA	Protein Interaction Property Similarity Analysis
PMP	Protein Model Portal
ProBiS	Protein Binding Site
ProSAT	Protein Structure Annotation Tool
P38 MAPK	P38 Mitogen–Activated Protein Kinase
RCSB	Research Collaboratory for Structural Bioinformatics
REST	REpresentational State Transfer
RMSD	Root Mean Square Deviation
SDA	Simulation of Diffusional Association
SI	Similarity Index
SIFTS DB	Structure Integration with Function, Taxonomy and Sequence database
SM	SWISS–MODEL
TRAPP	TRAnsient binding Pockets in Proteins
UHBD	University of Houston Brownian Dynamics
UIM	Ubiquitin–Interacting Motifs
UniProt	Universal Protein Database
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
VMD	Visual Molecular Dynamics
WebGL	Web Graphics Library
ZFLR	Zinc–Finger–Like–Region

1

Introduction and Theoretical Basis

Sections of this chapter are based on the following publication, the text was initially written by me and went through the review process of all authors:

Antonia Stank, Daria B. Kokh, Jonathan C. Fuller and Rebecca C. Wade. Protein Binding Pocket Dynamics. Account of Chemical Research, (2016) 49:809-815

1.1 Introduction

Proteins are one of the most important components in the cellular environment. They are responsible for many functions, and interact in a huge network with other proteins and macromolecules. All proteins are influenced by evolution, which affects all kinds of protein properties and therefore the big molecular symphony in living organisms. Small changes in the genes or the environmental conditions can change protein behavior and therefore have a big influence on the balanced system. The complex interplay of molecules can depend on the physicochemical feature of only one amino acid.

In recent decades, the development of computational methods enabled, besides many other things, the analysis of evolutionary relationships of proteins, and the simulation of molecular interactions. The development and application of computational methodologies to analyze the complex molecular interactions, the influence on functionality, and the relation to evolution, is the main focus of this work.

Protein information at the sequence level can already help to understand evolutionary relationships, or to find by mutagenesis identified functionally relevant amino acids. Nevertheless, merging the one–dimensional information with the three–dimensional protein structure enlarges the overall information content. One focus of this thesis is the development of a webserver that can visualize protein sequence annotations on protein structures, in an easy and user–friendly way. This data assembly allows for new discoveries in the field of protein functionality. As soon as the behavior and functionality of a protein is analyzed in detail it is possible to change this by specific treatments, for example drugs.

The highly complex binding mechanism between a protein and its binding partner (e.g. a ligand or another protein) takes place at an interaction site. Such interaction site can be a protein binding pocket, which is more frequently used for binding smaller molecules and drugs. The analysis of such binding pockets, especially regarding their dynamics, is a very important aspect in the field of computational drug design. A new classification of protein binding pockets regarding their dynamics is introduced in Chapter 3. A new webserver that can analyze structural changes in protein binding pockets, and the influence on the whole pocket, is also introduced in this thesis. An additional feature to visualize the conservation of evolutionary related proteins, together with their protein pocket dynamics, will be described, and explained with an example.

A biologically relevant application case of a protein–protein interaction analysis that includes protein structure modeling and analysis, interaction simulations, and the evolutionary analysis of protein functionality was done with the DnaJ heat shock proteins (HSP40). These proteins are chaperone molecules, and play an important role in the protein disaggregation and refolding processes of other proteins. The methodological details and challenges, as well as, the overall results and new discoveries, are explained and discussed.

Several theoretical methodologies and algorithms are applied in this thesis, which require explanation beforehand. This chapter serves as background information for the most relevant ones. After a short introduction into the world of proteins, it starts with some computational biology methods, and algorithms in the field of protein sequence analysis, followed by the problem of protein structure prediction and a common approach based on homology modeling. Afterwards, some background information about protein binding pocket analysis, including their detection, definition, and comparison is given. The theoretical background

of molecular interactions, including electrostatic potential calculations, provide the relevant information for the Protein Interaction Property Similarity Analysis (PIPSA) tool. The following brief introduction to Brownian dynamics provides the background information for the Simulation of Diffusional Association (SDA) software, which is explained as it is applied in Chapter 4. At the end of this chapter, a short introduction into the field of clustering will be given, to highlight the necessity to adapt the clustering procedure for SDA simulation results, which is explained in Chapter 4. The chapter finishes with an introduction to webservers and an applied webserver framework.

1.2 Proteins

Proteins represent a major component of living organisms and more than 25% by weight of a typical living organism [1]. They are biological macromolecules that are able to undertake an impressive variety of functions. The different three–dimensional shapes enable proteins to act, for example, as structural proteins, enzymes, antibodies, regulatory proteins, sensors, transporters and transducers [2].

1.2.1 From Sequence to Structure

The function of a protein is determined by its structure, and the fundamental building block of proteins are amino acids. The particular sequence of amino acids in the polypeptide chain influences the overall fold. There are 20 types of natural amino acids found in proteins, which can be distinguished by the nature of their side chain groups. Figure 1.1 shows the chemical structure of an amino acid, where R marks the position of the individual side chain. A polypeptide chain of amino acids is built via peptide bonds between two amino acids by removal of an OH from one, and an H from the next amino acid, as shown in Figure 1.2. Therefore, the two ends of a protein are called N– and C–terminus because the first has a unlinked NH_3^+ and the the last an unlinked COO^- group.

The 20 amino acids vary considerable in their properties. Table 1.1 lists the side chains of all natural occurring amino acids, their characteristics and names. The protein sequence is often represented by a list of residue names represented by a 1–letter code, for example, MGKDYYQTLGLARGASDEEIKRAY is the beginning of the human DNAJB1 J–protein (Uniprot accession: P25685).

Polar or charged amino acids can participate in hydrogen bonding and electrostatic interactions with other amino acid residues and with solvent. Non–polar side chains interact unfavourably with water, which is called the hydrophobic effect. Those amino acids are often located together in the interior of a protein [2].

The spontaneous folding of the polypeptide chain into a three–dimensional structure mainly depends on the amino acid sequence, which itself is encoded in the genes. The Central Dogma of Biology, stated by Francis Crick in 1958, describes the relation between genes and proteins, and can be summed up with the idea that DNA makes RNA, which then makes the protein [3]. Therefore, changes in the nucleic acid sequence in the genes (e.g. by mutation) can influence the protein sequence, its three–dimensional structure and possibly also its function. A group of proteins, called chaperone molecules, can assist proteins during their folding process.

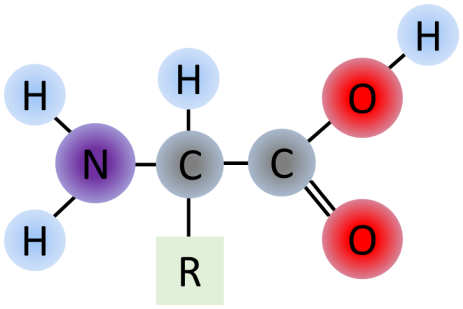

Figure 1.1: Chemical structure of an amino acid. The R marks the position of the individual amino acid side chain.

The protein architecture can be divided into four different structural levels: primary, secondary, tertiary and quaternary structure. An overview of these four levels is shown in Figure 1.3. The primary structure (Figure 1.3a) describes the amino acid sequence of a protein. The polypeptide chain is flexible and its local conformation is described by the secondary structure. Limitations in the angles of the peptide bonds lead to some preferred conformations: alpha–helices (Figure 1.3b) and beta–sheets. These two types of secondary structure occur because of hydrogen bonds between side chains and lead to a more packed conformation of the polypeptide chain. The alpha–helix is a local structure that depends on a set of consecutive residues in the amino acid sequence. However, beta–sheets form

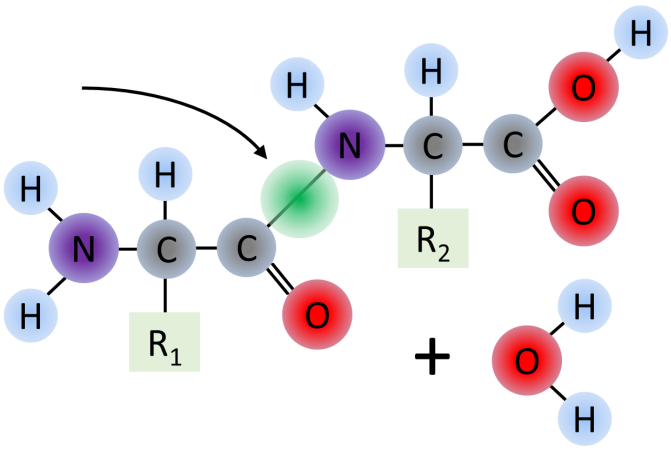

Figure 1.2: Peptide bond between two amino acids. The peptide bond is highlighted in green and marked by an arrow.

Table 1.1: List of the 20 naturally occurring amino acids and their physicochemical features.

Amino Acids (3-, 1-Letter Code)	Side-Chain Polarity	Side–Chain Charge at pH 7	Side–Chain (R)
Alanine (Ala, A)	nonpolar	neutral	$-CH_3$
Arginine (Arg, R)	basic polar	positive	$-CH_2CH_2CH_2NH-C(NH)NH_2$
Asparagine (Asn, N)	polar	neutral	$-CH_2CONH_2$
Aspartic acid (Asp, D)	acidic polar	negative	$-CH_2COOH$
Cysteine (Cys, C)	nonpolar	neutral	$-CH_2SH$
Glutamic acid (Glu, E)	acidic polar	negative	$-CH_2CH_2COOH$
Glutamine (Gln, Q)	polar	neutral	$-CH_2CH_2CONH_2$
Glycine (Gly, G)	nonpolar	neutral	$-H$
Histidine (His, H)	basic polar	positive/neutral	$-CH_2(C_3H_3N_2)$
Isoleucine (Ile,I)	nonpolar	neutral	$-CH(CH_3)CH_2CH_3$
Leucine (Leu,L)	nonpolar	neutral	$-CH_2CH(CH_3)_2$
Lysine (Lys, K)	basic polar	positive	$-CH_2CH_2CH_2CH_2NH_2$
Methionine (Met, M)	nonpolar	neutral	$-CH_2CH_2SCH_3$
Phenylalanine (Phe, F)	nonpolar	neutral	$-CH_2(C_6H_5)$
Proline (Pro, P)	nonpolar	neutral	$-CH_2CH_2CH_2-$
Serine (Ser, S)	polar	neutral	$-CH_2OH$
Threonine (Thr, T)	polar	neutral	$-CH(OH)CH_3$
Tryptophan (Trp, W)	nonpolar	neutral	$-CH_2(C_8H_6N)$
Tyrosine (Tyr, Y)	polar	neutral	$-CH_2(C_6H_4)OH$
Valine (Val, V)	nonpolar	neutral	$-CH(CH_3)_2$

by lateral interactions, and depend on sets of residues that can be distant in the amino acid sequence.

The spatial arrangement of the secondary structure elements, and their interaction patterns, are called tertiary structure, or protein fold (Figure 1.3c). The

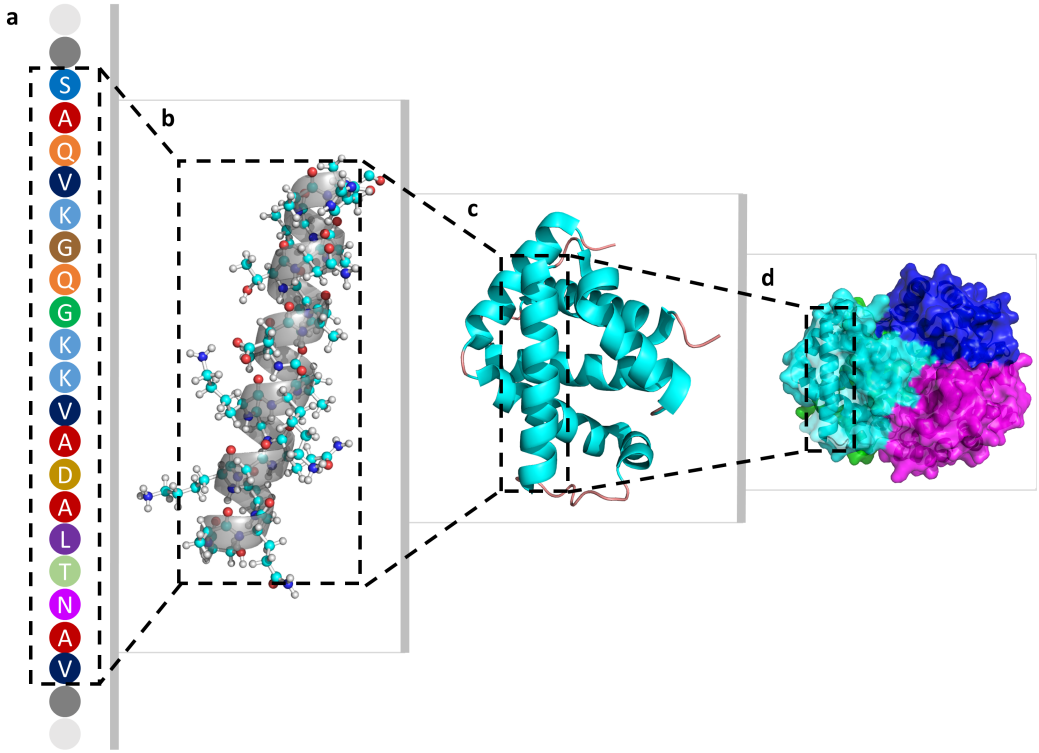

Figure 1.3: The four levels of protein architecture. (a) The primary structure describes the amino acid sequence of a protein. (b) The secondary structure characterizes the local three–dimensional sub–structures of a polypeptide backbone chain. The most common secondary structure conformations are alpha–helix (shown in figure) and beta–sheet. (c) The tertiary structure is about the spatial arrangement of the secondary structure elements of one polypeptide chain. (d) Quaternary structure describes the three–dimensional complex of a multi–subunit protein.

tertiary structure shows how a polypeptide chain can lead to versatile structured proteins with various functionalities and roles.

Proteins based on one polypeptide chain or subunit are called monomers. Many proteins contain multiple subunits, which can be the same or different monomers. The arrangement of those subunits to one complex is called the quaternary structure (Figure 1.3d). Complexes of multiple monomers of the same subunit are called homodimers, homotetramers and so forth. In case of different monomers the complexes are called heterodimers and heterotetramers, respectively.

Beside these four architectural levels, there are other descriptions of parts of the protein. One is the protein domain, which describes a conserved part of a protein that evolves, functions, and exists independently of the rest of the protein chain. Those domains often serve as functional or structural units, and can be

found across different proteins. A collection of known protein domains and functional units can be found in the Conserved Domain Database (CDD) [4], together with further annotations that provide insights into their sequence, structure, and function relationships.

Beside the specific composition of amino acids, there are other additional components in protein structures that influence the overall behavior of proteins. Those components, such as ions, water molecules and small organic ligands, can be an integral part of the three–dimensional structures, and relevant for interactions with other molecules.

The protein folding process is a complex mechanism, and at present it is not possible to simulate this overall process in detail with computational methods. Nevertheless, there are methods that can predict the secondary structure of proteins, for example PSIPRED [5] and JPred [6, 7]. For a summary about secondary structure prediction and available tools, see [8, 9]. The predicted secondary structures can be identified with the DSSP tool [10], when the three–dimensional structure is available. DSSP uses three–dimensional structures to detect hydrogen–bonds and geometric patterns, and translate these into protein secondary structures.

In Section 1.4, the prediction of three–dimensional protein structures with the homology modeling method is described. For further insights into protein synthesis, the protein structure, and the protein folding process, please read [11].

1.2.2 Features

The amino acid sequence and the folded three–dimensional structure are two, out of many, protein characteristics. Both of them influence other comparable features, such as the amino acid composition, protein weight and volume, and the size of the protein. Beside these, there are other more complex features, for example, the extinction coefficient, estimated half–life, instability index, aliphatic index and the hydropathicity. All of them depend mainly on the amino acid sequence and the environmental condition in which the protein is located.

One feature, which is of particular interest in this thesis is the electrostatic potential of proteins. It influences the electrostatic interactions of proteins and plays a crucial role in many biomolecular processes, including molecular recognition and binding. Details of how the electrostatic potential of a protein can be calculated, and which software is available are mentioned in Section 1.6.

All these mentioned features influence the dynamics and the functionality of a protein and therefore their interactions with other molecules.

1.2.3 Molecular Evolution

The highly specific functionality of proteins is the result of long term molecular evolutionary processes. The process of evolution is based on two factors that are essential for it to occur at the molecular level. First, genes encoding for molecules can be copied, which is relevant in the replication process of the organism. Nevertheless, this copying mechanism is not perfect and the copies of the genes and DNA are not identical to the template. These errors, as long as they occur not too frequently and always are non–critical, are necessary to introduce variation. Second, after replication the variation has to cause a change that increases the chance of the organism's survival, resulting in natural selection, because of a higher fitness [12]. Nevertheless, based on sequence analyses, it was found that the rate of protein evolution is predominantly influenced by its expression level rather than functional importance [13].

The mentioned errors are changes in the genes that can be passed on to offspring and are called mutations. Different types of mutation exists, for example, point mutations and insertions or deletions. The mutation in a gene can change the encoded amino acid, which then can influence different features of the synthesized molecule. Nevertheless, the relationship between amino acid sequence and protein structure is quiet robust, which allows to some degree the prediction of structure based on protein sequence. For this purpose, the comparison of sequences at the amino acid level is necessary.

1.3 Protein Sequence Analysis

The sequences of amino acids in naturally occurring proteins have been through an evolutionary selection procedure. This led to proteins with favorable properties and functions. These functions are often conserved between species, even if the sequences are not similar. Nevertheless, comparing the sequences of evolutionarily related proteins from different species can highlight conserved amino acids. These amino acids are highly likely to be involved in the functional mechanism of the protein. To identify those conserved amino acids, the protein sequences have to be aligned correctly and the right residues have to be compared with each

other. In the following, two algorithms essential in the field of sequence analysis are described in more detail. These algorithms are also frequently used in different ProSAT$^+$ background processes (see Chapter 2), by applying already existing software libraries.

1.3.1 Sequence Alignments

Protein sequence alignments are used to arrange sequences to enable the identification of similar sequential regions that can indicate structural, functional and/or evolutionary relationships. The alignment between two protein sequences can be done globally or locally. A global alignment helps to find the optimal correspondence between all amino acids of both sequences. Whereas the local alignment identifies local regions with high similarity between both sequences. The local alignment is biologically more relevant as it can help to identify evolutionary related conserved regions, even if the remaining parts of the proteins are different. These regions, for example protein domains, can contain functional units, which were subject to evolutionary pressures.

The Needleman and Wunsch algorithm was introduced in 1970 [14] and provides a way to generate a global pairwise sequence alignment of two proteins, which is a correspondence between the amino acids and appropriately inserted gaps in both sequences. The algorithm can be easily adapted to obtain a local sequence alignment, which does not necessarily take the whole length of both sequences into account. This adapted version is called the Smith and Waterman algorithm and was introduced in 1981 [15].

Often it is required to increase the information content of a sequence alignment by adding further sequences. This is done in a multiple–sequence alignment (MSA), which allows the comparison of N protein sequences. This also helps to identify evolutionary conserved regions by taking protein sequence information from several species into account.

1.3.2 BLAST

Even if the dynamic programming routines for sequence alignments are already optimized and very efficient, there is a need for an algorithm that is able to handle

the amount of data in the rapidly growing sequence databases. The search time for a specific and/or similar sequence should not increase with the database size.

In 1990 the program BLAST (Basic Local Alignment Search Tool) was introduced by Altschul *et al.* [16] and is by now one of the most popular tools and most cited papers in the field of Bioinformatics with more than 60,000 citations (source Google Scholar). This heuristic algorithm tries to find the highest–scoring ungapped local alignment between the query and the database sequence. The basic idea of the BLAST algorithm is to scan a database to find short words (often a length of 3) with a similarity score above a certain threshold when they are aligned to words in the query sequence. Those hits are then extended until the similarity is much worse than the best hit so far. This basic algorithm has been adapted to many variants that can be applied for specific sequence types and certain tasks. For more details and a detailed explanation of the algorithm itself, see [16] and [12].

1.4 Homology Modeling

Although the number of resolved structures is increasing steadily, there is still a huge gap between the number of sequenced proteins and the number of proteins with resolved three dimensional structures. This issue can be addressed by applying the homology modeling approach, also called comparative modeling. The fundamental idea of homology modeling is that protein structures are quite conserved for proteins with at least some degree of protein sequence similarity. If proteins descend from a common ancestor and, for example, were part of a divergence process they are called homologous or homologous proteins.

Besides homology modeling, other methods for protein structure prediction exist, and have improved in recent years. Nevertheless, homology modeling still produces the best results if an appropriate template structure is available. In case of incomplete template structures or even if no similar structures are available other, less precise methods, can be used. Those methods can be based on the idea of threading or even perform *ab initio* structure prediction. See Pavlopoulou and Michalopoulos [17] for a comparison and summary of available protein structure prediction tools and methodologies.

The Critical Assessment of protein Structure Prediction (CASP) (www.predictioncenter.org) experiments are aiming to establish the current state of the art in

protein structure prediction. The CASP experiments started in 1994 with the aim to challenge current protein structure prediction methods to identify what progress has been made and to show where future effort should be focused.

1.4.1 General Methodology

Even if many homology modeling methods exist, they often follow a common workflow. In the beginning, only the sequence of the protein that should be modeled is known. This sequence is used to find and select a good template structure via, for example, a BLAST search and a sequence alignment. This template structure is then used to build the model, often by using the backbone structure and side chain modeling. For the side chain arrangement, some tools also take advantage of side chain rotamer libraries. In case several appropriate template structures are available, multiple models can be generated and need to be benchmarked afterwards, often done with a structure quality score. Finally, the three–dimensional model structures are globally optimized.

1.4.2 SWISS–MODEL

The SWISS–MODEL webserver (www.swissmodel.expasy.org) is a popular webserver for homology modeling [18, 19, 20]. The provided webserver services allow expert and non–expert users to generate three–dimensional modeled protein structures. It assists the user in finding appropriate template structures and provides a quality score to evaluate the models.

The whole process includes four steps, starting with the identification of one or more structural templates, followed by the alignment of target sequence and template structure or set of template structures. Afterwards the models are built and finally the models are evaluated with a quality score. These four steps can be repeated until an appropriate protein model is found.

Some work in Chapter 4 takes advantage of the SWISS–MODEL webserver to model three–dimensional protein domains or complete three–dimensional structures.

1.5 Protein Binding Pocket Analysis

This section is taken from reference [21].

As early as 1894, Fischer introduced one of the first models of protein–ligand binding using the analogy of a lock for a rigid protein binding pocket and a key for a rigid and specific ligand to explain the interaction between an enzyme and its substrate[22]. The limitations of this model became clear when protein crystal structures were reported that showed a variety of pocket shapes for the same receptor cocrystallized with different ligands. Indeed, a model taking receptor flexibility into account was introduced in 1958 by Koshland [23] in which the protein binding site adapts by "induced fit" to bind the respective ligand. A further model, "conformational selection", in which the protein may adopt different conformations in its unbound state and a ligand binds selectively to one of these pre–existing conformations, was first employed to explain conformational changes of a protein binding site arising from the binding of an allosteric ligand [24]. This model has been supported by numerous experiments for both allosteric and non-allosteric ligands [25, 26].

Protein dynamics occur over spatiotemporal scales ranging from atomic fluctuations ($\sim$ femtoseconds) to protein folding and subunit association ($\sim$ seconds to hours)[27]. The size and flexibility of the individual components of a protein (e.g., amino acid side–chains or domains) influence the time scale over which their motions occur. The motions of different protein elements are often coupled, and both intrinsic protein flexibility as well as conformational adjustment to the interacting ligand may contribute to the binding process. The importance of the time scale of protein conformational transitions to distinguish between the two limiting cases, induced–fit and conformational selection mechanisms, has been highlighted in several studies [28, 29, 30]. The binding of glucose to human glucokinase and of geldanamycin to heat shock protein 90 (HSP90) are two examples where the predominant roles of the conformational selection [31] and induced fit [32] mechanisms, respectively, were demonstrated experimentally. Further examples are reviewed by Copeland [33]. More generally, protein–ligand binding can be considered to involve both mechanisms, but in any particular case, their relative importance varies, as does their effect on binding

1.5.1 Pocket Detection and Definition

To be able to bind one or more ligands, a protein pocket must possess or acquire a number of features that complement those of potential binders. Specifically, the volume should correspond to or exceed that of a ligand, the shape should enable the ligand to fit in, and the physicochemical properties should complement those of the ligand. Therefore, the main properties for characterizing a protein binding pocket are the overall geometry, the composition of amino acid residues, the type of solvation, the hydrophobicity, the electrostatics, and the chemical fragment interactions [34, 35]. These characteristics are also important for evaluation of the pocket's "druggability", i.e. its ability to bind a drug [36]. The position of a ligand in the holo–structure of a protein determined experimentally can be used to define the binding pocket and channels. Alternatively, computational methods can be applied for this purpose. Henrich *et al.* [34] reviewed the most popular computational approaches and tools to identify protein pockets and binding sites. These methods are based on either geometric or energetic analyses of the target protein structure, and some also use protein structure and/or sequence comparison. Often, the use of a combination of techniques improves predictions. This is exemplified by the Metapocket [37] webserver, which makes predictions of binding sites from a consensus from eight different shape–based binding site detection tools. Shape–based methods may fail if only unbound protein structures are available, and in this case, protein dynamics should be considered. Furthermore, many protein channels and tunnels are only partially open at any given moment, and thus approaches involving analysis of the protein flexibility are required for their detection, for example, using crystallographic thermal factors [38, 39] or molecular dynamics (MD) simulations [40]. These methods have recently been reviewed by Brezovsky *et al.*[41].

1.5.2 Pocket Comparison

Structural superimposition of protein pockets can enable the comparison of pocket functions, even if they have, for example, different shapes or chemical features. Whole protein structural alignment is the simplest, and a reliable approach for finding similar pockets even if the shape is not preserved. This method can be applied for proteins with high sequence similarity and for prion conformations

obtained in simulations. Bietz and Rarey improved this method by developing ASCONA, a tool to align multiple binding site conformations [42]. For cases with sequence conservation only within the binding pocket, a rough alignment of C–alpha atoms in combination with an analysis of amino–acid composition of the binding sites is a more effective procedure.[43] Sierk and Kleyweg summarized the main procedures for protein structure similarity comparison, including similarity measurements and a list of programs [44]. For proteins with only weak sequence/structure conservation within the pocket, comparison of the steric and physicochemical properties of protein cavities may be the only option that can be performed. There have been several proposed binding pocket similarity measures, using the positions of the atoms lining the pockets, specific amino–acids defining characteristic, contacts of the binding sites of interest, or their general physicochemical properties. Recently proposed comparison algorithms include methods using a convolution kernel for comparing clouds of atoms [45], distance histogram of specific reference points for a fast pocket comparison and classification [46], or graph–based algorithms to find pockets with similar physicochemical or functional behaviour [43, 47]. For more methods see also Kellenberger *et al.*[48] and Bartolowits *et al.* [49].

1.6 Molecular Interaction

Analyzing molecular interactions helps to understand the functions and behavior of proteins, and therefore assists in predicting the biological processes that a protein of unknown function is involved in. For example, if a protein of unknown function associates with one of known function it is likely that both play a role in the same biological pathway. Therefore, molecular interaction analysis helps to understand biological processes and pathways.

Molecular interactions can be analyzed using experimental or computational methods. Here, the focus will be on the computational methods. For the calculation of molecular interactions, suitable models for the corresponding forces and energetic contributions are required. Force fields provide mathematical equations to compute forces that act on an atom at various positions in space. In all–atom force fields, the forces are calculated for each atom in the whole system, whereas in a coarse grained or united–atom force fields the forces are calculated for groups of atoms. A simple force field considers bonded and non–bonded potential ener-

gies. The total energy (E_{total}) of the system can be described as the sum of the bonded (E_{bonded}) and non–bonded ($E_{non–bonded}$) terms as shown in Formula 1.1.

$$E_{total} = E_{bonded} + E_{nonbonded} \tag{1.1}$$

The bonded interactions (E_{bonded}) include bond stretching (E_{bond}), bond–angle bending (E_{angle}), and dihedral ($E_{dihedral}$) terms, as described in Formula 1.2. The bonded interaction term sums the forces which occur between atoms connected by a chemical bond. These short–range interactions are stronger than non–bonded interactions.

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral} \tag{1.2}$$

Non–bonded interactions ($E_{nonbonded}$) include the van der Waals (E_{vdW}) and the electrostatic ($E_{electrostatics}$) energies, see Formula 1.3. The interactions between atoms that are not connected by chemical bonds can be described by the Coulomb potential for the electrostatic interactions and the Lennard–Jones potential for the van der Waals interactions. This results in the equation shown in Formula 1.4.

$$E_{nonbonded} = E_{vdW} + E_{electrostatics} \tag{1.3}$$

$$E_{nonbonded} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_r \epsilon_0 r_{ij}} \right) \tag{1.4}$$

The first term describes the Lennard–Jones potential, which approximates the van der Waals interaction. ϵ_{ij} is the depth of potential well and σ_{ij} is the collision diameter. The second term describes the Coulomb potential, where q_i and q_j are the values of the charges, r_{ij} is their separation distance, ϵ_0 is the permittivity of free space and ϵ_r the relative dielectric constant of the medium in which the charges are placed.

The electrostatic forces, potentials, and interactions are a main focus in this thesis. In Chapter 4, electrostatic potentials are used to compare HSP40 proteins and the electrostatic interactions are the most relevant forces for the protein docking simulations. Therefore, the calculation for electrostatic interactions and potentials are discussed in more details in the following.

1.6.1 Electrostatic Potentials and Interactions

Electrostatic interactions are long range interactions. The molecular interaction of proteins is influenced by the electrostatic potential of each protein, and plays a crucial role in many biomolecular processes, including molecular recognition and binding. The electrostatic potential of a protein can be calculated by solving the nonlinear Poisson–Boltzmann equation (PBE), shown in Formula 1.5. Where $\varepsilon(r)$ is the position dependent dielectric permittivity, $\rho(r)$ is the molecular charge density, and q_i and n_i are the charge and the concentration of the i–th ionic species in the bulk, respectively [50].

$$-\nabla\varepsilon(r)\nabla\phi(r) = \rho(r) + \sum_i q_i n_i e^{-q_i/k_B T} \tag{1.5}$$

As solving the PBE is computationally demanding, tools based on fast numerical approximation of solutions to the linearized Poisson–Boltzmann equation (LPBE), shown in Formula 1.6, exist. In the Formula 1.6 the κ is the Debye–Hückle screening length and accounts for the distribution of mobile ions in the solvent. The effective charges are fitted to reproduce the electrostatic potential of a solute in a homogeneous dielectric computed via solving the linearized Poisson–Boltzmann equation in a heterogeneous dielectric [51]. Such LPBE solving tools include the Adaptive Poisson–Boltzmann Solver (APBS) [52] and the University of Houston Brownian Dynamics (UHBD) [53] tool, which is applied in Chapter 4. These tools calculate and store the electrostatic potentials of solute molecules on a grid with defined dimensions and grid spacing.

$$-\nabla\varepsilon(r)\nabla\varphi(r) + \varepsilon \cdot \kappa^2\varphi = \rho(r) \tag{1.6}$$

Electrostatic interactions describes the attractive and repulsive forces between two elements. In classical force fields, the interaction of two molecules is calculated by considering the partial charges of atoms in both molecules. For this purpose, the Coulomb's law can be applied on the atoms, partial charges respectively, in both molecules as described before. For an appropriate simulation of the influence of the solvent, it is recommended to explicit water molecules when using Coulomb's law and perform the simulation in a periodic solvent box.

1.6.2 Protein Interaction Property Similarity Analysis (PIPSA)

Comparing molecular interaction fields, such as electrostatic potentials, provides a way to analyze the interaction properties of proteins and can be performed with the PIPSA [54, 55, 56] software, which is also available as a webserver [57].

Electrostatic potentials play an important role during the molecular interaction process. The comparison of these potentials from different proteins can help to understand how and where molecules interact with each other. The PIPSA software assists the user during this analysis, and provides an automatic workflow to calculate the electrostatic potential of several proteins, compare them, and cluster them by their electrostatic similarity. This enables the detection of proteins with similar electrostatic potentials.

For a PIPSA analysis, the structures of the proteins need to be superimposed on each other, and the electrostatic potential grids are then calculated for every single protein. For this purpose, the APBS [52] or UHBD [53] tools can be used. For the electrostatic comparison, a "skin" of thickness δ is defined around the molecules, starting a distance σ from their van der Waals surfaces. A visualization of system definition for PIPSA is shown in Figure 1.4.

Now, a pairwise similarity index (SI) is calculated. Formula 1.7 shows the Hodgkin index, which is commonly used to measure the similarity of two molecular potentials.

$$SI_{12} = \frac{2(\mathbf{p}_1, \mathbf{p}_2)}{(\mathbf{p}_1, \mathbf{p}_1) + (\mathbf{p}_2, \mathbf{p}_2)} \tag{1.7}$$

$(\mathbf{p}_1, \mathbf{p}_2)$, $(\mathbf{p}_1, \mathbf{p}_1)$, and $(\mathbf{p}_2, \mathbf{p}_2)$ are the scalar products of the electrostatic potentials over the region where the potentials are compared. This means, if if the two potentials are identical SI_{12} is +1, if the potentials are uncorrelated SI_{12} is zero, and if they are anti–correlated SI_{12} is -1.

The pairwise SI is calculated by comparing each grid point within the intersection of the skins of the two molecules being compared. In some cases this intersection can be very small, or even empty. This is often the case if the protein structures being compared have very different shapes, due to high flexibility, bad structural alignments, or other shape influencing reasons.

The PIPSA algorithm allows a global comparison of the whole skin of the proteins or a local comparison. In a local comparison, an additional center and radius have to be defined. In this case, the similarity indices are only compared for those

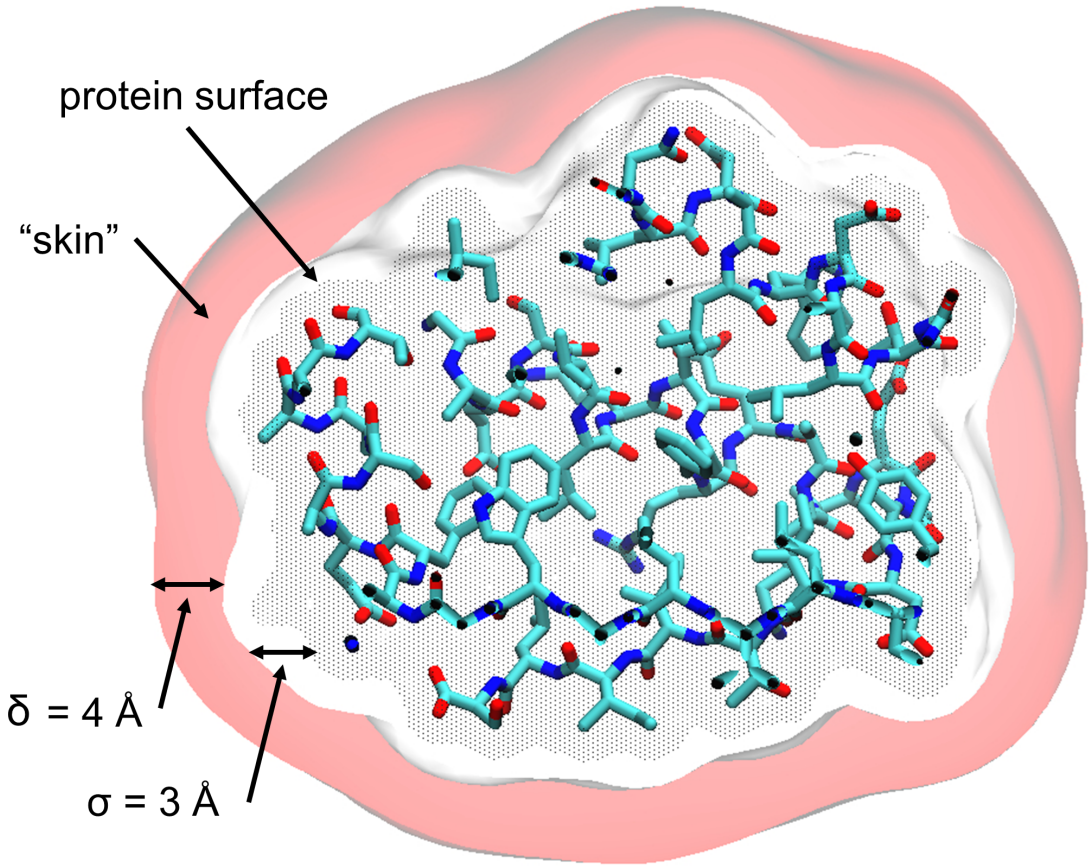

Figure 1.4: System definition for PIPSA. At all points in the "skin" of the molecule the SIs are computed. The "skin" (pink) has a thickness δ and is defined at a distance σ (white) from the van der Waals surface (dotted region) of the molecule. The used barstar protein (PDB ID: 1brs [58]) for this visualization is shown in licorice and the unlabeled representation was made and thankfully provided by Musa Özboyacı.

intersection grid points that are within the given radius around the specified center. This option can be helpful to compare protein binding pockets or interaction sites.

The similarity indices can be used to calculate a distance matrix and perform a clustering procedure, as described in Section 1.7. This clustering allows proteins with similar global or local electrostatic potentials to be group together. This is highly relevant for the comparison of the same protein from different organisms and allows an evolutionary analysis of electrostatic potentials, as is performed in Chapter 4.

1.6.3 Brownian Dynamics

Brownian motion describes the random motion of particles due to collisions with solvent molecules. This motion can be observed with a light microscope, as first reported by Robert Brown in 1827 [59]. The theory behind Brownian motion was later described by Einstein [60] and Smoluchowski [61].

The dynamics of solute particles, including macromolecules, can be simulated by the Brownian dynamics (BD) technique, which can be applied to investigate the diffusion–driven binding processes of molecules. Random diffusion of molecules can lead to the formation of, so called, 'encounter complexes' which can help to understand the initial recognition, interaction, and binding process of two or more molecules.

Ermak and McCammon developed an algorithm that describes the translational motion of a particle by calculating the displacement from the current position $\mathbf{r}_0$ during a timestep δt given the forces $\mathbf{F}(r_0)$ that are currently acting on the particle. A form of this algorithm, which ignores hydrodynamic interactions, is given in Formula 1.8 where $\mathbf{R}(\delta t)$ describes a random displacement that satisfy $<\mathbf{R}> = 0$ and $< \mathbf{R}^2 >= 6 \cdot D_T \cdot \delta t$. D_T represents the translational diffusion coefficient.

$$r = r_0 + \frac{D_T}{k_B \cdot T} \cdot F(r_o) \cdot \delta t + R(\delta t) \tag{1.8}$$

One advantage of the BD technique in comparison to, for example, molecular dynamics (MD) simulation, is that it allows the generation of trajectories on much longer temporal and spatial scales with lower computational cost. This also allows the simulation of bigger proteins and also multiple proteins, which is especially relevant for simulating the crowding effects of cell components.

Here, BD simulations are applied to investigate the protein–domain interactions of HSP40 domains (see Chapter 4). The most relevant reasons why the BD technique was chosen are the huge size of one of the binding partners, the unknown interaction site, and the number of required simulations. The Simulation of Diffusional Association (SDA) software is a package to, besides other things, simulate the docking process of proteins and protein–domains. A brief overview of the SDA software, which employs the BD technique is described in the following section.

1.6.3.1 The Simulation of Diffusional Association (SDA) Software

The Simulation of Diffusional Association (SDA) software was initially developed by Gabdoulline and Wade for the calculation of association rates of proteins [62, 63]. Since the first version, additional functionalities and improvements have been implemented into the SDA software package. The latest version (SDA7 [64]) is applicable to simulate the diffusional association of two solute molecules within a continuum solvent model. This also includes the association of a solute molecule to an inorganic surface. In addition, SDA can be used to simulate multiple proteins, which is helpful for studying the macromolecular crowding effect.

SDA can be used to calculate molecular association rate constants for known bound complex structures. In case the bound protein complex structure is not known, SDA can simulate a rigid docking and predict the diffusional encounter complex. For this purpose, one protein is kept rigid and fixed in the three–dimensional space, whereas the second protein's (often the smaller protein) diffusion is simulated by applying BD. The predicted encounter complexes can then be used for a refinement process (e.g. with a short MD simulation) to obtain a fully bound complex structure. The latest SDA version also allows multiple conformations for both solute molecules, which integrates some kind of intramolecular dynamics.

During the BD simulations in SDA, different force–field terms are used to model the effect of the solvent on the interaction energies and forces. This includes the electrostatic interaction energies (el) of two solutes, and their polar (edesolv) and non–polar desolvation (np) energies, shown in Formula 1.9.

$$\Delta G^{1-2} = \Delta G^{1-2}_{el} + \Delta G^{1-2}_{edesolv} + \Delta G^{1-2}_{np} \tag{1.9}$$

The electrostatic interaction energy between the two solute molecules are computed by using Formula 1.10.

$$\Delta G^{1-2}_{el} = \frac{1}{2}\sum_{i_1} q_{i_1}\Phi_{el_2}(r_{i_1}) + \frac{1}{2}\sum_{i_2} q_{i_2}\Phi_{el_1}(r_{i_2}) \tag{1.10}$$

The $\Phi_{el_2}(r_{i_1})$ and $\Phi_{el_1}(r_{i_2})$ represent the electrostatic potentials (explained before) of molecule 1 and 2 at the effective charge site q_{i_1} on the first molecule and q_{i_2} on the second. The electrostatic desolvation (edesolv) describes the effect that the interactions of charges that lie on the surfaces of the solutes with the solvent

decrease when two solutes come closer to each other. This can be calculated as shown in Formula 1.11. Where q_i is the effective charge on a solute and $\Phi_{edesolv}(r)$ is the electrostatic desolvation potential. The non–polar desolvation term is no considered in the simulations presented in this thesis, therefore they are not explained in more detail.

$$\Delta G^{1-2}_{edesolv} = \sum_{i_1} q^2_{i_1} \Phi_{edesolv_2}(r_{i_1}) + \sum_{i_2} q^2_{i_2} \Phi_{edesolv_1}(r_{i_2}) \tag{1.11}$$

The electrostatic desolvation potential ($\Phi_{edesolv}(r)$) is approximated by the following Formula 1.12:

$$\Phi_{edesolv}(r) = \alpha \frac{\epsilon_{solute} - \epsilon_{solvent}}{\epsilon_{solute}(2\epsilon_{solute} + \epsilon_{solvent})} \sum_j a^3_j \frac{(1+\kappa r_j)^2}{r^4_j} e^{-2\kappa r_j} \tag{1.12}$$

α is an empirical parameter used to scale the interaction potential strength, κ is the inverse Debye length, and ϵ_{solute} $\epsilon_{solvent}$ are the dielectric permittivity constants of the solute and solvent, respectively.

In Chapter 4, the SDA docking method is applied to the C–terminal domain and J–domain of the HSP40 protein family. By keeping the large CTD rigid and applying Brownian Dynamics on the the smaller J–domain, a configurational sampling of protein–domain docking is obtained and potential interaction sites are analyzed.

1.7 Clustering

In the last few years, the topic of "Big Data Analysis" has become very popular, as it provides the possibility to extract information from a huge set of data. This large amount of data cannot be analyzed by hand any more, and requires new methodologies to handle and process data. The idea to cluster data is not a new method, and might not be appropriate for a very large amount of data because of its computational complexity, but it is an important topic as it provides one way to handle data and extract relevant information. Nevertheless, one should remember, that during a clustering procedure, the information content always gets reduced so as to be able to extract some kind of core structure or feature of the dataset. The data that is not part of the core elements might be noise, or maybe important

outliers that should be considered. To find a balance between data reduction and information content is one of the biggest problems in clustering.

All clustering methods can be divided into two basic forms: partitional and hierarchical. The basic idea of a partitional algorithm is to take a dataset with multiple elements *N*, and a predefined number *k* of partitions to be obtained, and return *k* disjoint subsets of the data representing the clusters of the dataset. The hierarchical clustering algorithm constructs a tree, either from the bottom up (*agglomerative*) by merging single data elements and groups into larger groups, or from the top down (*divisive*) by splitting the data into smaller groups. Often the root of the tree represents the whole data set and the leaves are the single data elements. In those trees, the branch lengths are often directly related to the measured distance between the joined or split elements/groups. By cutting the tree at a meaningful level a relevant partitioning of the data can be derived. Beside these two basic clustering forms there exist many variants on one of the two, or a mixture of both.

In the following, general information about clustering will be explained, for a more detailed overview about clustering procedures in the field of bioinformatics and drug discovery, read the book by John D. and Norah E. MacCuish [65].

1.7.1 Data Types and Distance Measures

The type of data plays an important role in deciding which clustering method and distance measure should be used. There are four different basic data types to be distinguished: binary, ordinal, continuous and categorical data. Depending on the data type an appropriate measurement for the similarity or dissimilarity of two data elements is necessary. Actually, many measures exist that quantitatively measure how similar or dissimilar two data elements are. Most of these measures are symmetric, which is also required for many clustering algorithms. This means that the distance between point *x* and *y*, *d(x,y)*, is the same as between point *y* and *x*, *d(y,x)*. Some distance measures are, for example, the Tanimoto measure (binary data), Manhattan distance (continuous data), or the Pearson correlation coefficient (continuous data). For more details and further measurements, see [65]. Another example for a measurement of continuous data is the Euclidean distance. It is a popular, intuitive, and easy to understand geometry based measure that describes how near or far two elements are in space. The distance *d* for

elements representing data points *x* and *y* in Euclidean space with dimension *P* can be calculated as follows:

$$d = \sqrt{\sum_{i=1}^{P}(x_i - y_i)^2} \tag{1.13}$$

In this context one of the most commonly used distance measures in the field of protein structures should be mentioned – the root–mean–square deviation (RMSD). It is an adaption of the Euclidean distance (see Formula 1.14) to measure the average distance δ between *N* atoms in (superimposed) protein three–dimensional structures.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\delta_i)^2} \tag{1.14}$$

1.7.2 Clustering Algorithms

Many algorithms are available for partitional and hierarchical clustering. The most appropriate one has to be chosen according to the data type, distance measure, and the required type of results. Here, the basic ideas of the most commonly used ones, and those applied in this thesis will be briefly described. This includes *K*–means as a partitional algorithm, and average linkage and Ward's clustering as hierarchical algorithms.

1.7.2.1 *K*–Means Clustering

The most popular partitional algorithm is *K*–means. The goal of the *K*–means clustering algorithm is to divide a data set into *k* meaningful partitions. This means, that the number of clusters needs to be defined at the beginning, and the number of clusters is not influenced by the data itself.

At the beginning, the algorithm generates *k* arbitrary centers within the feature space of all elements in the data set. These centers can be actual, or artificial, data points. In an initial process, all data points are assigned to the nearest neighbor center. For the assigned groups around each center, a new center is calculated based on the means. In an iterative process, the data points are assigned to the updated centers and afterwards the centers are updated again. Therefore, the elements in the each group will change until the process converges and the groups do not change anymore. It is also possible to define a maximum number of iterations.

1.7.2.2 Average Linkage Clustering

The average linkage clustering, also known as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is an hierarchical clustering method, which defines the distance between two clusters as the average distance between all pairs of cluster elements. The idea of the average linkage criterion originated with Sokal and Michener in 1958 [66].

As in many agglomerative hierarchical clustering algorithms, the average linkage method starts with the definition of a triangular matrix showing the similarity or distance between each pair of the *N* data elements. Afterwards the pair with the most similar elements, meaning with the smallest distance are grouped together. The matrix is updated according to the Formula 1.15. The linkage criterion, which is determined by the average distance of each pair of elements (d(a,b)) in both clusters, is applied in every clustering repetition step for the merged cluster (A and B). Based on the linkage criterion, the average linkage clustering tends, in general, to join clusters with small variances, and is slightly biased towards producing clusters with the same variance.

$$\frac{1}{|A| + |B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \tag{1.15}$$

1.7.2.3 Ward's Clustering

Similar to the average linkage clustering, Ward's clustering is also an agglomerative hierarchical algorithm. It was introduced by Joe Ward in 1963 [67]. This clustering is based on Ward's minimum variance criterion, which minimizes the total variance within each cluster. During the initialization step, the distances in the matrix are often defined to be the squared Euclidean distance between points. Afterwards, in every clustering step, two clusters are merged together that lead to a minimum increase in the total variance of all clusters. For an implementation of the Ward's clustering method, the Lance–Williams algorithm can be applied, as described in detail in [68, 69].

1.8 Webserver Interfaces

Scientific webservers have become more and more popular in recent decades, as they provide easy access to scientific software. Scientific software often requires

some knowledge and experience regarding the execution and appropriate input parameters. Therefore, new users or beginners often have problems to use new scientific software.

Webservers can help to overcome these problems by assisting the user, guiding through the preparation and analysis steps. This also allows a restriction on the user input, and therefore reduces commonly occurring errors, or meaningless results because of wrong user inputs. Providing scientific software via webservers makes them easily accessible to expert and non–expert users in an user–friendly way. Furthermore, when software is updated, an update is only necessary on the webserver, and all users have immediate access to the new software version. This allows quick updates, even for small changes, and makes it easy to maintain the software.

A new webserver should be user friendly, clearly structured, and simple to use. To accomplish this, it is helpful to use a framework to build the webserver. All new developed webservers in this thesis use the Play Framework (www.playframework.com), and take advantage of JavaScript (JS) and cascading style sheets (CSS). Furthermore, all of the presented webservers use one of the most popular HTML, CSS, and JS frameworks for webservers, called Bootstrap (www.getbootstrap.com). This additional framework makes it easier to design modern–looking, well–structured user interfaces, with additional JS based functionality and animations. This is especially useful to keep a neat page layout, even if a lot of data needs to be displayed to the user.

1.8.1 Play Framework

The Play Framework (www.playframework.com) is based on the model–view–controller (MVC) architecture pattern, which is popular for desktop graphical user interfaces in web applications. MVC divides the application into three interconnected parts, and separates server internal information (stored in the models) from those visible for the user (shown in the view). The communication between the "models" and the "view" is done by the "controller". This allows a clear structure of the web server that can easily be extended.

The Play Framework is based on Java and Scala, and integrates existing Java libraries for core functionality, such as data storage. Furthermore, integrating commonly used Java and Java Script libraries, as well as, CSS is also possible.

2

Visualization of Sequence Annotations on Protein Structure

Sections of this chapter are based on the following publication:

Antonia Stank, Stefan Richter, Rebecca C. Wade. ProSAT^{+}: visualizing sequence annotations on 3D structure. Protein Engineering, Design and Selection, (2016) 29:281-284

Visualizations shown in some figures in this chapter are part of the above mentioned reference and were initially done by myself and are published in this thesis with a license agreement with Oxford University Press.
Text sections taken and adapted from the above mentioned references are marked in the beginning of a section. This text was initially written by me and went through the review process of all authors.

2.1 Introduction

The growing amount of protein sequence, structure and annotation data provides an ever richer basis for making discoveries and solving research problems. However, to make sense of these data, it is necessary to assemble information on the sequence, structure and annotation for simultaneous visualization and analysis. Doing this by hand can be quite tedious and error-prone, as it is necessary to ensure that the data assembling is correct. It is a complex process that requires expert knowledge about the different data sources and data formats.

Despite the complexity, it is worth while to invest in this data assembling process as the two web servers, ProSAT [70] and ProSAT2 [71], could already show. ProSAT uses sequence annotation data from SwissProt [72] and predicted sequence patterns and feature/functional sites from Prosite [73] to map them via sequence alignment onto the 3D protein structure. Following the success of ProSAT, the ProSAT2 web server was introduced that further includes sequence annotations from the UniProt KnowledgeBase [74] and the BRENDA enzyme information system [75]. ProSAT2 also introduced new functionalities to select residues annotated with certain criteria and visualization of user-prepared annotations.

Over the last ten years, the amount of sequence entries including sequence annotations stored in the UniProt KnowledgeBase grew enormously. In 2006, when ProSAT2 was published, there were less than 5 million sequence entries in the UniProt database, whereas the latest release (November 2016) contained more than 71 million sequence entries (source: http://www.ebi.ac.uk/uniprot/TrEMBL stats). This impressive amount of sequence information leads to the fact that the UniProt database is probably the best known data source for protein sequence information today.

In the case of 3D protein structure data, the Protein Data Bank (PDB) [76] (www.rcsb.org) is the most prominent data source. The content of this database also increased in the last ten years from around 40,000 structures in 2006 up to more than 125,000 in 2017 (source http://www.rcsb.org/pdb/statistics/content GrowthChart.do?content=total).

Existing web servers using databases containing sequence and structure data automatically benefit from the continuous increase in data. Furthermore, the significance of using, assembling and analysing those data becomes more and more important. Therefore, other web servers for visualizing sequence annotations on protein structures were developed, such as AMASS [77] and Aquaria [78]. Both display the protein structure with the Java-based Jmol application (www.jmol.org), which requires Java to be installed on the client side and often leads to security issues, meaning that the user can encounter problems trying to run the application in the browser. The RCSB PDB web server [76] provides several viewers including the JavaScript-based JSmol (www.sourceforge.net/projects/jsmol) and the WebGL (Web Graphics Library) based NGL viewer [79], allowing an easy-to-use and fast 3D structure inspection. However, annotated sequence features cannot be displayed on these structural views. Protter [80] is a web tool

for visualizing protein sequence and topology, but not the 3D structure, together with sequence or other annotations. The UCSF Chimera visualization program for molecular structures [81] provides sequence feature annotations from UniProt on 3D structures but, as a standalone program, it requires installation and user experience.

A new web server was developed for visualizing sequences and sequence annotations simultaneously on a protein structure that incorporates key features of ProSAT and ProSAT2 as well as significant new capabilities for handling, visualizing and sharing of the data. The new Protein Structure Annotation Tool-plus (ProSAT$^+$) [82] web server was developed to assist both expert and non-expert users with analyzing protein sequence annotations in the context of the corresponding 3D protein structure. It is as available at: http://prosat.h-its.org/ and provides a combination of some of the features of the above-mentioned tools. The index page of the ProSAT$^+$ webserver contains the initial input fields to search for a protein structure and is shown in Figure 2.1. In addition, ProSAT$^+$ allows mapping of user-defined and functionally classified sequence annotations on the structure, the opportunity to export these via URL, and the transfer of sequence annotations from similar sequences found with BLAST+ [83] (version 2.2.29). Furthermore, the specific URLs enable the easy integration in any other web application and can therefore be used as a module that provides a set of additional features.

In the following sections, the data sources used in ProSAT$^+$ are explained in more detail. Then the whole workflow and technical details are described. This is followed by an example usage, and a discussion and an outlook on future features is given.

2.2 Data Sources

The main focus of ProSAT$^+$ is to assemble data from different sources and visualize them simultaneously in an easy to use way. The UniProt database serves all available sequence annotations for a specific protein, which are then mapped on the 3D structure that is available from the PDB. This mapping procedure requires an alignment of the two sequences, one in the UniProt entry and one in the PDB entry. The SIFTS database [84] contains, besides much other information, precalculated sequence alignments. This allows a quick and reliable mapping of all sequence annotations available in the UniProt entry on the 3D protein structure

from the PDB. All three databases are introduced and some technical functionalities that are used in ProSAT$^+$ are discussed in detail in the following sections.

2.2.1 Universal Protein Resource (UniProt)

The Universal Protein Resource (UniProt) [85] was launched in 2003 and published for the first time in 2004 [86]. It is an initiative of the UniProt consortium, including the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). It hosts about 71 million sequences. It can be reached under www.uniprot.org and oday it is the most prominent resource for protein sequence and annotation data. UniProt consists of four main components: UniProtKB, UniRef, UniParc and UniMes [87].

The fundamental component is the UniProt Knowledge Base (UniProtKB) containing the sequence data. Besides the UniProtKB TrEMBL where the data is mined automatically by advanced computer algorithms, the UniProtKB is comprised of the UniProtKB Swiss-Prot. Here, the high quality data is manually curated and non-redundant. The aim of UniProtKB is to provide all available information about a protein including splice variants, polymorphims or post translational modifications. Also protein families and groups are provided.

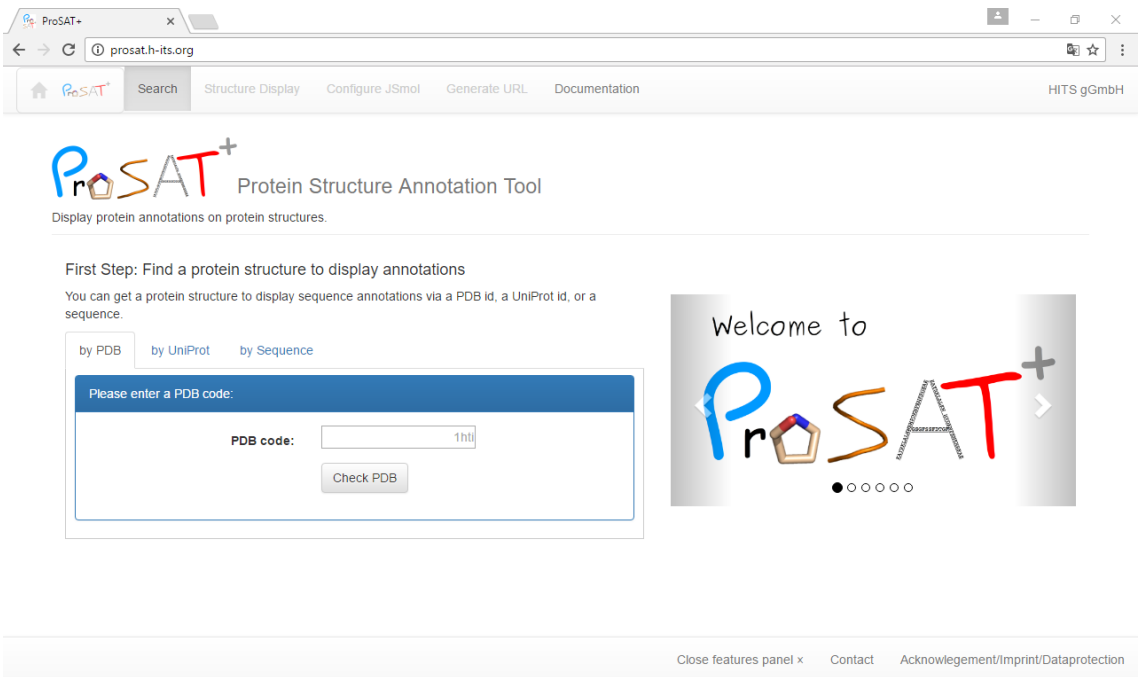

Figure 2.1: Index page of the ProSAT$^+$ webserver. Structures for mapping sequence annotations can be searched by PDB ID, UniProt accession number, or DNA/Protein sequence.

Each sequence that is included to the database is assigned an unique and conserved accession number (AC), which is a stable identifier of 6 or 10 alphanumerical characters. The entry name is another unique identifier often containing relevant information about the entry. The accession number enables a unique URL request to the UniProt server to obtain a file in XML format containing all data of this entry. This so called representational state transfer (REST) request is used in ProSAT$^+$.

Additional services of the UniProt webserver provide an easy usage and comparison of sequence data for all researchers. The reviewed and non-reviewed data are regularly updated in the database and the amount of data is growing every month. UniProt provides a state of the art web interface making it convenient for researchers from different scientific fields to access and find relevant data and information.

2.2.2 Protein Data Bank

The Protein Data Bank (PDB) [76] was established in 1971 at Brookhaven National Laboratory and is the most popular archive for information about the three dimensional molecular structures, of proteins and nucleic acids. This data comes from all found organisms and different species and is freely available to all users. Three dimensional structures of molecules enable a detailed analysis of the molecular function and interaction with other molecules, for example, drugs. The PDB includes molecular structures of several sizes, including small sub-domains of proteins up to huge complexes of multiple proteins.

Each structure entry has a unique ID (PDB ID) that helps to find the 3D structure in the database at: www.rcsb.org. The PDB archive is updated weekly and the website assists users to perform simple and complex queries on the data, analyze, and visualize the results. In addition the webserver provides several services including RESTful web services that allow an easy and structured access to PDB entries and their information and data. One of these is called 'describeMol' and it is a querying tool which sends descriptions of the entities that are contained in a PDB file. This service can be accessed, for example, via http://www.rcsb.org/pdb/rest/describeMol?structureId=4hhb for the PDB ID 4hhb. This 'describeMol' service is used in the ProSAT$^+$ webserver to receive and extract general information for the displayed structure.

2.2.3 Structure Integration with Function, Taxonomy and Sequence

The Structure Integration with Function, Taxonomy and Sequence database (SIFTS DB) [84] was established in a collaboration between the Protein Data Bank in Europe (PDBe) and UniProt. The database provides a mapping between UniProt and PDB entries on residue-level and is updated every week after the PDB update. It also includes annotation data from the IntEnz [88], GO [89], Pfam [90], InterPro [91], SCOP [92], CATH [93] and PubMed [94] resources.

For each PDB entry, the SIFTS pipeline is applied, which consists of two main components. First, an automated process to identify the correct UniProtKB cross-references for each protein chain in the structure is applied. Afterwards, an automated pipeline produces a residue-based correspondence between the protein sequence in the PDB and the corresponding UniProtKB sequence. The data is stored and distributed as a XML file in the FTP server ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts. For more details see Velankar *et al.* [84].

The residue based mapping of the PDB and UniProt protein sequences is used in the ProSAT+ webserver as it provides a fast, reliable and easy to access resource for mapping each residue in the protein sequence to the correct residue in the protein structure.

2.3 ProSAT+ Workflow

This section is taken from reference [82].

The ProSAT+ web server is implemented in Java on the server side and uses the Play framework 2.4 together with the Twitter Bootstrap (version 3.3.6) for the HTML front-end, thereby allowing a simple-to-use and clear interface.

ProSAT+ is designed to be modular and extensible. The overall workflow of ProSAT+ is illustrated in Figure 2.2. Initially, it is necessary for the user to define the protein of interest. This user input can be in the form of PDB ID codes, UniProt [85] accession/entry names or a protein or DNA sequence. In the case of a PDB ID code, the referenced UniProt accession in the RCSB PDB entry, obtained from the RCSB PDB RESTful web service called describeMol, is used for linking to the Uniprot entry and extracting the sequence and sequence annotations. If a UniProt

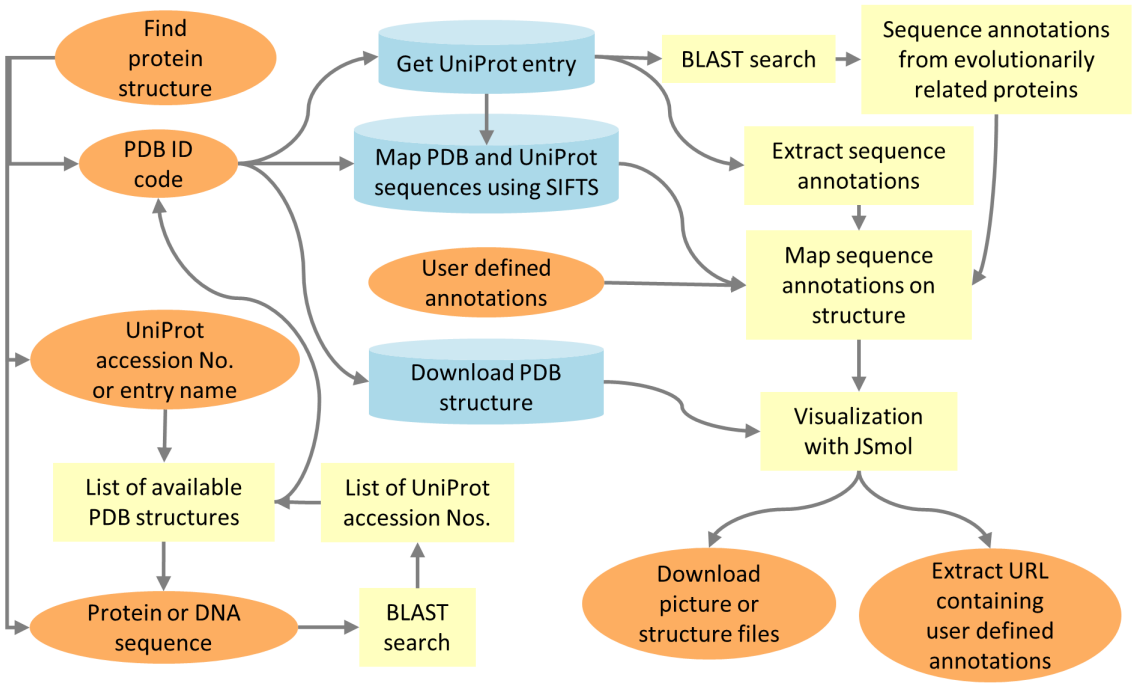

Figure 2.2: The ProSAT+ pipeline integrates user input data (ellipses) and data from databases (cylinders) by data processing (rectangles) to enable visualization of the protein structure and sequence annotations (rectangle) and output of the associated data (ellipses). Figure was published in reference [82].

accession/entry name is provided, its link to the PDB is followed to access structural information. If several PDB entries are given in the UniProt entry, the user can select one PDB entry from a given list. If no protein structure is listed in the UniProt entry, the sequence is extracted and a BLAST search against the section of UniProt containing links to the PDB is performed. If a protein or DNA sequence is provided initially, the sequence is again used for a BLAST search against the UniProt section with links to the PDB. In both cases, the user can select a matching UniProt accession for which a 3D structure is available.

Once the PDB and UniProt entries are defined, the corresponding SIFTS [84] mapping XML file containing the validated sequence mapping between the sequence in the PDB structure file and the sequence in the UniProt entry is downloaded by ProSAT+. This mapping file is important since the mapping of the PDB and UniProt sequences regarding the sequence numbering can be shifted. Such sequence numbering differences can present a major problem for manual mapping of sequence annotations to structures and can lead to the wrong mapping of these annotations on the structures. A correct mapping of the two sequences

can be achieved with the regularly updated and precalculated SIFTS mapping. All sequence annotations are extracted from the UniProt entry using its XML format and, together with the SIFTS mapping file, all annotations can be displayed at the correct residue number on the visualized PDB entry using JSmol. In some cases, there are only a few sequence annotations in the UniProt entry for a specific protein. Therefore, ProSAT+ offers the possibility to search for evolutionarily related proteins using BLAST+ and to map their sequence annotations onto the protein structure of interest. For this UniProt sequence mapping, the alignment module of the BioJava project (substitution matrix: PAM250, gap open penalty: 6, gap extension penalty: 1) [95] is used. The UniProt annotations are shown using a functional classification [71] in a sliding side panel on the web page as well as on a sequence display in the main window.

The 3D protein structure visualization with the Java Script-based JSmol can be enriched by user annotations that can be defined by the user directly in a URL and shared with others, for example, for publishing or sending to a project partner. In this case, it might be helpful to copy the URL into an external URL shortener (e.g. http://goo.gl) and use the short URL afterward. The use of JSmol, rather than the Java-based Jmol application for visualizing the protein structure, which was used in the predecessors of ProSAT+, ProSAT and ProSAT2, means that there is no requirement for Java to be installed on the client side and the associated security issues are avoided.

ProSAT required the prior assembly of data from many sources, e. g. the PQS database [96] and SwissProt [72]. In ProSAT+, the approach taken is to collect and pool the data only when it is needed. This has the advantage that the data are always up to date and do not need any regular internal database updates of the web server except for the UniProt sequences in FASTA format used for BLAST+ searches. In the case of data format changes in external data sources, some work might be necessary in ProSAT+ to maintain compatibility. However, since the well-defined XML formats for UniProt, SIFTS and PDB header information are used, this is anticipated to be straightforward.

In the structure visualization mode, the user is provided with information about the protein of interest and can change the structure representation. A sequence scroll bar contains information about the UniProt and PDB sequence mapping, the residue numbers, and available sequence annotations. A short motif search interface makes it possible to find specific protein sequence motifs via a

regular expression search and to highlight them on the protein structure. A side panel gives an overview of all available sequence annotations for different categories, for example, binding sites or residues annotated as natural variants. Using a check box, it is possible to select all annotations of one category at once. The selected annotations are visualized and labeled on the 3D structure. Various predefined structure representations can be selected from the menu, but it is also possible to use specific JSmol commands by opening the JSmol console. The final visualization can be exported as a picture or a 3D model (.x3d format). A documentation page is included in the web server where all elements of the visualization are described in detail.

2.4 Example Usage

This section is taken and adapted from reference [82].

The usability of ProSAT$^+$ is demonstrated by way of an example application to hemoglobin, shown in Figure 2.3. Hemoglobin is a well studied protein because of its relevance in oxygen transportation and sickle-cell anemia. Assuming there is an interest in a specific variant and there is a need to understand why this variant is associated with a certain phenotype. The PDB ID code 1j7y [97] represents the structure of the hemoglobin heterotetramer containing two alpha subunits (UniProt accession: P69905) and two beta subunits (UniProt accession: P68871). This PDB ID code can be entered in the ProSAT$^+$ search interface, the structure is visualized and the sequence information for the four chains is shown.

One feature of ProSAT$^+$ can be seen by selecting the second residue ('V') in the sequence scroll bar. The residue is highlighted and labeled in the 3D structure and one can see that the residue numbers in the UniProt and PDB sequence differ by one. Further, the two different types of subunit also have different UniProt accessions meaning that they have different sequence annotations. ProSAT$^+$ automatically extracts all data from both UniProt entries and they can be visualized simultaneously on the 3D structure. This functionality can be important for sequence annotations in the binding interface of the different subunits. To look up a hemoglobin variant called 'M-Boston/M-Osaka' in the alpha subunit one can open the side panel by using the ProSAT$^+$ logo or the 'show features panel' button and select chain A, the different categories of sequence annotations are listed and, in

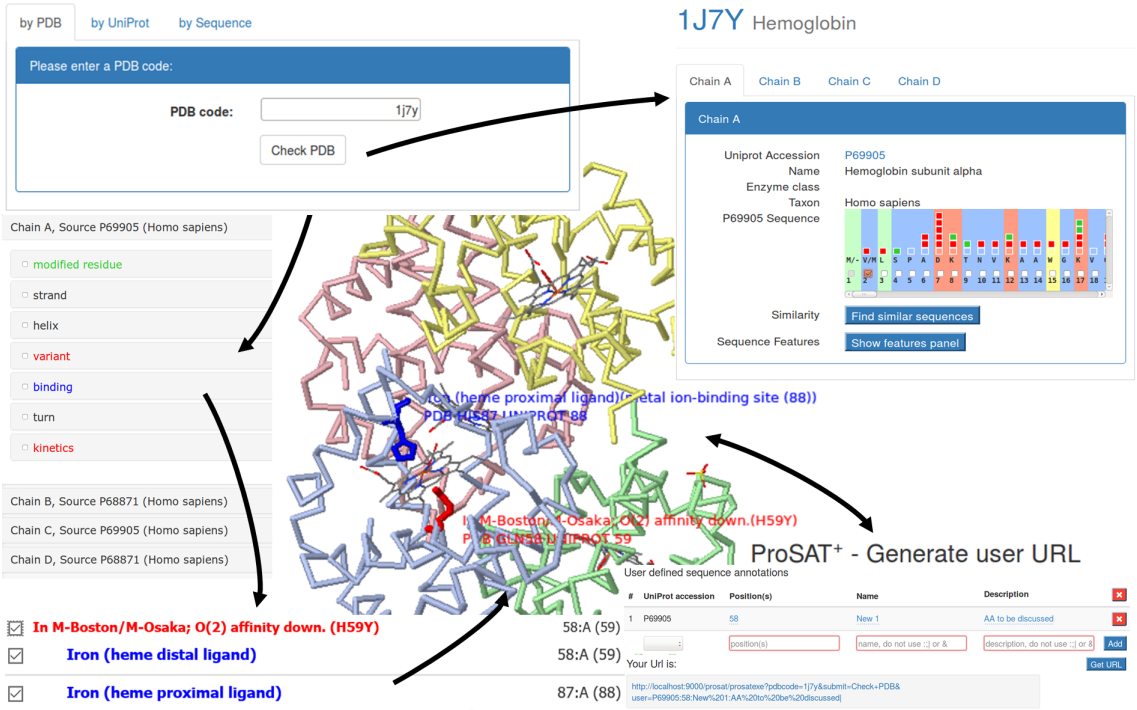

Figure 2.3: Screenshots illustrating the use of ProSAT$^+$. In this example, functional sites in the hemoglobin heterotetramer are explored. The user enters a PDB ID code, here 1j7y, and is provided with an annotated visualization of the protein structure. The sequence scroll bar and the sliding side panel can be used to select specific functional residues and find sequence annotations. The four subunits (Chain A-D) of this heterotetramer are represented with different colors (yellow, pink, blue and green) and the sequence annotations of the two corresponding Uniprot entries are listed in the side panel. See text for details. Figure was published in reference [82].

the 'variant' section, the one of interest can be found. The additional information that this variant leads to a decrease in oxygen-binding affinity can be verified through the link to SwissVar [98]. By selecting this annotation, the corresponding histidine at position 59 is visualized and labeled in the 3D structure. The amino acid is near to the bound heme (HEM) ligand of the corresponding subunit. In the category 'binding', two annotated residues can be found and by selecting both, one can see that not only does the histidine at position 59 bind the heme iron at the distal position but the histidine at position 88 is also annotated as relevant for binding and binds the heme iron at the proximal position. This multiple residue annotation can also be seen in the sequence scroll bar. ProSAT$^+$ allows the user to share relevant annotations with collaboration partners by simply defining the identified amino acid residue, e.g. at position 59, together with an annotation, and create a URL using the Generate URL menu function. This URL can be sent

around or published (e.g. after the use of a URL shortener like http://goo.gl). Opening this URL will show all manually added annotations visualized on the 3D structure.

2.5 Discussion and Outlook

The immense growth of available data makes it more and more imperative to combine data from different sources, which in turn allows new discoveries and answers in fields which would not be able to be answered with the single data sources.

The ProSAT$^+$ webserver is one example where the combination of data from different sources leads to an increased information content. Protein sequence annotations often represent the state of the art knowledge about the molecular functionality of a protein. The mapping of this data on the three-dimensional protein structure allows researchers to better investigate and understand the overall function of a protein. ProSAT$^+$ provides users with additional features, for example sequence motif search, to simplify common tasks while analyzing proteins regarding their sequence, structure, and function. Furthermore, the design of ProSAT$^+$ allows the creation of a specific URL that directly links to a structure visualization that can include known and user-defined sequence annotations. This enables a ProSAT$^+$ link to be referenced whenever structure visualization is helpful or specific amino acid residues are discussed, such as in a publication or another webserver. This specific URL allows the integration of the ProSAT$^+$ service into any other webserver with a pre-defined structure and highlighted, annotated residues. This feature is already applied in the LigDig [99] and TRAPP webservers [100]. In case of the LigDig webserver, residues that were applied during a protein binding superposition procedure are forwarded via this specific ProSAT$^+$ URL and can be inspected in context of sequence annotations in the ProSAT$^+$ webserver. In case of the TRAPP webserver, the ProSAT$^+$ functionality is incorporated as an iframe module (an html element with content of another website) and allows the analysis of protein binding pocket dynamics in parallel with protein sequence annotations.

Even if the amount of data is increasing steadily, ProSAT$^+$ highly depends on the availability of sequence annotation data in databases or from the user. Many databases do not allow the visualization of their data on external webservers, which unfortunately limits the possibility to include such data in ProSAT$^+$. Nev-

ertheless, the modular design of ProSAT$^+$ means that in the future other sources of sequence and structural annotation data can readily be incorporated for display in ProSAT$^+$. These annotations could provide diverse types of information, for example, on sequence conservation and protein dynamics, which would complement the information on sequence variants. The ProSAT$^+$ webserver thus provides a user-friendly tool that can be adapted for different application scenarios and should be of value for non-expert and expert users alike.

3

Protein Binding Pockets

Sections of this chapter are based on the following publications and manuscript:

Antonia Stank, Daria B. Kokh, Jonathan C. Fuller and Rebecca C. Wade. Protein Binding Pocket Dynamics. Account of Chemical Research, (2016) 49:809-815

Jonathan C. Fuller, Michael Martinez, Stefan Henrich, Antonia Stank, Stefan Richter and Rebecca C. Wade. LigDig: a web server for querying ligand–protein interactions. Bioinformatics, (2015) 31:1147-1149

Antonia Stank, Daria B. Kokh, Max Horn, Elena Sizikova, Rebecca Neil, Joanna Panecka, Stefan Richter and Rebecca C. Wade. TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets. (2017) Submitted.

Visualizations shown in some figures in this chapter are part of one of the above mentioned references and were initially done by myself, or made by the person mentioned in the figure legend.
Text sections taken and adapted from one of the above mentioned references are marked in the beginning of a section. This text was initially written by me and went through the review process of all authors.

3.1 Introduction

Computational simulations of protein interactions became more and more relevant in the last decades, as they allow a precise analysis and provides the possibility to understand molecular interactions in more detail. Protein binding pockets are also in focus of these simulations. The additional knowledge of how a protein interacts with and reacts to another molecule is very important to understanding the overall protein functionality. Therefore, it is also the starting point to finding new ways to prevent specific protein interactions or to inactivate the whole protein, which is relevant to designing new drugs.

The field of computer aided drug design (CADD) has its beginning some decades ago. CADD often covers and assist several steps during the modern drug design process. One part of CADD is based on protein structure and sequence data. Several methodologies exist that, for example, can detect binding sites, calculate physicochemical properties of a binding site and a drug molecule, or simulate the whole binding process. High throughput docking simulations can help to reduce the number of molecules that are considered experimentally as possible candidates for a new drug. The computational docking of a protein–ligand (e.g. a drug) interaction takes many parameters into account (e.g. volume, size, shape, hydrophobic residues and charge) and can therefore be used to screen several thousands of candidates and discard those where the ligand features do not fit with those of the protein binding pocket. The more parameters can be taken into account during these simulation processes, the more realistic are the obtained results. One factor that was ignored for a long time, because it was computational too demanding, is the protein dynamics and its flexibility in the binding pocket. Additional computer power and better algorithms nowadays allow to simulate this flexibility. Available software and algorithms are discussed later in this chapter. Nevertheless, the simulation of an individual binding process for several thousands of drug candidates in high throughput screening is still not possible, but is the overall goal in the field of CADD.

This chapter gives a brief introduction to protein interactions and the alignment of protein binding pockets. Then it focuses on the computational analysis of protein binding pockets, their dynamics and classification, as well as the influence of the dynamics on the binding kinetics. At the end, a new webserver is presented

that allows the simultaneous analysis of protein binding pockets, the detection of transient pockets, and the visualization of sequence conservation and annotations.

3.1.1 Significance of Protein Interactions

In living organisms, the communication on the molecular level is highly dependent on interactions. Biological pathways are a cascade of several interactions that allow, besides many others, the regulation of reactions, processes, and level of concentration of molecules and cellular components. Without these interactions the whole molecular network in living organisms would not be possible and most probably life on earth would not exist. A better understanding of this highly complex interaction network helps to understand different kinds of diseases where, for example, a key component of a biological pathway is missing or not functional anymore, because of a mutation. The analysis of a specific molecular interaction is the beginning of finding ways to influence it with, for example, a new drug.

3.1.2 Functional Impact of Protein Interactions

Proteins can form several different types of molecular interactions, which may influence the molecular functionality of the protein itself and/or of the corresponding binding partner. The cascade of molecular reactions and complete biological pathways are highly dependent on these specific molecular interactions, which take place at the binding interface and provide compatible molecular properties for the respective binding.

The protein–protein interaction describes the binding of two proteins to form a molecular complex. Big molecular machineries are often built of several proteins that bind together to form a functional machine that is only able to perform as a complex. The protein–protein interaction sites are often shallow depressions or protrusions on the surface of a protein.

Another very important type of protein interaction is the binding of small molecules, including drugs and other low molecular weight ligands. This binding is especially relevant for the activation, inhibition, or substrate binding of enzymes. This interaction often takes place in more cavity like regions of the protein, as the ligand can bind more easily for a longer time in this protected area. These so called binding pockets are discussed and classified in more detail in the following sections of this chapter with a focus on their dynamics.

The functionality of proteins can also depend on the binding of ions, which often play a crucial role in the interaction with other molecules. One example are the class A HSP40 proteins, whose protein–domain interaction is analyzed in detail in chapter 4. This class of proteins bind zinc (ZN) ions at a zinc–finger–like region (ZFLR) that also contributes to substrate recognition and binding [101].

Another type of protein interaction is the binding to inorganic surfaces. This is relevant for different organisms that live attached to surfaces (e.g. mussels to mineral rocks [102]), but also plays a crucial role in nanobiotechnology and therefore simulating such protein–surface interactions also became an important research field [103, 104]. One main question for the interaction with a surface is if it influences the interaction with other proteins and therefore affects experimental results.

With the increasing computing power, the simulation of larger systems becomes possible. Therefore, the interaction of a protein with a membrane is also in focus of researchers. This is relevant to understand the functionality of membrane-bound proteins and especially transmembrane proteins. Examples for this special group of proteins are transmembrane receptors, ion channels, or membrane transport proteins. Cytochrome P450 is a membrane-bound protein and is relevant for drug metabolism and sterol biosynthesis. A detailed analysis of the functionality of Cytochrome P450 in the context of the membrane was done by Yu *et al.* [105] and helps to understand the whole drug metabolism process.

3.2 Protein Binding Pocket Alignment

The comparison of protein binding sites can be useful to understand the functionality of a protein for which no molecular functionality is known. In case the binding site has a high similarity to another protein for which, for example, binding partners are known, this can be helpful to understand the functionality and binding behavior. For this purpose, different algorithms are available to find and align protein binding pockets. A brief introduction to protein pocket comparison tools was given in section 1.5. Here, the focus is on one specific protein binding pocket alignment tool called ProBiS [43], which is integrated in a protein pocket comparison module in the LigDig webserver [99]. After a detailed description of the ProBiS algorithm, the LigDig webserver is briefly introduced with the main focus on the binding site alignment module.

3.2.1 ProBiS

The Protein Binding Site (ProBiS) [43] algorithm is designed to take advantage of locally similar three–dimensional patterns of physicochemical properties on the surface of a protein to detect similar protein binding sites. The great benefit of this tool is that it can find and also superimpose binding sites, which have a low sequence identity or a low global structural conservation. The ProBiS algorithm is available as a webserver and can also be downloaded and executed locally (http://probis.cmm.ki.si/).

The basic function of ProBiS is to take a query protein structure and compare it pairwise with other protein structures. For this purpose, a preparation step is required in which all residues on the protein surface are identified, based on an algorithm that finds solvent accessible surface atoms [106]. These residues are then represented as a graph of vertices and edges in a three–dimensional space. Each vertix replaces one functional group of a protein surface residue and has a label based on the following physicochemical classification: hydrogen bond acceptor (AC), hydrogen bond donor (DO), mixed acceptor/donor (ACDO), aromatic (PI) and aliphatic (AL) [47]. Afterwards, these graphs are used to produce a product graph for each pair of proteins. One vertex in this product graph represents a pair of vertices with identical physicochemical properties in the two compared graphs. Vertices in the product graph are connected with an edge if the corresponding vertices in the two initial graphs are positioned below a threshold distance apart (default <2 Å) [107]. The resulting product graph can be considered as a representation of all different rotations and translations of one protein graph onto the other. The maximum clique in the product graph represents the rotational and translational change that is required to superimpose the largest number of vertices in the two protein graphs onto each other. The two protein structures are superimposed by using the information stored in the product graph. For more details about the algorithm see, [43].

3.2.2 The LigDig Webserver

The LigDig webserver is designed to assist researchers to answer questions that previously required several independent queries to diverse data sources. Furthermore, it performs basic manipulations and analyses of the structures of protein–ligand complexes. The LigDig webserver is designed to be modular and consists

of seven tools, which can be used separately, or by linking the output from one tool to another, in order to answer more complex questions. These tools allow users to: (1) perform a free–text compound search, (2) search for suitable ligands, particularly inhibitors, of a protein and query their interaction network, (3) search for the likely function of a ligand, (4) perform a batch search for compound identifiers, (5) find structures of protein–ligand complexes, (6) compare three–dimensional structures of ligand binding sites and (7) prepare coordinate files of protein–ligand complexes for further calculations [99]. Figure 3.1 shows the homepage of the LigDig webserver with the seven different modules. In addition, a link to the external ProSAT$^+$ webserver is also included on the front page, as it provides a service that is related to the overall functionality of the LigDig webserver. LigDig is freely available and the source code can be downloaded here: http://mcm.h-its.org/ligdig. The integration of the ProSAT$^+$ feature and the implementation of the binding site based superposition of protein structures (module six) were developed in context of the presented work and therefore described in more detail here.

Module number six assists users to compare the three–dimensional structures of ligand binding sites and for this purpose the previously described ProBiS algorithm is applied. The user can enter a list of at least two PDB IDs. The corresponding protein structures are downloaded from the RCSB website and information about the number of chains, residues, and bound ligands are extracted. This information is listed in a table and the user can select the query protein or the query binding site by selecting a bound ligand. Afterwards the other protein structures are superimposed on the query protein by using the ProBiS algorithm. The resulting aligned structures and their highlighted binding site residues and ligands are viewed online using the JSMol viewer (www.sourceforge.net/projects/jsmol). In Figure 3.2, an example result and the visualization in the LigDig webserver is shown.

Based on the modular design of the LigDig webserver, the binding site comparison tool could be linked with module number five that assists users to search for structures of protein–ligand complexes. This enables a user to search for structures binding to a certain ligand and afterwards to select several protein structures to align their binding sites. For this purpose, the structures that were selected by the user are automatically forwarded to the binding site comparison tool. With these

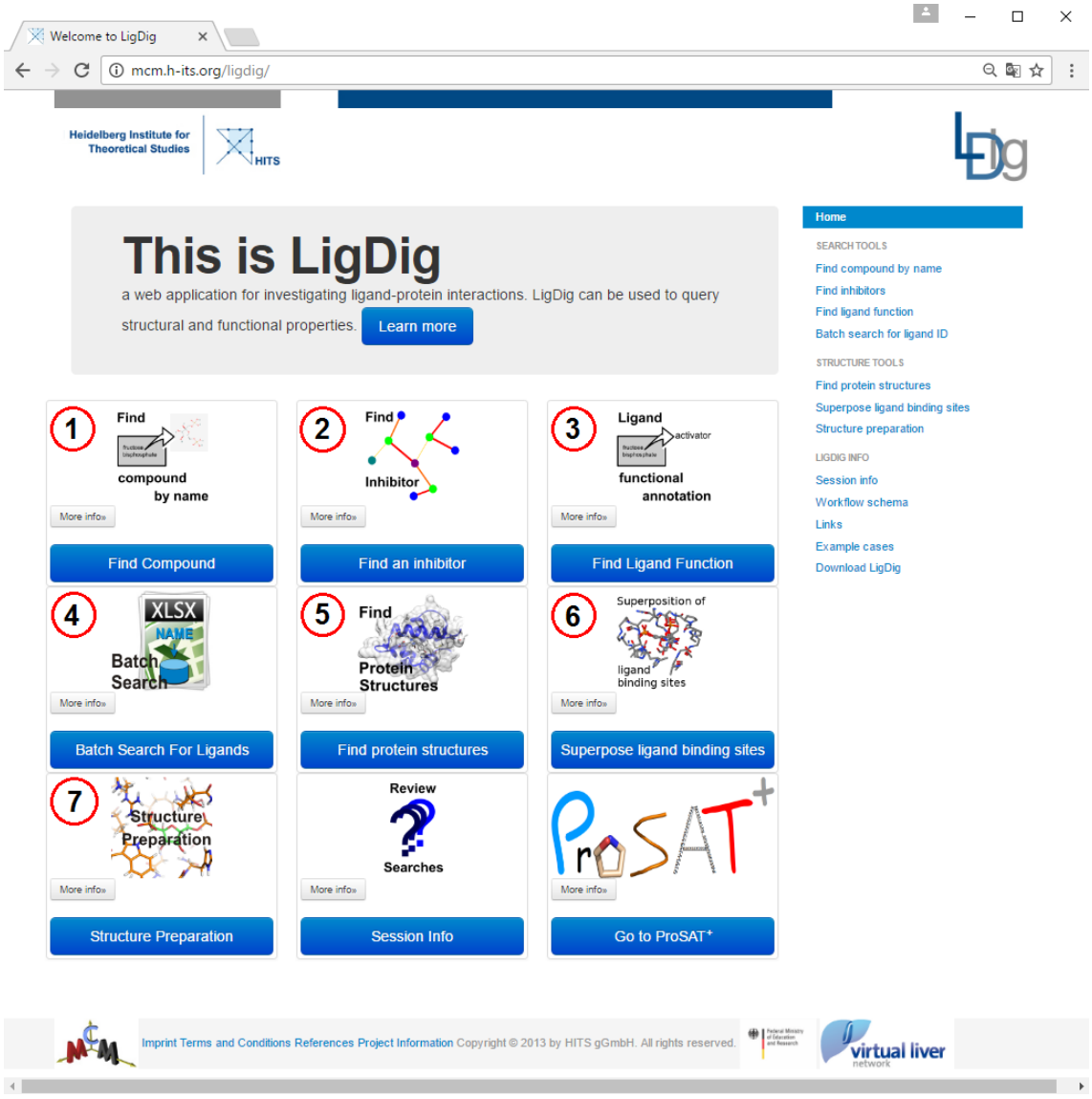

Figure 3.1: Homepage of the LigDig webserver showing the seven modules (labeled in red circles (1–7)) and the direct link to the ProSAT$^+$ webserver. Source: http://mcm.h-its.org/ligdig

combined features it is easy to compare binding sites regarding similar physico-chemical features, or detect new, unknown binding sites in proteins.

As described in Chapter 2, the ProSAT$^+$ webserver allows the visualization of user annotated residues in the context of known sequence annotations on a protein structure by building a specific URL. This feature is used and connected with the binding site comparison tool in the LigDig webserver. The residues of the proteins, which were used to superimpose the binding sites onto the reference structure are extracted and included in an URL. The user can click for each structure on this

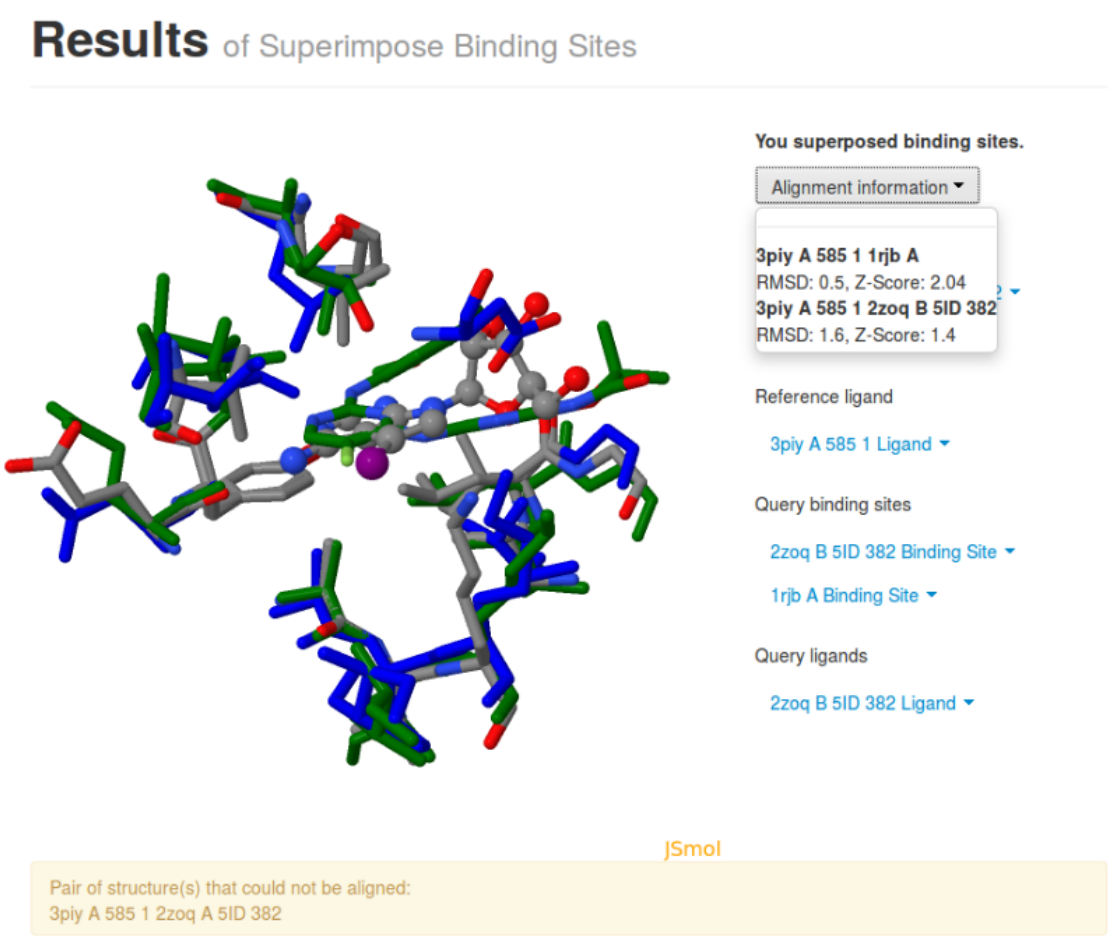

Figure 3.2: Example visualization of the LigDig webserver module number six to compare protein binding sites, based on the ProBiS [43] algorithm. In the dropdown list of the binding sites for each structure, a specific ProSAT$^+$ session link including the binding site residues and the PDB ID is provided (not shown).

link in the dropdown list (see blue links in Figure 3.2) and is then forwarded to a ProSAT$^+$ session showing the respective binding site residues. Now, the user can check if sequence annotations are known for these binding site residues or maybe for residues located near by. In addition, this allows to display the residues with similar physicochemical properties in different structures (because of the applied structures in the ProBiS algorithm) in the context of known sequence annotations. Further details about the ProSAT$^+$ webserver and the applied data are described and discussed in Chapter 2.

3.3 Protein Pocket Dynamics

This section is taken and adapted from reference [21]. I wrote the article and the content was discussed with all other authors who also reviewed the manuscript. Daria B. Kokh designed the example cases for the five classes of pockets and contributed to the drawing of the figures, as labeled.

The dynamics of protein binding pockets are crucial for their interaction specificity. Structural flexibility allows proteins to adapt to their individual molecular binding partners and facilitates the binding process. This implies the necessity to consider protein internal motion in determining and predicting binding properties and in designing new binders. Although accounting for protein dynamics presents a challenge for computational approaches, it expands the structural and physicochemical space for compound design and thus offers the prospect of improved binding specificity and selectivity.

A cavity on the surface or in the interior of a protein that possesses suitable properties for binding a ligand is usually referred to as a binding pocket. The set of amino acid residues around a binding pocket determines its physicochemical characteristics and, together with its shape and location in a protein, defines its functionality. Residues outside the binding site can also have a long-range effect on the properties of the binding pocket. Cavities with similar functionalities are often conserved across protein families. For example, enzyme active sites are usually concave surfaces that present amino acid residues in a suitable configuration for binding low molecular weight compounds. Macromolecular binding pockets, on the other hand, are located on the protein surface and are often shallower. The mobility of proteins allows the opening, closing, and adaptation of binding pockets to regulate binding processes and specific protein functionalities. For example, channels and tunnels can exist permanently or transiently to transport compounds to and from a binding site. The influence of protein flexibility on binding pockets can vary from small changes to an already existent pocket to the formation of a completely new pocket.

3.3.1 Five Classes of Protein Pocket Dynamics

A new classification of protein binding pocket dynamics with five classes was introduced by Stank *et al.* [21]: subpocket, adjacent pocket, breathing motion,

channel/tunnel, and allosteric pocket. These classes are illustrated schematically in Figure 3.3 relative to a reference binding pocket. The five classes are distinct but nonexclusive, meaning that overlaps between these classes are possible. An allosteric pocket, for example, may also be in close vicinity to the reference pocket and, thus, considered as an adjacent pocket. An important feature of an adjacent pocket is that it is positioned such that one bivalent ligand could bind in both the adjacent and the reference pockets simultaneously. Breathing motion refers to the enlargement or contraction of the original pocket, roughly retaining the original pocket shape. Such breathing motion may precede subsequent motions to form a distinct subpocket.

Examples of each of the five classes are shown in Figure 3.4. The first case, subpocket formation (Figure 3.4A), is illustrated by the binding site of the N–terminal ATP–binding domain of HSP90. It is lined by an unstable α–helix3 (blue) that undergoes distortion, converting to two short helices connected by a loop (orange). The inset shows a purine–based inhibitor that occupies both the ADP/ATP binding site and a hydrophobic transient subpocket formed under α–helix3. Figure 3.4B shows an adjacent pocket in interleukin 2 (IL–2), which has a highly adaptive protei–protein binding site that can be blocked by a small molecule. Arkin *et al.* [108] detected a flexible hydrophobic subpocket on the surface adjacent to the protein–protein binding site, which provides an additional space for binding a small molecule inhibitor [109]. This adjacent binding site is formed due to side–chain rotation accompanied by backbone adaptation. Pocket breathing motion is illustrated in Figure 3.4C for the B–cell lymphoma–extra–large (BCL–XL) protein, where considerable variation of the binding pocket shape is caused by movement of the α–helices lining the binding site. A set of NMR structures of nonspecific lipid transfer protein (ns–LTP) in complex with prostaglandin B2 demonstrates high plasticity of the hydrophobic binding pocket, including the opening and closing of a channel that enables ligand binding, as shown in Figure 3.4D for two models from the NMR ensemble [110]. Figure 3.4E shows the formation of an allosteric pocket in P38 mitogen–activated protein kinase (P38 MAPK) due to motion of the highly conserved Asp–Phe–Gly motif [111]. Opening of the allosteric binding pocket requires flipping of the Phe side–chain toward the ATP/ADP binding site. This reduces the volume of this binding site, and thus, binding of an inhibitor at the allosteric site inhibits ADP/ATP binding.

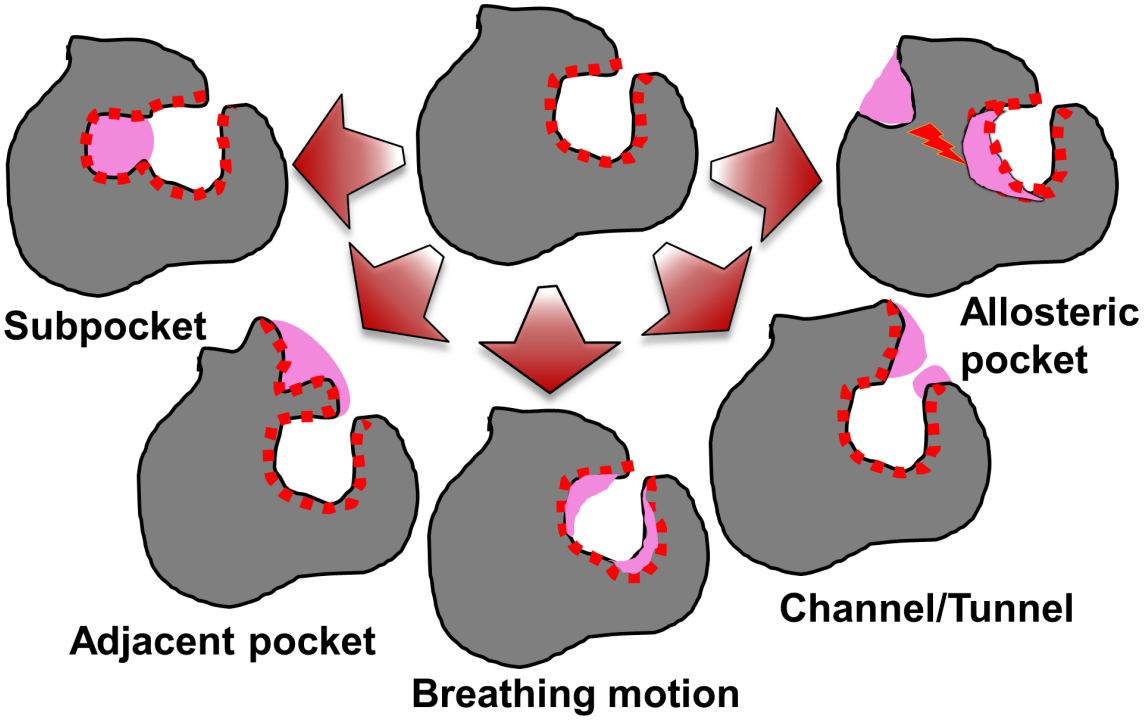

Figure 3.3: Cartoon representation of five different classes of pocket dynamics: subpocket, adjacent pocket, breathing motion, channel/tunnel, allosteric pocket. Regions colored in pink indicate pocket variation relative to the reference structure (shown in the center); the red dotted lines show the pocket shapes. For allostery, the shape of the original binding site is affected by a molecule binding at a distinct binding site. Figure was published in reference [21] (open access article under an ACS AuthorChoice License) and designed together with the authors and drawn by Daria B. Kokh.

3.3.2 Effect of Protein Binding Pocket Dynamics on the Thermodynamics and Kinetics of Ligand Binding

Target flexibility is one of the main factors that affects receptor–drug binding kinetics (see reviews by Pan *et al.*[120], Romanowska *et al.* [121], and Klebe *et al.* [122]). This subsection focuses on the influence of protein dynamics on the thermodynamics and kinetics of ligand binding in the context of the different binding pocket classes. Protein–ligand binding is often described by the one– or two–step models illustrated in Figure 3.5. These models are the most commonly used, but other models with more barriers or with a downhill binding free energy landscape are sometimes applicable. The observed rate constants depend on the kinetics of all the steps involved in the binding/dissociation process [121]. The free energy barriers to binding and unbinding may in part arise from the conformational rearrangement of the ligand and the protein. In particular, a barrier can be asso-

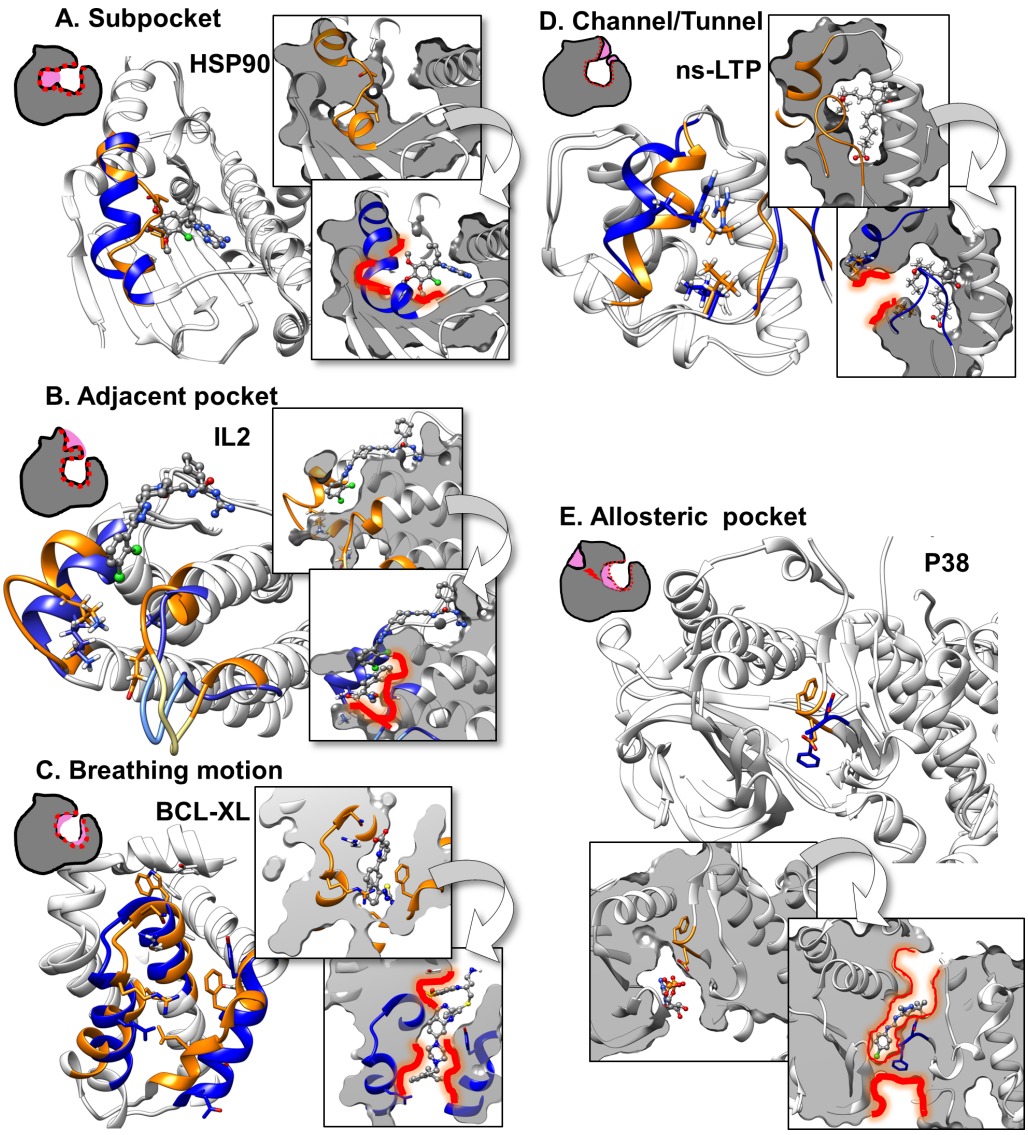

Figure 3.4: Examples of protein binding sites that illustrate the five different classes of pocket dynamics represented in Figure 3.3. For each case, two structures with different pocket conformations are shown in cartoon representation with flexible elements responsible for pocket changes shown in orange and blue. The binding pocket variations are visualized in the insets in cross sections of the protein structures going through the pocket of interest. Protein interiors are shown in gray. Transient opening of protein cavities is highlighted in red. In A, an unstable part of α–helix3 is shown in blue. In B, a flexible loop (Gln74–Leu80 shown in light blue and light orange) is missing in the crystal structures and was modeled using PRIME software (Schrödinger LLC, version 4.1).[112, 113] In 3.4E, the flipping Phe and Asp residues are shown in orange (open active site pocket with ADP bound) and blue (occupied allosteric pocket and blocked ADP–binding pocket). The structures (protein name, PDB ID) are (A) HSP90, 1yer,[114] 1uyd;[115] (B) IL–2, 1pw6, [116] 1m4a; [108] (C) BCL–XL, 3zln, [117] 3qkd; [118] (D) ns–LTP, 1cz2 (models 2 and 8); [110] (E) P38 MAPK, 1kv1, [111] 1ny3; [119] highlighted in orange and blue, respectively. See text for details. Figure was designed together with the authors and drawn by Daria B. Kokh and published in reference [21] (open access article under an ACS AuthorChoice License).

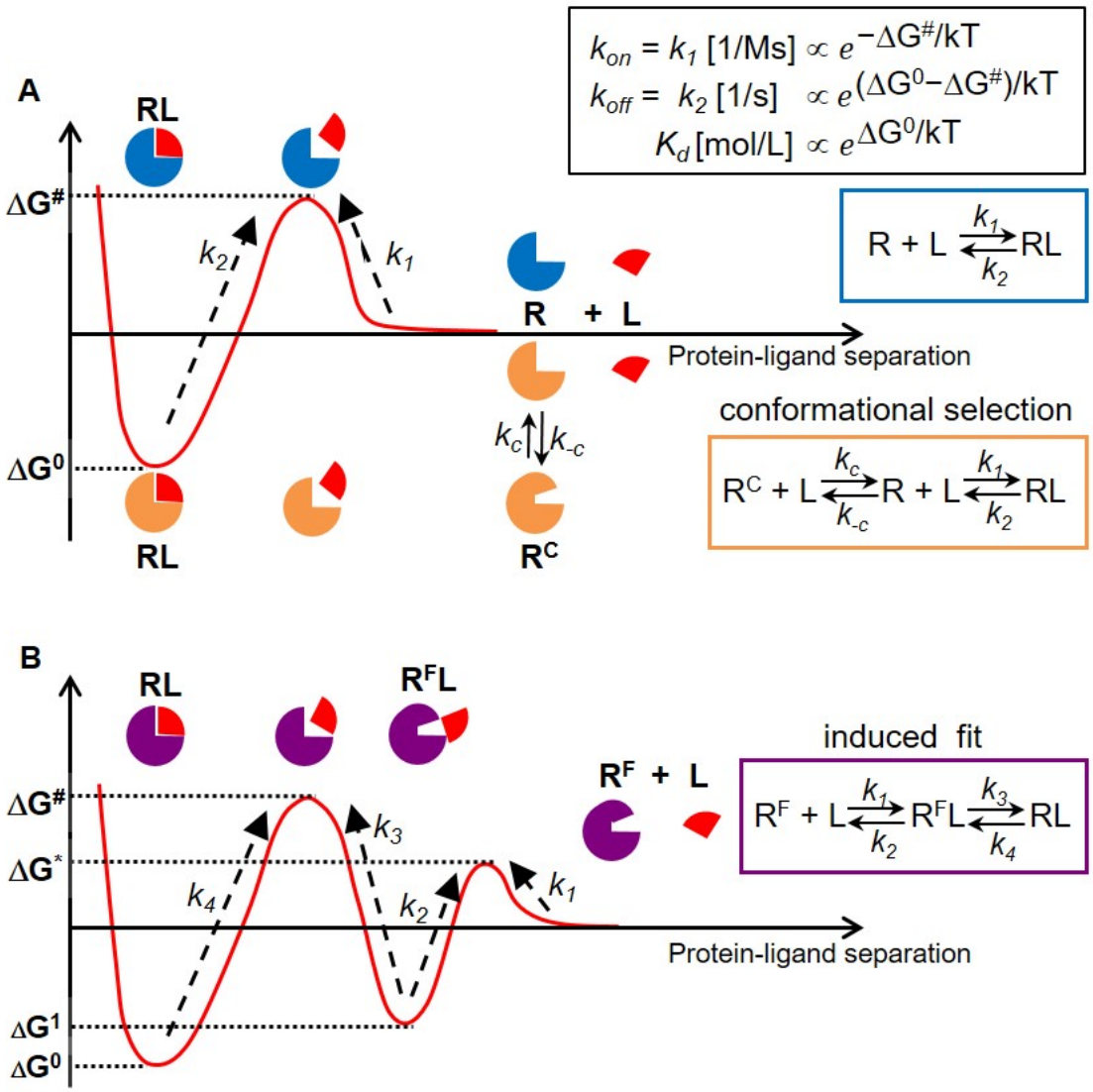

Figure 3.5: (A) Schematic illustration of free energy profiles for protein–ligand binding. Binding free energy and transition state energy are denoted by ΔG^0 and $\Delta G^{\#}$, respectively. k_{on} and k_{off} are association and dissociation rate constants; K_d is the equilibrium dissociation constant for receptor R and ligand L in the one–step binding model with or without conformational selection. R^C describes the conformational ensemble of the receptor, and k_c and k_{-c} describe the corresponding transition rate constants. (B) Two–step binding model with induced fit. R^F describes the receptor conformation in the free, unbound form and $k_1 - k_4$ indicate the respective rate constants for transitions between the states. Figure was published in reference [21] (open access article under an ACS AuthorChoice License) and designed together with the authors.

ciated with stochastic protein motion causing closing and opening of the pocket itself (e.g., breathing motion) or the pocket entrance (in the case of a channel or tunnel). These cases can be considered with a conformation selection binding mechanism, and the association kinetics can be described using a gating model

[28, 123, 28]. This model describes how the binding rate is modulated by the relation between the time scale of accessibility (or gating) of the binding site and the time scale for ligand binding. The frequency of gate opening and the fraction of time the gate is open are influenced by the type of structural dynamics necessary to open the gate. The closed state is often ascribed to a lid motion that blocks ligand entrance. Movement of a lid can be energetically expensive, explaining the long residence times of many ligands bound through a gating mechanism (see reviews [33] and [124]). For example, the rate of enzyme–inhibitor complex formation of Mycobacterium tuberculosis enoyl–ACP–reductase (InhA) was found to correlate with motion of the substrate binding loop [125]. In the two–step model (Figure 3.4B), a weakly bound transient complex (an encounter complex) is formed after passing the first barrier to binding, which increases the chance of a ligand to bind to a specific protein pocket that only opens stochastically or enables slow conformational changes required for the ligand binding (by induced fit). For example, binding of a bivalent compound to an adjacent pocket may proceed in two steps: first to the original pocket and then to the adjacent one due to induced fit.

Channel and breathing pocket dynamics often fall into the category of gating or conformational selection processes, whereas the induced–fit mechanism often plays a leading role in the formation of small subpockets. On the other hand, ligand–induced stabilization of breathing motions can alter binding kinetics and thermodynamics, as was observed for the protein kinase inhibitor Gleevec, which showed different binding kinetics to Abl and Src driven by its different ability to stabilize the P–loop [126]. The case of an allosteric pocket is the most complex, as both the allosteric and orthosteric binding sites can be considered to be gated, and the binding rates of the two pockets are not independent of each other. Ligand binding in the allosteric pocket can increase (activate) or decrease (inhibit) the k_{on} of the ligand in the orthosteric binding site and vice versa for k_{off}.

3.3.3 Detection of Transient Binding Pockets

Considering multiple conformations of a protein can increase the accuracy of ligand docking calculations and enable the detection of novel binding pockets. In addition, it can give insights into the kinetics of ligand binding or transition channel opening. Experiments may not be able to access all the conformations that could affect compound selectivity. Computational methods to simulate pocket dynamics can fill these gaps. Here, different computational sampling methods and

pocket analysis tools applicable for detecting the five classes of pocket dynamics depicted in Figure 3.3 are discussed.

3.3.3.1 Algorithms for Sampling of Protein Pocket Conformations

MD simulation is often used to explore variations in protein pocket shape and physicochemical properties due to protein dynamics. For example, Eyrisch and Helms applied 10 ns MD simulations to several proteins, including BCL–XL and IL–2, and successfully identified transient pockets [127]. Although MD simulation can be used as a sampling method for all of the pocket classes shown in Figure 3.3, it is computationally expensive, and the binding pocket dynamics may not be adequately sampled during the simulation time. Therefore, other methods that are more computationally efficient, although less accurate, have been used to study protein binding pocket dynamics arising from large–scale protein motions. In particular, normal mode analysis (NMA) provides a means to quickly explore the possible motion of a protein around a given input structure and is particularly useful for low–frequency interdomain harmonic breathing motions. It may be applied to atomic–detail molecular mechanics models or to coarse–grained elastic network models. For example, Ahmed *et al.* used a normal mode–based geometric simulation approach on an adenylate kinase structure to generate a pathway of conformational changes that describe domain movements leading to binding site closure [128]. However, NMA may not be appropriate when higher frequency or anharmonic motions have an important influence on binding pocket dynamics. Another method for exploring protein mobility is tCONCOORD [129], which performs geometric constraint–based sampling of protein conformations. This method enables fast sampling of large–scale motions (such as loop or domain motions). Ashford *et al.* [130] showed that the ensemble of BCL–XL structures generated by tCONCOORD revealed a transient binding subpocket, and Seeliger and de Groot [131] were able to generate transitions from apo–to–holo conformations for several proteins displaying interdomain pocket breathing motions using a combination of tCONCOORD and MD refinement. The framework rigidity optimized dynamic algorithm, FRODA [132], is another sampling algorithm for examining the internal mobility of proteins in a short time while respecting the stereochemistry and defined constraints. Metz *et al.* applied FRODA to sample hydrophobic transient pockets in IL–2 with better results than a comparable MD simulation, which shows the applicability to the subpocket and adjacent pocket

classes [133]. A disadvantage of the latter methods is that the generated structures are not energetically validated, and additional MD equilibration is generally required before using them further, for example, in a ligand docking procedure.

3.3.3.2 Protein Binding Pocket Analysis Algorithms

Simulations or experiments may provide a large ensemble of protein structures with a variety of binding pocket conformations. To distinguish a suitable pocket for ligand docking, the analysis of pocket dynamics in multiple structures is required. MDPocket [134] is one of the first methods designed for analysis of an ensemble of structures or MD trajectories. It provides the user with a pocket frequency map for visualization of pocket opening and for tracing some of the characteristics (e.g., pocket volume and accessible surface area) of a selected cavity along an MD trajectory. In another approach, EPOS^{BP} [127], protein cavities defined by pocket lining atoms are clustered to identify conserved or transient pocket regions. A deficiency of this approach is that the numbering of subpockets changes from snapshot to snapshot, hindering analysis of the pocket dynamics. Principal component analysis of pocket shapes mapped on a three–dimensional grid is used in the PocketAnalyzer [135] approach aimed at detection of transient pocket regions in MD trajectories. In TRAPP (TRAnsient binding Pockets in Proteins) [136], conserved and transient regions are defined with respect to a starting reference structure. TRAPP can be applied to the detection of larger conformational changes, such as might be revealed by a tCONCOORD or NMA calculation. This is achieved by using only binding site residues for structure alignment together with a robust algorithm for pocket detection that does not require parameter adjustment when going from buried to shallow cavities. The new TRAPP webserver [100] allows the user to have easy access to the TRAPP algorithm and is described in the next section. Unlike previously discussed methods, where pocket shapes are used for the analysis of pocket dynamics, the PPIAnalyzer method [133] clusters protein conformations with respect to the RMSD of heavy atoms, followed by the identification of transient binding sites using the PocketAnalyzer program.

All of the methods described above are in principle applicable to exploring the formation of adjacent pockets, subpockets, and pocket breathing motions. Detecting allosteric pockets and analyzing the structural and dynamic relationship between allosteric and orthosteric pockets poses further requirements. Methods that reveal energy transport or networks in proteins that might relate to allosteric

effects have been reported, such as SPACER [137] and MCPath [138]. Furthermore, the FRODA method has been applied for sampling structures to reveal the relationship between allosteric and orthosteric pockets [132]. For channel detection, further tools are available. One example is Caver 3.0 [40], which uses a Voronoi diagram to describe tunnels in a MD trajectory. Other tools are trj_cavity [139], Cavity analysis [140], and MOLE [141], which employ MD trajectories to generate conformations for analysis.

3.4 TRAPP Webserver

This section is taken, adapted, and extended from reference [100].

The analysis of protein binding pocket dynamics is an important step in the analysis of protein function and therefore also in designing new drugs. The new TRAnsient Pockets in Proteins (TRAPP) webserver provides an automated workflow that allows users to explore the dynamics of a protein binding site and to detect pockets or subpockets that may transiently open due to protein internal motion. These transient or cryptic subpockets may be of interest in the design and optimization of small molecule inhibitors for a protein target of interest. The TRAPP workflow consists of the following three modules: (i) *TRAPP Structure* – generation of an ensemble of structures using one or more of four possible molecular simulation methods (ii) *TRAPP Analysis* – superposition and clustering of the binding site conformations either in an ensemble of structures generated in step (i) or in PDB structures or trajectories uploaded by the user; and (iii) *TRAPP Pocket* – detection, analysis, and visualization of the binding pocket dynamics and characteristics, such as volume, solvent–exposed area and properties of surrounding residues. A standard sequence conservation score per residue or a differential score per residue, for comparing on– and off–targets, can be calculated and displayed on the binding pocket for an uploaded multiple sequence alignment file, and known protein sequence annotations can be displayed simultaneously. An abstract graphical representation of the general functionality of the freely available webserver at http://trapp.h-its.org is shown in Figure 3.6.

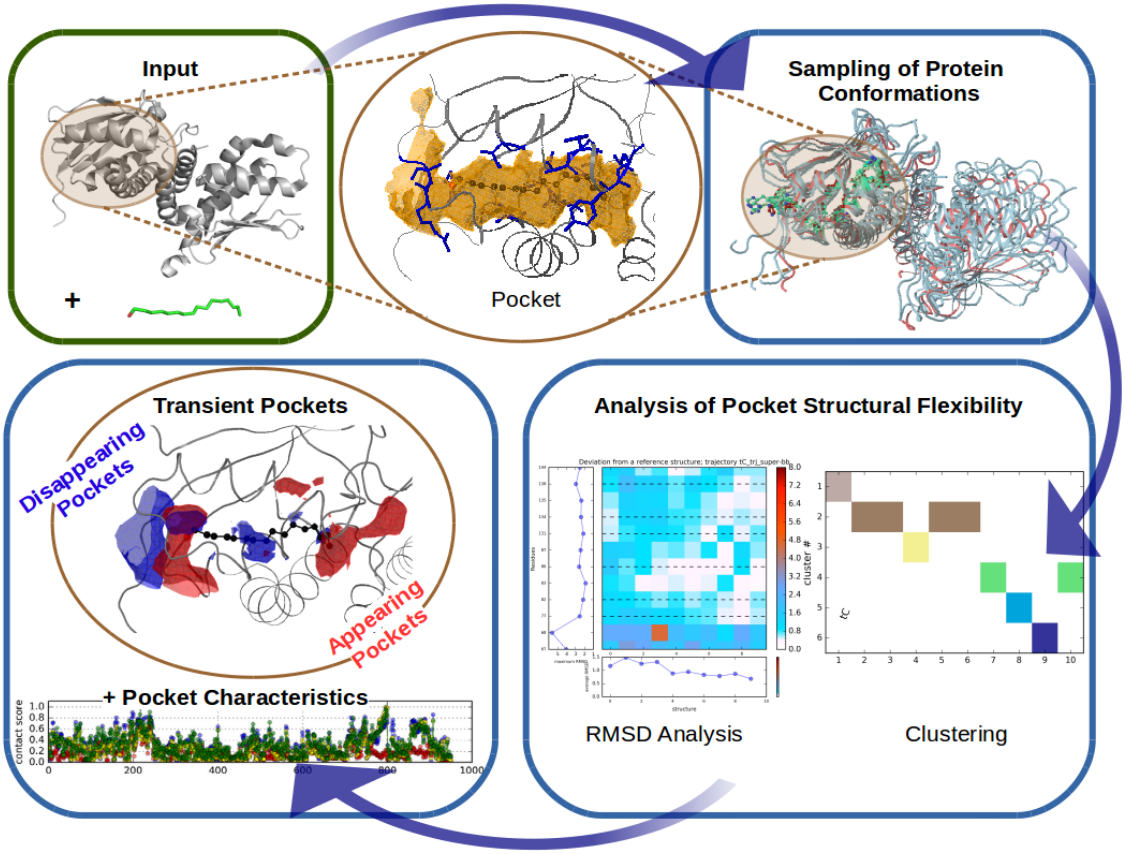

Figure 3.6: Graphical representation of the general functionallity of the TRAPP webserver. Figure is taken from submitted manuscript [100].

3.4.1 Introduction

Protein flexibility plays a key role in molecular recognition but is often neglected in protein structure–based drug design projects. Thus, transient or cryptic pockets that are not visible in available protein crystal structures but may bind ligands are missed. Computational approaches to identify transient binding pockets or subpockets provide a means to reveal druggable pockets and to expand the possibilities for improving the specificity and diversity of designed compounds. Indeed, consideration of protein binding pocket dynamics has played an important role in drug discovery [21]. For example, consideration of pocket dynamics in p38 mitogen–activated protein kinase helped to find an inhibitor [111]. Another example is the identification of a cryptic pocket in HIV integrase, adjacent to the known active site, in molecular dynamics (MD) simulations [142]. This pocket was exploited in the discovery of HIV integrase inhibitors, leading to the development of the drug Raltegravir [143]. MD simulations have also revealed a tran-

sient and potentially druggable binding pocket at the dimeric interface of HIV–1 protease [144].

Most often, ligand design is undertaken for known binding sites rather than novel sites. The known binding sites may be sites where natural ligands bind or where existent drugs (or active compounds) bind. Therefore, the TRAPP webserver was designed as a tool for studying the dynamics of a known binding pocket or any other protein cavity of interest and for identifying and characterizing transient subpockets. These transient subpockets may be considered for ligand design and optimization, and therefore the TRAPP webserver provides information on their physicochemical and sequence properties, as well as shape and dynamics.

A range of computational tools for detecting binding pockets on protein structures and analysing their static structures is available (see https://bioinformatic tools.wordpress.com/tag/pocket-finder/ and the recent review by Zheng *et al.* [145]). Several of them, trj_cavity [139], MDpocket [134] based on FPocket [146, 147], EPOCK [148], PocketAnalyzerPCA [135], POVME 2.0 [149], EPOSBP [127], as well as the previous version of TRAPP web server [136], are designed for analysis of binding pocket dynamics in MD simulation trajectories. Some of them also provide pocket characteristics, such as the volume or hydrophobicity, and analysis of changes in pocket shape (as defined by pocket occupancy [149], principle component analysis of the pocket motion [135] or clustering of pocket regions [127]). They all, however, require the user to provide and, in some cases to align and superimpose, snapshots of the trajectory to be analyzed. The trajectory is usually a standard MD trajectory but these are typically too short to sample all the cryptic pockets that are potentially interesting for drug design. Indeed, protein binding pocket dynamics are dependent on a wide variety of motions ranging from small side–chain vibrations and rotations to large–scale changes in secondary structure and domain movements. Complete sampling of such motions is difficult to achieve in standard MD simulations. On the other hand, several approaches have been developed that enable protein flexibility to be explored in a relatively short simulation time [129, 150]. In particular, it has been demonstrated that the opening of cryptic pockets appearing on the microsecond time scale can be revealed in several nanosecond L–RIP simulations [150]. These simulations, though less accurate than standard MD, give the user insight into the transient hot spots of the binding pocket. Thus, the main motivation of the present development of the TRAPP webserver was to provide the user with an automated workflow

to invoke a toolbox of methods to explore pocket flexibility arising from protein conformational changes on a wide range of temporal and spatial scales. The current TRAPP webserver provides a choice of several methods to efficiently generate conformational ensembles representing binding pocket dynamics, in addition to the option of uploading conformational ensembles, e.g. crystal structures or MD trajectories, generated by other tools. Furthermore, the integrated alignment procedure makes it possible to compare binding pockets in proteins from different species, e.g. with gaps or insertions, or with mutated residues. Additionally, the TRAPP webserver provides new tools to analyse the geometric, dynamic and physicochemical properties of the transient pockets, along with protein sequence conservation and differential conservation between on–targets and an off–target. The present TRAPP webserver also allows simultaneous visualization of residue–based mutations by integration of queries with the ProSAT$^+$ webserver [82]. Thus, the TRAPP webserver provides capabilities for a wide–ranging interactive analysis of pocket dynamics for transient subpocket detection and characterization on the basis of a single input protein structure. Here, the structure and implementation of the TRAPP webserver, the user input and results produced are described. The use of the TRAPP webserver is illustrated by means of an example application to a protein target for anti–parasitic drug design.

3.4.2 Applied Methodologies and Workflow

The overall workflow of the TRAPP webserver consists of three modules, as illustrated in Figure 3.7: (i) *TRAPP Structure*, (ii) *TRAPP Analysis*, and (iii) *TRAPP Pocket*. For each of these steps, the user can select and define additional simulation parameters in the user interface. After an initial user input, *TRAPP Structure* and *TRAPP Analysis* are started sequentially. The *TRAPP Pocket* module is run after further user input. Here, short descriptions of the applied methods and the simulation results are provided. A detailed documentation of each module, input parameters and output data, as well as several usage examples are available on the TRAPP webserver.

Initially, the user is required to upload a reference protein structure in PDB format, and define the center of the binding pocket of interest. The latter is specified either by uploading a PDB file containing the coordinates of a ligand, or by manually defining the coordinates of the center of the pocket together with the distance within which protein residues are considered as binding site residues. They are

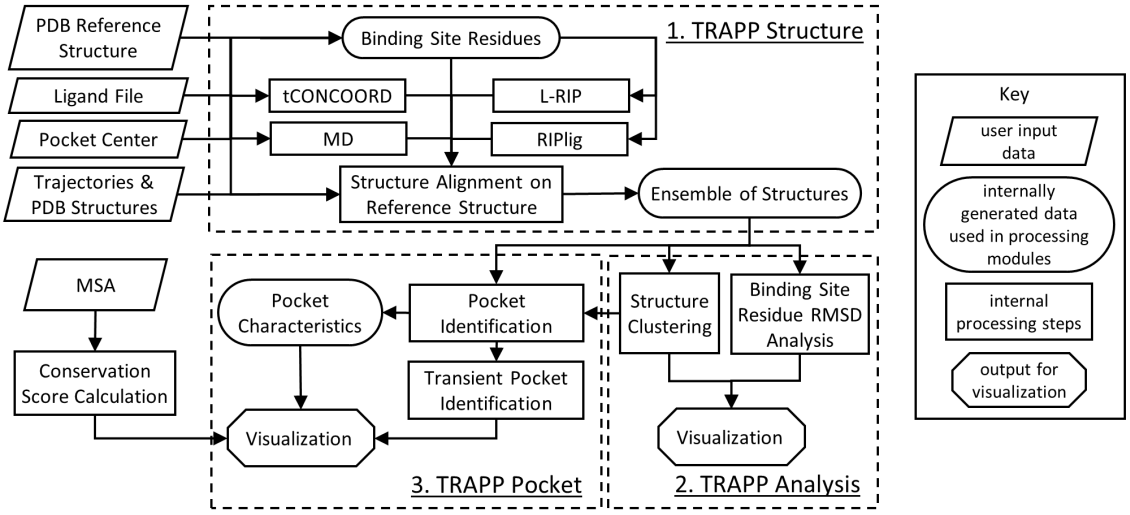

Figure 3.7: Workflow of the TRAPP webserver. Figure is taken from submitted manuscript [100].

used later for structure alignment and superposition. Further, the user can select one or several simulation methods to be used in the *TRAPP Structure* module for exploring protein conformations, or upload trajectories or PDB files. Optionally, the user may provide a multiple sequence alignment (MSA) file in FASTA format to analyse the sequence conservation of the binding pocket.

3.4.2.1 Functionality and Output of the Workflow Modules

The aim of the *TRAPP Structure* module is to enable fast generation of ensembles of protein structures that represent the conformational diversity of the binding pocket. The user can choose to use one or more of the following methods to generate an ensemble of structures: short implicit solvent MD simulation [151], tCONCOORD [129], L–RIP and RIPlig [150]. All the methods are described in detail in the respective publications. The short MD simulations mainly allow for the sampling of side chain movements, whereas the MD–based perturbation approaches, L–RIP and RIPlig, are useful for sampling larger scale motions of the binding pocket, including backbone motions, domain motions and changes of secondary structure elements. tCONCOORD provides a constraint–based sampling methodology for side–chain and loop motions, but can also be useful for exploring slow domain movements [131]. The generated trajectories or ensembles of

snapshots are directly transferred to the *TRAPP Analysis* module for clustering and analysis of protein flexibility.

The *TRAPP Analysis* module provides tools for the comparison of the binding pockets in the protein structures uploaded or generated in the previous step, *TRAPP Structure*. To this end, all structures are aligned and superimposed using the backbone atoms of the binding site. For clustering of the binding site conformations, one of two different metrics can be selected: RMSD of backbone atoms (default) or RMSD of geometric centre of the non–hydrogen atoms of the binding site residues. By default, a simplified single–linkage hierarchical clustering algorithm [136] is used with a RMSD threshold of 3 Å. The clustering procedure can be refined by using a smaller threshold and/or *K*–means clustering [136]. The set of binding site residues can be increased or reduced at this step. A binding site residue RMSD matrix for all structures, the RMSD averaged over all binding site residues for each structure, and the maximum RMSD of each binding site residue in all structures are calculated and plotted. All three–dimensional structures of the cluster representatives are displayed together with the reference structure in a JSmol applet.

The *TRAPP Pocket* module tracks pockets in the protein structures and identifies transient regions. Protein cavities located at the binding site are calculated for each structure as described in reference [136] and stored on a grid. Additionally, the physicochemical properties of the side chains that directly contact the binding cavities (e.g., positive/negative charge, H–bond donors/acceptors), pocket volume, and surface area are calculated and displayed. Furthermore, the pocket lining residues are analyzed for all structures and compared graphically to those in the reference structure. The pocket dynamics are analyzed and transient regions (those that either appear or disappear relative to the reference structure provided as an input by user) and conserved regions (defined by the percentage presence in snapshots or structures) are identified. Appearing and disappearing regions are displayed by isocontours in JSmol along with plots, showing in which structure or snapshot they occur. Transient pocket regions occurring in 25% and 50% (default values that can be altered by the user) of the structures are split into compact sub–regions and displayed. The extent of the opening of each of the transient subpockets is computed for each snapshot and plotted.

If the user uploads a MSA, a conservation score is calculated per residue using a python–based tool [152] that applies the Jensen–Shannon divergence score.

This allows a quantification of the similarity between probability distributions, by taking a background amino acid distribution (BLOSUM62) into account. This score is rescaled to the range between 30 and 70, mapped on the protein structure, and displayed by colour (blue: low conservation, red: high conservation) together with the *TRAPP Pocket* results. Another option allows the user to upload a MSA file containing a set of protein sequences representing an on–target group and one sequence representing an off–target. The conservation score is calculated twice: once with the MSA containing the off–target sequence, and once without it. Afterwards, the absolute difference of these two conservation scores per residue is calculated, rescaled, and mapped on the three–dimensional structure. In this case, blue highlights residues with high similarity between on– and off–target sequences and red indicates residues with low similarity. The MSA is displayed with MSAviewer [155] and the sequence conservation can be shown by colour coding the reference structure. The TRAPP webserver also provides a coupled visualization of residue–based annotations in ProSAT$^+$. More details about the sequence conservation feature and the integrated ProSAT$^+$ service are provided in the following section.

All data and graphics generated by the *TRAPP Analysis* and *TRAPP Pocket* modules can be downloaded as raw data or as a compressed archive. Additionally, a PyMOL session is generated and can be downloaded with other data as a single archive for in–depth visual inspection of the protein structures and pocket analyses.

3.4.3 Sequence Analysis in the Context of Protein Pocket Dynamics

The sequence conservation visualization feature and the integrated ProSAT$^+$ service in the TRAPP webserver is unique, as it provides the user the possibility to combine the protein pocket dynamics analysis with evolutionary and sequence based information. This combination of data makes sense, as changes in sequence can have an influence on the function and/or the dynamics of a pocket and therefore on the whole protein. In addition, the evolutionary data provide information about the evolution of protein, meaning the change of a certain amino acid in one species might explain a different level of activity or binding affinity of the protein.

The integrated sequence annotation analysis feature is provided by the service of the previously described ProSAT$^+$ webserver (see Chapter 2). The sequence annotations mapped on the three–dimensional structure can be inspected in parallel to the protein pocket dynamics results. An example of how the different data and results are combined and displayed to the user is discussed and shown in an example application case in Subsection 3.4.5.

3.4.3.1 Sequence Conservation

The TRAPP webserver offers the possibility to analyze the sequence conservation per residue together with the *TRAPP Pocket* results. The user can choose between an average conservation score or the difference in the conservation score with and without an off–target sequence. For this purpose, the user has to upload a multiple sequence alignment (MSA) in FASTA format. The FASTA format is a common file format to store (multiple) sequence information, including sequence alignments. The files begin with a single–line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a '>' symbol at the beginning and in the next line the sequence starts.

For calculating the conservation score for the MSA, the 'Protein Residue Conservation Prediction' tool by Capra *et al.* [152] using the Jensen–Shannon divergence is applied with default parameters except those listed in Table 3.1.

Table 3.1: List of the parameter values differing from the default values in the 'Protein Residue Conservation Prediction' tool by Capra *et al.* [152]. The gap cutoff is set to the maximum value to force the consideration of each alignment position, independent of the number of gaps. The window size parameter is changed only in the On–/Off–target option.

Parameter	Value	Function
-g	0.99	gap cutoff, scores columns that contain more than gap cutoff fraction gaps are not scored
-m	/matrix/blosum62.bla	similarity matrix file
-s	js_divergence	conservation estimation method
-w	0	window size (default=3)

To enable better visualization in JSmol, the conservation scores are re–scaled into the range between 30–70. This is done be extracting the minimum and maximum conservation score of all residues and applying the following formula:

$$f(x) = (((70 - 30)(x - min))/(max - min)) + 30 \tag{3.1}$$

To visualize the conservation scores, the scores are placed at the B–factor position in the reference structure file, which provides an easy way to show the level of conservation by color scheme. The applied color scheme can be seen here: http://jmol.sourceforge.net/jscolors/#Positional%20Variability.

Average Sequence Conservation
For running the average conservation analysis in TRAPP, the uploaded MSA in FASTA format is used to run the above mentioned conservation calculation tool that results in a conservation score (between 0–1) for each aligned position in the MSA. To map the conservation score on the three–dimensional reference structure, the sequence from the reference PDB file is extracted and aligned to the MSA with the MUSCLE tool [153, 154]. This new MSA is used to map the conservation scores to the correct residue in the reference structure.

Sequence Conservation between On–/Off–Target Sequences
This option allows the comparison of one protein sequence class with another single sequence regarding the conservation score. This is often useful when comparing on– and off– target proteins. In this case the user can upload a MSA in FASTA format including several on–target sequences and one off–target sequence. The user needs to define the off–target sequence by entering the FASTA Header name to the webserver interface. The conservation tool is applied two times: once with all sequences (on– and off–target sequences) and once without the off–target sequence. Afterwards, the absolute difference of the conservation score per position is calculated and re–scaled again between 30–70 as explained before.

Visualization
The results of the conservation calculations are displayed on the reference structure together with the *TRAPP Pocket* results in JSMol. For this purpose, the scores are added to the PDB file at the B–factor position and visualized via the temperature colouring option in the JSMol application. For the average conservation,

the colour gradient starts at blue (low conservation) and goes via white to red (high conservation). For the on–/off–target option, blue means no difference between the conservation scores with/without the Off target sequence. Red highlights residues with a high difference, meaning these residues are not conserved between the on– and off–target sequences. As the scores are re–scaled, the gradient starts with ice–blue and only goes up to pink. This is necessary for a better contrast between the isocontours visualizing opening and closing pockets and the conservation colouring. The conservation scores (or the differential scores) can be switched on by clicking a checkbox in the *TRAPP Pocket* result page.

Additionally, the MSA is visualized in the *TRAPP Pocket* result page. For this purpose, the MSAViewer [155] is used. Two additional lines are added to the MSA: the corresponding amino acid sequence in the uploaded reference PDB file and the residue numbers extracted from the reference PDB File. The user can double–click on the residue numbers to highlight them in the JSMol viewer.

3.4.3.2 Protein Sequence Annotations

The ProSAT$^+$ [82] service is integrated as an iframe into the final visualization of the *TRAPP Pocket* results. This allows the comparison of pocket dynamics simultaneously with protein sequence annotations mapped on the three–dimensional structure. The user can show and hide the ProSAT$^+$ frame with a button, which provides a convenient way to switch between the different analyses and visualized data.

In the beginning of a TRAPP analysis, the user can enter the corresponding PDB ID of the uploaded reference structure. This automatically creates a link to a ProSAT$^+$ session showing the annotation data for the corresponding protein. Otherwise, the PDB ID, Uniprot accession number, or protein sequence can be entered later to the ProSAT$^+$ interface when it is shown with the other TRAPP results.

The combined visualization of the TRAPP and ProSAT$^+$ results allows an easy analysis of specific binding site residues which are highly flexible or play a role in the pocket dynamics and for which a sequence annotation (e.g. mutagenesis site) is known. This enables the first interpretation of how the protein binding pocket dynamics influence the interaction with other binding partners.

3.4.4 Technical design of the TRAPP webserver

The TRAPP webserver is implemented in Java and uses the Play framework 2.5. For the HTML front–end, the Twitter Bootstrap (version 3.3.7) is used to provide a clear layout of the user interface. All visualisations of the three–dimensional structures and pocket information employ the Java Script–based JSmol (https://sourceforge.net/projects/jsmol/). This enables a Java–free visualization in most common web browsers, which has the advantage that no Java installation is required on the client site and associated security issues can be avoided. The TRAPP modules are implemented in Python and Fortran [136]. After each submission of a TRAPP job, the user is provided with a session identifier link, which can optionally be sent by email. This session link allows the user to monitor the status of their current job and access their results. This session link, which is only provided to the user who started the analysis, also allows the user to share their results with other people. The runtime of the submitted jobs depends on which method is selected in the *TRAPP Structure* module and can vary between a few minutes and several hours. Submitted jobs are automatically distributed on a compute cluster. The user can select to receive a notification of job completion by email. The TRAPP results are stored for a limited period during which they can be accessed and downloaded.

3.4.5 Example Application Case

Dihydrofolate reductase (DHFR) is a key enzyme in the folate pathway, and therefore in DNA synthesis. It is a known target for anti–cancer antifolate drugs, but antifolates are also being investigated as potential drugs against human trypanosomatid parasites [156]. However, the development of selective anti–parasitic antifolates without side effects is hindered by the fact that binding pocket of trypanosomatid DHFR is very similar to that of human DHFR (hDHFR). Therefore, it is essential to identify DHFR binding pocket features that could differentiate parasitic variants (on–targets) from the human one (off–target). Here, an example of how this problem can be addressed by analysis of the binding site flexibility with the TRAPP webserver is provided and visualized in Figure 3.8.

There are several crystal structures available for the DHFR of Trypanosoma cruzi, a human parasite causing Chagas' disease. Thus, an analysis of the conformational variability of the binding site of Trypanosoma cruzi DHFR (TcDHFR) is

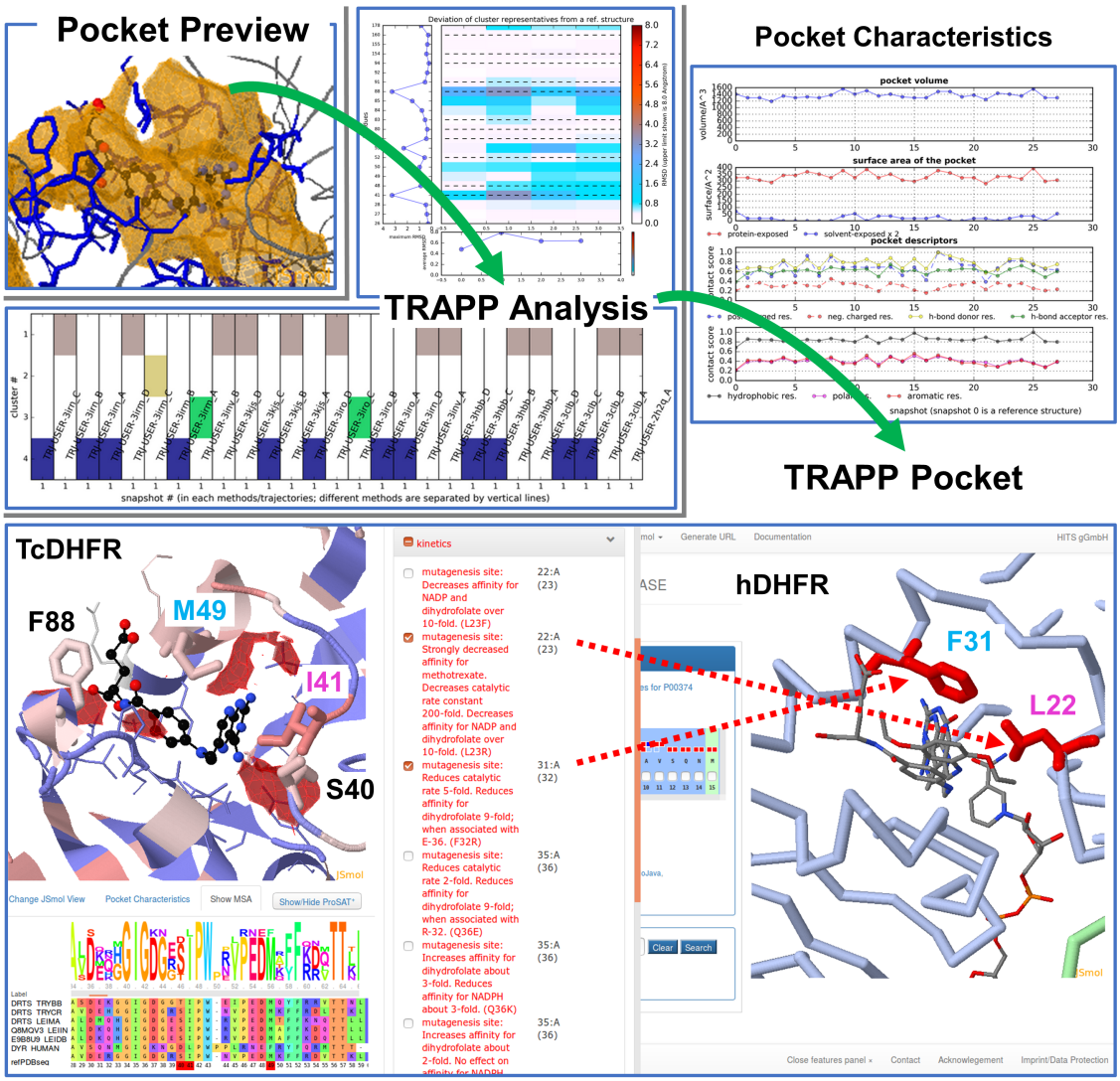

Figure 3.8: Screenshots from an example application of the TRAPP webserver to the analysis of the binding site of a parasite enzyme to identify transient subpockets that are selective with respect to the human homologue. The binding pocket is shown by orange isocontours in the reference structure of TcDHFR (upper left). For the *TRAPP Analysis* module (upper middle), the clustering of structures and RMSD binding site analysis of the four cluster representatives is shown. In the *TRAPP Pocket* results (bottom right), pocket characteristics are shown for all analyzed structures (upper right). Appearing subpockets are shown by red isocontours (bottom) with close residues that are not conserved between the on– and off–targets (TcDHFR and hDHFR) labelled. Residues I41 and M49 in TcDHFR correspond to L22 and F31 in hDHFR, whose mutation affects inhibitor binding and catalytic activity in hDHFR (bottom right). See text for details. Figure is adapted from the one in the submitted manuscript [100].

done, and afterwards its binding site sequence is compared with that of the off–target protein, hDHFR. By including the DHFR sequences of other representative trypanosomatid species (other potential on–targets) in the alignment, amino acid residues conserved in trypanosomatids can be identified.

The 9 crystal structures available for TcDHFR are (resolution and number of chains in parentheses): 2h2q (2.4 Å, 2), 3cl9 (3.3 Å, 1), 3clb (3.0 Å, 4), 3hbb (3.0 Å, 4), 3kjs (2.5 Å, 4), 3inv (2.4 Å, 2), 3irm (2.1 Å, 4), 3irn (2.6 Å, 4), 3iro (2.8 Å, 4). To prepare these structures for TRAPP, the PDB files are split into 29 files, each containing a single protein chain with only ATOM records retained. (Chain B of 2h2q is discarded because of many missing residues around the binding site). The PDB file 3cl9 is assigned as the reference structure, the bound ligand (methotrexate) is extracted into a separate ligand file in PDB format, and is used to define the binding site (5 Å radius). To examine the on–/off–target selectivity, a multiple sequence alignment (MSA) in FASTA format is generated for five trypanosomatid DHFR sequences, to represent the on–target group (UniProt: Q27793, Q27783, P07382, Q8MQV3, E9B8U9), and for the off–target hDHFR (UniProt: P00374).

With these input data prepared, the *TRAPP Analysis* can be started. Screenshots that illustrate the analysis and results are shown in Figure 3.8. After uploading the prepared data into the user interface, a summary of all parameters, a preview of the pocket in JSmol (see Figure 3.8), and a list of all defined binding site residues can be inspected. The *TRAPP Analysis* results show the RMSD values of the binding site residues for each uploaded structure to those in the reference structure. In addition, the RMSD values for the four cluster representatives (k–means, 1.5 Å threshold using geometric centers of side–chains) are shown together in one plot (Figure 3.8). This analysis reveals that the greatest structural variability is at the binding site residues Ile41, Arg53, Pro85 and Phe88.

Next, the *TRAPP Pocket* with the default parameter values can be run. In the summary of the results, a checkbox to color the reference structure according to the per–residue differential sequence conservation can be selected (blue - conserved, red - not conserved between the on–targets and off–target). In the dropdown list for appearing pockets, the threshold is set to 50% to show in the JSmol applet structure view red isocontours for the regions that appear in 50% of all the uploaded structures compared to the reference structure. An opening of several small subpockets near Ser40/Ile41, Met49, and Phe88 can be observed. The displayed MSA shows full conservation of amino acid positions 41, 49, and 88, and partial conservation of position 40, between the representative trypanosomatids (Figure 3.8), and the coloring of the structure shows that these positions differ from the corresponding residues in hDHFR. Notably, these residues are characterized by differing size and/or charge in parasite vs. human DHFR (Met49 vs. Phe,

Phe88 vs. Asn and Ser/Thr40 vs. Asp). The integrated ProSAT$^+$ tool further shows that the residues identified as flexible, non–conserved, and close to an appearing subpocket, Ile41 (Leu22 in hDHFR, both labelled in magenta in Figure 3.8) and Met49 (Phe31 in hDHFR, both labelled in cyan in Figure 3.8), are known mutagenesis sites in hDHFR (Leu22 to Arg and Phe31 to Arg) that affect the binding affinity of the anti–cancer drug methotrexate [157] or the catalytic activity [158].

In summary, this quick analysis with the TRAPP webserver highlights transient pocket regions and residues in the structure that can provide starting points for the design of selective drugs targeting parasite DHFR but not human DHFR. In fact, Schormann *et al.* could design an inhibitor about 14 times more active against TcDHFR than hDHFR (in terms of K_i value) by targeting the region close to Met49 [159].

3.5 Conclusion and Discussion

This section contains text adapted from references [100] and [21].

This chapter highlighted the importance of considering protein pocket dynamics especially in the drug design process. New webservers were introduced that provide different functionalities to analyze protein binding pockets regarding their similarity, dynamics, conservation and known sequence annotations. The introduced classification of protein binding pocket dynamics into five classes – subpocket, adjacent pocket, breathing motion, channel/tunnel, and allosteric pocket – assists to distinguish different kinds of pockets.

The LigDig webserver and its binding site comparison tool allows to find protein interaction sites of different and similar overall protein structures. Based on similar physicochemical features, similar binding sites are detected and the three–dimensional structures are aligned. Together with the linked ProSAT$^+$ webserver features, it assists users in the initial analysis of protein functionality and protein binding sites.

A range of computational methods is available for simulating and analyzing protein pocket dynamics. However, the treatment of protein dynamics in structure–based drug design projects is generally very limited, and better computational tools are required for studying binding pocket dynamics for this purpose. An automated tool that combines simulations of protein mobility using a variety of ap-

proaches with the analysis of pocket dynamics would help users to choose the most appropriate strategy for designing new inhibitors. The assessment of the druggability of a protein pocket should include measures of how the physicochemical properties of a pocket are affected by the pocket dynamics derived from analysis of simulations and NMR or crystal structures. Treatment of protein pocket dynamics may increase the space of potential binders, e.g., through ligand binding in transient subpockets. However, it may also provide a basis for identifying specific compounds that can bind to the structurally conserved parts of a pocket despite its mobility. Therefore, the consideration of the class of protein binding pocket dynamics can facilitate the choice of appropriate computational approaches to sample and analyze the pocket dynamics as well as ligand design strategies.

The new TRAPP webserver provides several functionalities, which actually assist the user with different computational approaches to sample protein pocket structures, and also in the post analysis of the structural data. It is a platform to facilitate exploration of the dynamics of a protein binding site, as well as variations in pocket physicochemical properties and detection of transient pockets and subpockets. Conformational variations of the binding site may arise due to protein internal flexibility ranging from side–chain rotations to large–scale conformational changes in the secondary and tertiary structure. For this purpose, an ensemble of different experimental or simulated protein structures is analyzed, which can be either provided by the user or generated by using several provided methods. Additionally, the TRAPP webserver provides an analysis of the binding site residue conservation and known sequence annotations in the context of transient binding pocket analysis results. The TRAPP webserver has a modular structure, which currently provides four methods to generate conformational ensembles and is designed so that further methods can be added in the future. A challenge for the webserver design is the fact that some of these methods are computationally demanding, meaning that, despite parallelization, job run times prevent uninterrupted interactive use of the TRAPP webserver for an analysis. Future improvements in the computing abilities of the back–end, increases in compute cluster size, or linking of the TRAPP webserver to cloud computing resources will increase the capacity and speed of the TRAPP webserver and ameliorate these issues. The modular design of the TRAPP webserver also permits the addition of further pocket analysis features, e.g. a subpocket specific druggability score or a pocket–ligand similarity score. Both features would provide additional quanti-

tative data to aid the design of more specific compounds that take advantage of certain subpockets and their biochemical features. Other features that would increase the usability of the TRAPP webserver include the possibility of an automatic download of the reference protein structure by PDB ID and a post–processing to prepare the structure and the potentially bound ligand. This post–processing step would require, similarly to what is done in the binding site comparison tool in the LigDig server, to show the user structural information and let the user select which ligand to use to define the binding site. Another useful feature would be to automatically download all available structures for a specific protein, for example by using a UniProt accession number. These structures could be superimposed and merged to a trajectory and would provide an easy and fast way to use already available structural data that include protein dynamics information.

To further improve the usability of the TRAPP webserver, one could change the protein structure visualization by using the new and faster NGL viewer [79], but this would require a complete re–writing of all visualization scripts and the displayed data would need to be compatible with the viewer. Another improvement would be to combine the visualized data of the TRAPP and ProSAT$^+$ webservers. The current and main problem for this is the data exchange between the two webservers, especially the mapping of the data to the correct residues. In case the user uploads a protein structure which has different residue numbers or residues, it becomes difficult to map the sequence annotations from ProSAT$^+$ for a specific protein on the corresponding structure in the TRAPP webserver. Nevertheless, with additional sequence alignments, the transmission of the ProSAT$^+$ data (e.g. as a json file by a RESTful system) to the TRAPP webserver, and an adaption of the visualization interface would be possible. In case the previously mentioned PDB ID download option is implemented, this step would be much easier as the same structures in TRAPP and ProSAT$^+$ are guaranteed.

Overall, the current TRAPP webserver enables the analysis of the structural plasticity of a binding pocket, and the identification of transient pocket regions and residues important for selectively targeting a specific protein. Together with the sequence–based conservation analysis of on–target and off–target groups, it enables the identification of residues that distinguish the two groups, providing a basis for developing more selective drugs.

4

Analyzing the Role and Function of J-Proteins

Sections of this chapter are based on the following publication and manuscript:

Nadinath B. Nillegoda, Janine Kirstein, Anna Szlachcic, Mykhaylo Berynskyy, Antonia Stank, Florian Stengel, Kristin Arnsburg, Xuechao Gao, Annika Scior, Ruedi Aebersold, D. Lys Guilbride, Rebecca C. Wade, Richard I. Morimoto, Matthias P. Mayer and Bernd Bukau. Crucial HSP70 co-chaperone complex unlocks metazoan protein disaggregation. Nature, (2015) 524:247-251

Nadinath B. Nillegoda, Antonia Stank, Duccio Malinverni, Niels Alberts, Anna Szlachcic, Alessandro Barducci, Paolo De Los Rios, Rebecca C. Wade and Bernd Bukau. Evolution of an intricate J-protein network driving protein disaggregation by the Hsp70 chaperone machinery. (2017) Submitted.

Visualizations shown in some figures in this chapter are part of one of the above mentioned references and were initially done by myself.
Text sections taken and adapted from one of the above mentioned references are marked in the beginning of a section. This text was initially written by me and went through the review process of all authors.

4.1 Introduction

Protein aggregates can occur in stressed and ageing cells, and characterize several pathophysiological states [160, 161]. Usually, these protein aggregates are effectively eliminated in healthy metazoan cells [162, 163], which shows that an effective disaggregation process exists. In comparison to non–metazoan organisms, metazoans do not have the heat–shock protein disaggregase HSP100. HSP100 is an important protein in the HSP70–dependant disaggregation system that leads to high efficiency in non–metazoans [164, 165]. Even with the additional crucial HSP110 nucleotide exchange factor, the human HSP70 system shows a poor disaggregation activity *in vitro* [163, 166].

The HSP70 chaperones are highly relevant in protein folding and disaggregation processes. Their broad spectrum in function is achieved through diverse members of the J–protein (HSP40) co–chaperone family, which preselect substrates for the HSP70 interaction. These co–chaperone molecules regulate the ATP–dependent substrate binding and release cycle of HSP70 interacting chaperones. J–proteins have a J-domain (JD), which is named after the Dna**J** protein from *E. coli*, where it was first identified [167]. The JD contains a conserved HPD tripeptide as the signature motif for interaction with HSP70. Three classes of J-proteins (A, B, and C) with more than 50 members in humans exist. All three classes target HSP70 to substrates, but with some functional redundancy among the three classes [168]. With increasing organism complexity, the number of J–protein members increases.

For a long time, the relation between J–protein class and function was unknown. The discovery of complex formation between class A and class B J–protein members through transient interactions in metazoans changed the understanding of the role of J–proteins [169]. The mixing of human class A and B J-proteins leads to an increase in the disaggregation activity. This gain in protein disaggregation power through the interclass J–protein network provides the human HSP70–based disaggregase with a similarly efficient mechanism to the non–metazoan HSP100-HSP70 bichaperone disaggregase system [170]. The discovery of the J–protein network provides an explanation of the similar efficiency that is found in solubilization of amorphous aggregates in HSP100–lacking higher organisms. Recruiting multiple ATP dependent HSP70 proteins by complexed J–proteins leads to formation of an oligomeric chaperone complex. This molecular complex is proposed to

build up entropic pulling forces that allow an efficient extraction of trapped and aggregated proteins [170]. However, the evolutionary process and origin of this J–protein network is still unknown.

The work presented in this chapter, together with experimental results, provides data that give hints on an evolutionary conservation of J–protein interclass complexation, that is relvant for the efficiency in the disaggregation processes of eukaryotes [171]. Before achieving this overall result, the following questions had to be answered:

- *Is structural data available for J–proteins from different organisms to cover the evolutionary process, and if not can these structures be modeled?*
- *Experiments have shown evidence that electrostatics play a role in the J–protein interaction, is it possible to visually and quantitatively prove this?*
- *Can the experimental results, showing interactions between J–proteins, be proven by simulations?*
- *Can differences in the electrostatics be highlighted that might explain the complexation of eukaryotic, but not of prokaryotic J–proteins?*
- *Can electrostatic differences in two groups (canonical and non–canonical) of class B J-domains be observed and quantitatively analyzed?*

To answer these questions, different computational methodologies had to be applied. In the following, a more detailed description of how these questions were addressed is given.

Experimental results showed a synergy between class A and B J-proteins of humans, *C. elegans*, and yeast, that leads to an increase in protein disaggregation eficiency [169, 171]. Similar experiments with prokaryotic *E.coli* J-proteins showed no increase in the protein disaggregation activity when both J–protein classes were mixed [171]. This opens the questions of how the different J–protein classes interact with each other, and which biochemical differences in the J–proteins lead to the synergy and this highly efficient protein disaggregation activity. To investigate these complex questions, computational modeling and simulations were performed. First, the three–dimensional structures of the JDs and the CTDs of the J–proteins from human, fungi, nematodes and bacteria were modelled, because available structural data was limited. Only these two domains were modeled for

two main reasons. Firstly, these two domain are linked by a highly flexible linker region, which is difficult to resolve by crystallography, and therefore structural data is rarely available. Secondly, by using two separate domains it is possible two perform docking simulations and to analyze inter– and intramolecular interactions of J–proteins. Furthermore, the experiments have shown interactions between the JDs and CTDs, therefore is was reasonable to perform the docking simulations with these domains.

Considering the protein structures from different organisms, which can be grouped into pro– and eukaryotes, provides a meaningful way to analyze the possible evolutionary effects on the interclass J–protein interaction. The electrostatic potentials of all domains were computed, and Brownian dynamics simulations with the Simulation of Diffusional Association (SDA) software package [62, 63, 64] were performed to generate diffusional encounter complexes of human (class A (DNAJA1, DNAJA2) and class B (DNAJB1, DNAJB4)) and bacterial (*E.coli* DnaJ and CbpA) JDs and CTDs in detail, and compare them with the experimental results [169, 171]. In the next step, a local electrostatic potential comparison around the interaction sites identified by docking, and a clustering based on electrostatic similarity was performed with the Protein Interaction Property Similarity Analysis (PIPSA) tool [54, 55, 56] to quantitatively compare the the electrostatic potentials of the J–proteins. A subset of the J–proteins helped to identify clear differences between pro– and eukaryote J–proteins.

The human class B family of J–proteins contains canonical and non–canonical J–proteins [168]. In contrast to canonical J–proteins (e.g. DNAJB1 and DNAJB4), non-canonical J–proteins, for example, DNAJB2 and DNAJB8 are able to prevent protein aggregation of amyloidogenic proteins via ubiquitin–interacting motifs (UIMs) and serine–rich stretches located in the respective C–terminal domains (CTDs) [168, 172, 173, 174]. Even if all J–proteins have a JD, the overall three–dimensional structure can be very different, for example DNAJB8 occurs as a poly–dispersed oligomeric complex, instead of a dimer like DNAJB1. Experiments have revealed that, in comparison to DNAJB1, both DNAJB2 and DNAJB8 are incapable of reactivating aggregated luciferase, even when mixed with class A J–proteins [171]. The question of which differences present in canonical and non–canonical J–proteins, especially in the JDs, lead to this difference in function is addressed in this chapter by another PIPSA analysis of canonical and non–canonical class

B JDs. This helped to identify possibly relevant features of class B JDs, in their complexation with class A CTDs.

4.1.1 Three-Dimensional Structures of class A and B J-Proteins

Three classes of J–Proteins exist – A, B, and C. Here, only classes A and B will be considered. The class A and B J–proteins are homodimers whose monomers consist of two main domains; the C–terminal domain (CTD) and the N–terminal J–domain (JD). Figure 4.1 displays representative structures of a class A (DNAJA2, green) and a class B (DNAJB1, blue) CTD. In the following, class A protein domains are always colored in green and class B domains in blue. Both classes contain CTD–I and CTD–II domains in each monomer. At the C–terminus of the CTD–II domain, both classes have a dimerization site that connects the two monomers to form a homodimer. In comparison to the class B CTDs, the class A CTDs contain an additional zinc–finger–like region (ZFLR), that is connected to the CTD–I domain. The class A CTD together with ZFLR provides substrate specificity [175, 176].

At the C–terminus of each JD, a highly flexible, disordered and G/F rich linker region (not shown in Figure 4.1) connects the JD with the CTD–I domain. In Figure 4.1, a representative structure of a JD (DNAJA2, human) is shown on the right. The three–dimensional structures of class A and B JDs are highly similar, which is why only one is shown. The JDs mainly consist of four α–helices (I–IV, see Figure 4.1) that are connected by short loop regions. At the end of helix II, a short, three residues long HPD motif can be found (marked in Figure 4.1). This motif is highly conserved in all J–protein structures, and is important for the HSP70 interaction.

The J–proteins are highly flexible, and in the dimer, the distance between the CTD monomers can vary, while the protein is only restrained bound at the dimerization domain at the bottom of the protein. The JD of each J-protein can move around, as the JD is only bound to the CTD by a flexible linker. The JDs can interact with their respective CTD, or another CTD, but this binding is transient and not permanent. For a more detailed overview of the structure and function of J-proteins, see [168, 177].

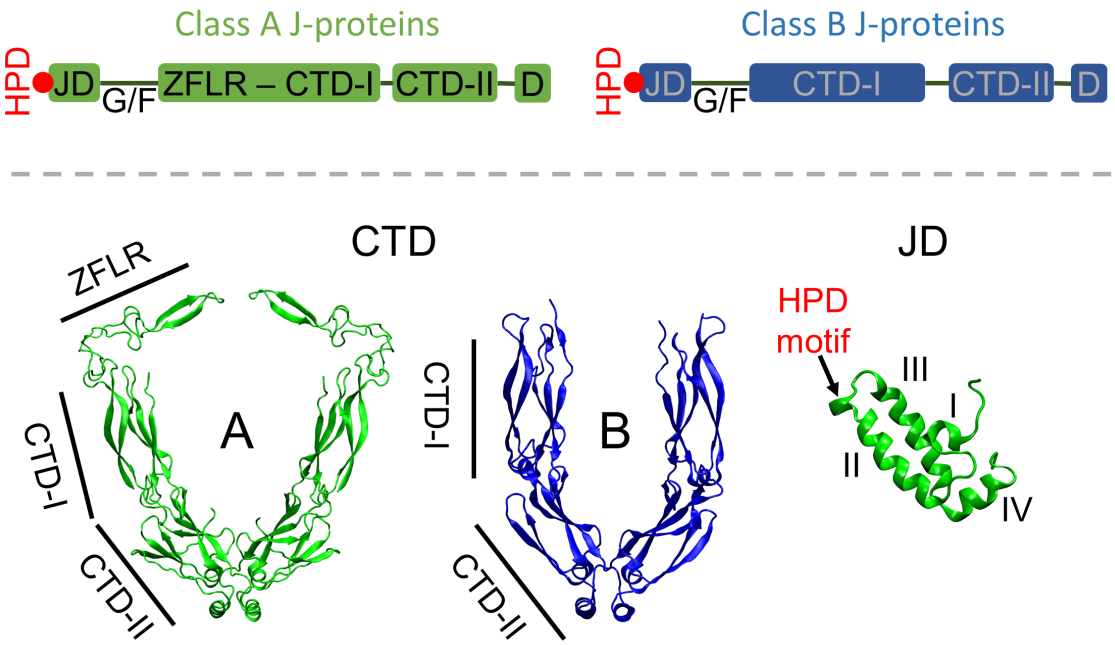

Figure 4.1: The topologies of class A (top left) and B (top right) J–proteins. Below this, the cartoon representations of the two distinct classes of J–proteins A (green) and B (blue) are displayed, showing the highly conserved domain organization. The HSP70–intertacting HPD motif is labeled in red in both representations and is located in the N–terminal J–domain (JD). The JD is connected by a Gly/Phe–rich flexible region (G/F) (not shown in cartoon representation) with the C–terminal β–sandwich domains (CTD–I and II). The monomers are connected via the dimerization site (D). Class A J–proteins are distinguished mainly by a zinc–finger–like region (ZFLR) that inserts into the CTD–I subdomain [168, 177]. The scaling of the CTDs and the JD is not the same to enable a bigger representation of the JD. The four helices (I-IV), similar in class A and B JDs, are labeled in the cartoon representation.

4.1.2 Methodological Challenges

Protein–protein docking simulations are an established procedure in the molecular modeling field. Nevertheless, the docking of J–domains to the respective CTD is a special challenge. A protein–protein docking simulation with the whole structures of the J–proteins was not possible. Reasons for this include the missing G/F linker region between the JD and CTD, and the fact the CTDs are highly flexible, meaning rigid body docking of this huge J–proteins would not have been able to simulate the CTD interaction with the JD, which is flexibly bound to the CTD.

Experiments have shown evidence that electrostatics play a role in the interactions of the J–proteins [169]. This fact suggested the application of the SDA software, which considers the electrostatics of simulated proteins, and additionally provides the ability to incorporate constraints. This feature of adding constraints to the simulation was helpful in one specific simulation set–up. Previously, the SDA

software has been applied to globular macromolecules, but the three–dimensional structure of the CTD homodimer has an approximate scale of 110 Å x 100 Å x 30 Å and an overall V–shaped structure. Additionally, J–proteins are highly flexible and therefore can have a huge variety of structural conformations. This made appropriate simulations of the whole, flexible J–proteins in the SDA simulations difficult and therefore required the JD–CTD docking simulations. This simulation set–up allowed the inclusion of some kind of flexibility, to represent the missing G/F linker region. Another uncertainty was whether the quality of the many modeled protein structures was good enough for these types of simulations to reproduce the experimental results.

To analyze the docking results, the existing clustering workflow in SDA required modification to handle the challenging analysis of this docking data. This modified clustering procedure is explained in detail in Section 4.4.2. However, together with the experimental results the overall methodology that was applied could be verified and tested for this challenging class of molecules.

4.2 J-Protein Structural Modeling

This section is taken and adapted from reference [169].

For the computational analysis of the J–proteins, three-dimensional structures of class A and B J–proteins from different organisms are required. Therefore, the necessary protein structures were either crystal or NMR structures taken from the RCSB Protein Data Bank (PDB; http://www.rcsb.org) or comparative models that are either present in the SWISS–MODEL database and can be found using the Protein Model Portal (PMP, www.proteinmodelportal.org) or were modelled with SWISS–MODEL (SM) [178, 179, 180, 19]. Tables 4.1 and 4.2 give an overview of all structures, and the corresponding organism and UniProt accession numbers. A fully annotated list with details about the modeling procedure and the PDB ID can be found in Tables 6.1 and 6.2 in the Appendix.

The structure of the CTD DNAJB1 dimer was taken from the crystal structure (PDB ID 3agz, resolution: 2.51 Å) [181] and that of JD DNAJB1 from the first entry of the NMR structure (PDB ID 1hdj) [182]. Since the N–terminus of chain B in the CTD DNAJB1 dimer is missing three residues compared to chain A, the N–terminal nine residues from chain A were superimposed on the N–terminus of

Table 4.1: List of analyzed class A J–protein structures. For DNAJA4 only the CTD was modeled and analyzed.

Name	Organism	UniProt Accession Number
DNAJA1	*H. sapiens*	P31689
DNAJA2	*H. sapiens*	O60884
DNAJA4	*H. sapiens*	Q8WW22
Ydj1	*S. cerevisiae*	P25491
DNJ–12	*C. elegans*	O45502
DnaJ	*E. coli*	P08622
DnaJ	*S. typhi*	P0A1G8
DnaJ	*K. pneumoniae*	C4T9C4
DnaJ	*P. oryzihabitans*	A0A0D7F716
DnaJ	*B. pertusis*	Q7VVY3
DnaJ	*A. aceti*	A0A063X4A7
DnaJ	*Sphingomonas sp.*	Q1NCH5
DnaJ	*C. ultunense*	M1ZGL1
ATJ3	*A. thaliana*	Q94AW8

Table 4.2: List of analyzed class B J-protein structures. The last three J–proteins are only analyzed based on their JDs. DNAJB1^{RRR} is a triple mutant of DNAJB1 (D4R, E69R, E70R) that fails to interact with opposite class A CTDs [169].

Name	Organism	UniProt Accession Number
DNAJB1	*H. sapiens*	P25685
DNAJB4	*H. sapiens*	Q9UDY4
Sis1	*S. cerevisiae*	P25294
DNJ–13	*C. elegans*	Q20774
CbpA	*E. coli*	P36659
CbpA	*S. typhi*	P63262
CbpA	*K. pneumoniae*	W9BQH2
CbpA	*P. oryzihabitans*	A0A0D7FE35
CbpA	*B. pertusis*	J7RE62
CbpA	*A. aceti*	A0A063XA16
CbpA	*Sphingomonas sp.*	Q1NEX3
CbpA	*C. ultunense*	M1ZLZ3
At5g25530	*A. thaliana*	F4JY55
DNAJB2	*H. sapiens*	P25686
DNAJB8	*H. sapiens*	Q8NHS0
DNAJB1^{RRR}	*H. sapiens*	P25685, D4R, E69R, E70R

chain B to obtain coordinates for the missing three residues. Comparative models of the CTD DNAJB4 dimer and JD DNAJB4 were both found in the PMP and are based on the template structures 3agz and 1hdj, respectively. To add the missing three residues at the N–terminus of chain B, the same procedure as for CTD DNAJB1 was used.

Comparative models of the CTD monomers of DNAJA1 and DNAJA2 were both taken from the PMP. Both structures were modelled with SM based on the template crystal structure, 1nlt (resolution 2.70 Å)[175]. The structure of JD DNAJA1 is the first entry of the NMR structure, 2lo1 in the RCSB PDB. A comparative model of JD DNAJA2 was taken from the PMP and is based on the same template structure, 2lo1. Structures of the CTD dimer were generated for DNAJA1 and DNAJA2 as follows: the dimerization site was modelled with SM based on the template crystal structure, 1xao (resolution 2.07 Å) [183]. Then, the structures of the CTD monomers were superimposed on the corresponding dimerization site model and only the C–terminal missing residues of the dimerization site were added to the CTD domains. The structure of the J–domain of DNJ–12 was taken from the crystal structure PDB ID 2och (resolution 1.86 Å). The CTD dimer of DNJ–12 was modeled based on the crystal structure, 1nlt, using SM. The J–domain of DNJ–13 was modeled based on 1hdj and the dimer structure of the CTD of DNJ–13 was based on 3agz A/B. Further editing of the following structures was performed to generate a set of comparable structures of the J–proteins. The N–terminal Gly was deleted in JD DNAJA1, because it is not part of the UniProt entry P31689. The last seven residues of JD DNAJB1 were deleted to have a comparable C–terminal end to the JDs of DNAJA1 and DNAJA2. Similarly, the last four C–terminal residues in JD DNAJB4 were deleted to obtain comparable C–terminal ends to the JDs of DNAJA1 and DNAJA2.

For the following proteins, three dimensional structures are available and were used: J–domains of Sis1 (PDB ID: 4rwu, resolution 1.25 Å, [184]), *E. coli* class A (PDB ID: 1xbl, [185]), *E. coli* class B (PDB ID: 3ucs, resolution 1.87 Å), human DNAJB8 (PDB ID: 2dmx), human DNAJB2 (PDB ID: 2lgw, [186]) and CTD of Sis1 (PDB ID: 1c3g, resolution 2.7 Å, [187]) and Ydj1 (PDB ID: 1nlt, 1xao). The structure of the CTD dimer of Ydj1 was built, similarly to DNAJA1 and DNAJA2, using the structure of the CTD monomer (PDB ID: 1nlt), which was superimposed twice on a crystal structure containing the dimerization site (PDB ID: 1xao) by using PyMOL (http://www.pymol.org). For the remainder of the proteins studied, no

crystal or NMR structure was available. Therefore, three–dimensional structures of the domains of these proteins were built by comparative modeling using the SM webserver (http://swissmodel.expasy.org) [18]. For all class A CTD models, the modelled CTD dimer of Ydj1 was used as a template structure and the two Zn^{2+} ions were transferred afterwards by superimposing the structures with PyMOL. For the class A CTD of *Pseudomonas oryzihabitans* (Uniprot accession: A0A0D7F716), a less conserved loop close to the Zn^{2+} binding region was modeled with different backbone coordinates from the template structure but these prohibit realistic Zn^{2+} binding because of a too large binding distance. Therefore, three residues were changed in the sequence to force the SM algorithm to model the same backbone coordinates as in the template structure (KIIPEP –> DIIKDP). Afterwards, the three residues in the model structure were back–mutated using the mutagenesis tool in the PyMOL software. This model was then used as a template structure in the SM webserver to slightly adapt the side chains in the mutated region. Only in the case of the *Sphingomonas sp* (strain SKA58) DnaJ (UniProt accession: Q1NCH5), is the model of *Acetobacter aceti 1023* DnaJ (UniProt accession: A0A063X4A7), which was built using the Ydj1 model, used as a template because the less conserved loop around the Zn^{2+} binding region was modeled better for Zn^{2+} ion binding than when the Ydj1 model was used. In the case of the *Bordetella pertussis* (UniProt accession: Q7VVY3) class A CTD, the sequence alignment was manually adapted to enable the modeling of the C–terminal region. For this purpose, a multiple sequence alignment of the four gamma and the beta bacterial sequences and the yeast sequence (template structure) was considered using the software DeepView [178]. A DeepView project with the adapted alignment was uploaded to the Swiss–Model webserver.

For the class B CTDs, the Swiss–Model webserver was used to find a template structure and, if multiple templates were found, the one with the highest sequence identity to the target structure was chosen and then, in the case of more than one structure for this sequence, the corresponding structure with the highest QMEAN4 score. For the following class B CTDs (UniProt accession numbers), the template structure 3lz8.B (resolution 2.9 Å) was used: P36659, P63262, W9BQH2, J7RE62, F4JY55. The PDB ID 3lz8.A was used as a template structure for the following class B CTDs (UniProt accession numbers): A0A0D7FE35, Q1NEX3, M1ZLZ3, O75953. For the Type B CTD of A0A063XA16, the structure with the PDB ID 4j80.A (resolution 2.9 Å, [188]) was used as a template. The dimer structure of Sis1 was built

by superimposing the crystal structure of the monomer (PDB ID: 1c3g) twice on the 19 C–terminal residues of the crystal structure of the DNAJB1 dimer (PDB ID: 3agz).

For the class A J–domain of *Acetobacter aceti 1023* DnaJ (UniProt accession: A0A063X4A7), the structure with the PDB ID 4j80 was chosen, and for the *Sphingomonas sp* (strain SKA58) DnaJ (UniProt accession: Q1NCH5) and the ATJ3 (UniProt accession: Q94AW8), the structure with PDB ID 4rwu was chosen as the template structure. In the case of DNAJA4 (UniProt accession: Q8WW22), the structure with PDB ID 2lo1 was taken as the template. For all other class A J–domains, the structure with the PDB ID 1xbl from *E. coli* was used as the template. For the class B J–domains, the following templates were used (UniProt accession: PDB ID of template structure): A0A063XA16:4j7z, Q1NEX3:2dmx, M1ZLZ3:2yua, F4JY55(At5g25530):2m6y, O75953(DNAJB5):4wb7. For all other class B J–domains, the *E. coli* structure with the PDB ID 3ucs was used. For the DNAJB1^{RRR} JD mutant, the DNAJB1 structure was used and the following three mutations were integrated with the PyMOL mutagenesis tool: D4R, E69R, E70R.

4.2.1 Results

Fully annotated lists of all structures are given in Tables 6.1 and 6.2 in the Appendix. Table 6.1 contains detailed information about available crystal structures, including the resolution and corresponding UniProt accession number, available NMR structures are also listed. Table 6.2 contains information about the homology modeled structures. For each modeled domain it provides the applied template structure, the sequence identity, and the QM4 score provided by the SWISS–MODEL webserver. Graphical representations of all structures are provided in the next sections together with the electrostatic potentials.

The quality of the modeled structures highly depends on the sequence identity to the template structure and the quality of the template structure itself. For example, the class A CTD homodimer of yeast (Ydj1, *S. cerevisiae*, P25491) was built by aligning a crystal structure of a yeast class A monomer two times on a crystal structure of the dimerization site. This constructed homodimer is limited in quality, but was the only way to generate a structure of a class A CTD dimer that was used many times as a template structure. But this led to an overall quality reduction of all class A CTDs.

Based on the QM4 score, the quality of the modeled JDs were in generally better than those from the CTDs. This is probably related with the higher structural conservation of the four helices in the JDs, and the fact that more structural data is available. This allowed the selection of better template structures with higher sequence identity for the individual modeled JDs.

In the case of the class A CTDs, often a sequence identity around 30 % was obtained. This sequence identity is already at the lowest limit of the criterion to decide if a structure can or should be modeled. However, based on the fact that all class A CTDs are based in general on the same template structure, this uncertainty of model quality is contained in all class A CTDs and therefore still enables the comparison between these structures. In the case of *Sphingomonas sp.* (Q1NCH5), the structure was modeled with the modeled structure of *A. aceti* (A0A063X4A7) as template, which itself was modeled with the structure of *S. cerevisiae*. This means that the structure of *Sphingomonas sp.* is also based on the structure of *S. cerevisiae*.

In the case of class B CTDs more structural data was available. This allowed the selection of template structures with higher sequence identity, leading to better model qualities in comparison to those of class A CTD models. However, in some cases also a low sequence identity of around 33 % was received, see class B models of *A. aceti*, *Sphingomonas sp.*, and *C. ultunese* in Table 6.2.

Overall, the quality of the modeled structures is limited, which should be considered when analyzing any simulation or computational data that used these models, but they also provide the possibility for further analyses and comparison of the J–proteins from different organisms.

4.3 Electrostatic Potential Comparison

This section is taken and adapted from reference [169].

As previously explained in Chapter 1, the electrostatic potentials of proteins play an important role for protein interactions and molecular function. The attraction and repulsion influences where and how strongly the proteins interact. Therefore, the analysis and visualization of the electrostatic potentials of proteins

provides a way to obtain insights into the protein binding features and compare them.

Experiments with triple charge–reversal variants of the J–domain of DNAJB1 that replaced negatively charged residues with the positively charged Arg residue around the helices I and IV, showed a strong reduction of interclass J–protein cooperation and disaggregation efficiency [169]. These results demonstrate that electrostatics might play a role in the J–protein interaction and are worth analyzing in more detail to understand how relevant they are.

Here, the electrostatic potentials of J–proteins were calculated and compared. As previously explained, the structure of these proteins is highly flexible and contains a G/F–rich linker region that is disordered. This region is not considered in the electrostatic potential comparison, because of missing structural data. Therefore, the electrostatic potentials are compared separately for the JD and the CTD structures of the class A and B J-proteins.

4.3.1 Methods

All structures were prepared by adding polar hydrogen atoms to the protein structures with WHATIF5 [189] assuming pH 7.2. Afterwards, the electrostatic potentials of each protein structure was calculated by numerically solving the linearized Poisson–Boltzmann equation with UHBD [53]. Electrostatic potential grids with 250^3 grid points with a 1 Å spacing were used for all proteins. The relative dielectric constants of the solvent and the protein were set to 78.0 and 4.0, respectively, and the dielectric boundary was defined by the protein's van der Waals surface. The ionic strength was set to 50 mM at a temperature of 300K, with an ion exclusion radius (Sternlayer) of 1.5 Å. The protein atoms were assigned OPLS atomic partial charges and radii [190].

All electrostatic potential figures were generated by using the Visual Molecular Dynamics (VMD) software [191]. The isopotential levels for all figures is defined to 1 (positive charge, cyan) and -1 (negative charge, red) kcal/mol/e. An overview of the electrostatic potential visualizations, which will be discussed in the next section, is shown in Figure 4.2 for the class A CTDs (top) and JDs (bottom) and in Figure 4.3 for the class B CTDs (top) and JDs (bottom).

4.3.2 Results and Discussion

The three–dimensional structures of JDs and CTDs of the class A and B J–proteins from human to bacteria allow the analysis of the degree of conservation of class specific electrostatic potentials. In this section, a visual inspection is described for the JDs and CTDs of class A and B, which are separated into prokaryotic and eukaryotic J–proteins. A quantitative comparison of the electrostatic potentials of the CTDs is performed in Section 4.5 with the PIPSA tool.

The prokaryotic J–protein sample includes class A J–protein DnaJ and class B CbpA structures from a wide range of bacteria (see Figure 4.2 and 4.3). The eukaryotic sample consists of non–metazoan (plants and yeast) and metazoan (nematode, human) J–proteins belonging to class A and B. The representative for the plants come from *A. thaliana* and is only included in the more detailed PIPSA analysis in Section 4.5 and the electrostatic potential is shown in Figure 4.10.

Visual inspection of JDs (Figure 4.2 and 4.3) shows a general conservation of the protein structure and electrostatic potentials within each of the J–protein classes throughout evolution. Both class A and class B JDs display a bipolar charge distribution, which is however more prominent among the class B JDs. The positive patch around α–helix II, which is implicated in HSP70 binding [192, 193], is the most prominent feature of the electrostatic potential of the JDs (see Figure 4.2 and 4.3).

Among the CTDs however, clear class–dependent differences are observed and also differences between prokaryotic and eukaryotic structures are recognizable. Qualitatively, the eukaryotic class B CTDs (Figure 4.3) are dominantly positively charged (cyan), while in prokaryotic structures, a mixture of positively (red) and negatively charged patches are observed. The structure of the prokaryotic class B *C. ultunense* (Figure 4.3, M1ZGL1) represents a different electrostatic potential compared to the other prokaryotic class B CTDs and has, similar to the eukaryotic structures, a dominant positively charged patch at the outer region of the CTD.

In eukaryotic class A CTDs (Figure 4.2), the ZFLR and CTD–I region is peppered with exposed positively and negatively charged patches, while the CTD–II is predominantly negatively charged (red). In contrast, there is a switch of these electrostatic potential patterns in prokaryotic class A J–proteins: the ZFLR and CTD–I regions are predominantly negative, while the CTD–II regions show clusters of both positive and negative patches.

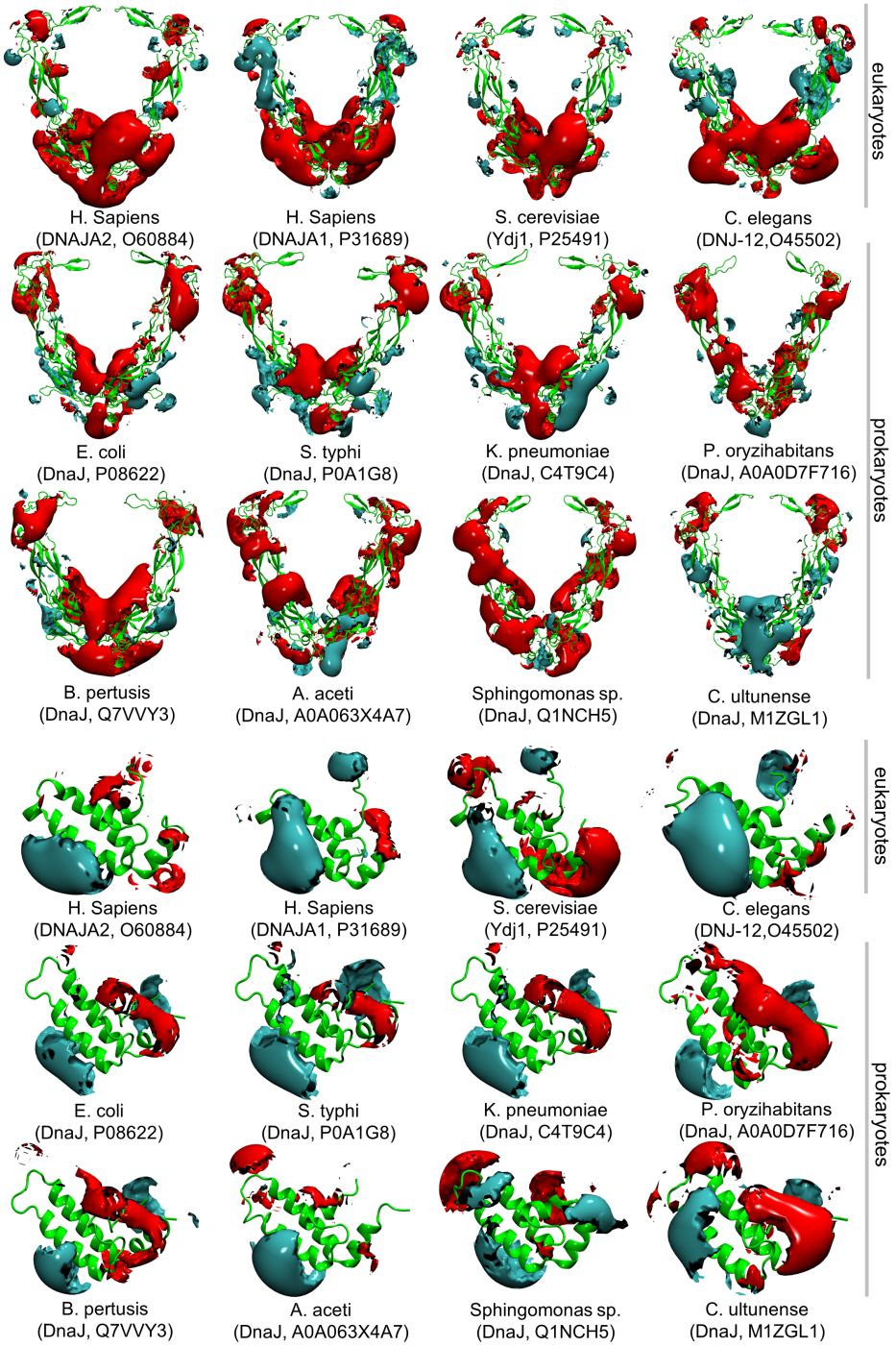

Figure 4.2: Electrostatic isopotential contour maps of class A CTD dimers (top) and JDs (bottom) from human, fungi, nematodes and bacteria. The protein structure is depicted in green cartoon representation and the electrostatic potential around the proteins is contoured at +1 (positive,cyan) and -1 (negative,red) kcal/mol/e. J–protein name and corresponding UniProt accession number are given in parentheses for each organism. Human class A J–proteins are represented by DNAJA1 (P31689) and DNAJA2 (O60884). S. cerevisiae (Ydj1, P25491) and C. elegans (DNJ-12, O45502) represent fungi and nematodes, respectively. Bacterial DnaJ are represented from the following subgroups: alphaproteobacteria (A0A063X4A7, Q1NCH5), betaproteobacterium (Q7VVY3), gammaproteobacteria (P08622, P0A1G8, C4T9C4, A0A0D7F716) and firmicute (M1ZGL1). Figure elements are based on those in reference [171] and are rearranged.

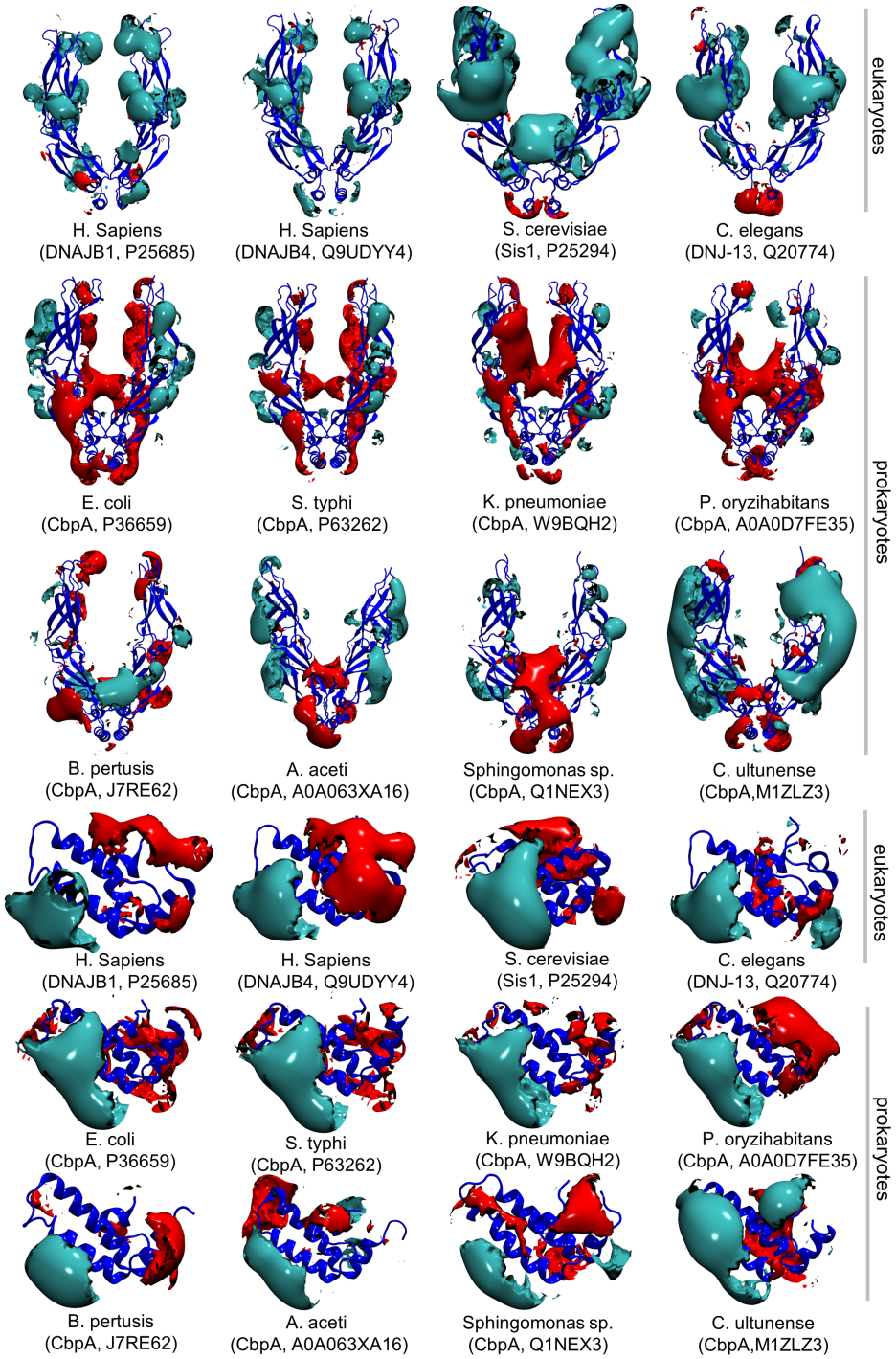

Figure 4.3: Electrostatic isopotential contour maps of class B CTD dimers (top) and JDs (bottom) from human, fungi, nematodes and bacteria. The protein structure is depicted in blue cartoon representation and the electrostatic potential around the proteins is contoured at +1 (positive,cyan) and -1 (negative,red) kcal/mol/e. J–protein name and corresponding Uniprot accession number are given in parentheses for each organism. Human class B J–proteins are represented by DNAJB1 (P25685) and DNAJB4 (Q9UDY4). S. cerevisiae (Sis1, P25294) and C. elegans (DNJ-13, Q20774) represent fungi and nematodes, respectively. Bacterial CbpA are represented from the following subgroups: alphaproteobacteria (A0A063XA16, Q1NEX3), betaproteobacterium (J7RE62), gammaproteobacteria (P36659, P63262, Q9BQH2, A0A0D7FE35) and firmicute (M1ZLZ3). Figure elements are based on those in reference [171] and are rearranged.

Taken together, eukaryotic–like features are observed emerging in class B CTDs of some bacteria such as *C. ultunense*, *A. aceti* and *Sphingomonas sp* (Figure 4.3), but not in the partnering class A CTD (Figure 4.2). Experiments showed interclass J–protein complexation in eukaryotes, but not in prokaryotes [169, 171]. The electrostatics of the interacting J–proteins are similar in both classes, those from non–interacting J–proteins are also similar, but at least in one class different to the interacting J–proteins. These features suggest J–protein networking via interclass complex formation, which is influenced by electrostatic interaction, in both animals and simpler eukaryotic unicellular organisms such as yeast, but not in bacteria.

4.4 Protein Domain Docking Simulations

This section is taken and adapted from reference [169] and [171].

To further investigate the conjecture of a protein domain interaction network of eukaryotic J–proteins, all–against–all rigid body Brownian dynamics docking simulations were performed. For this purpose, JDs were docked to inter– and intramolecular CTDs. These analyses help to understand the interactions of JDs and CTDs and detect possible interaction sites.

Previously performed experiments with DnaJ and CbpA of *E. coli* revealed a high disaggregation activity independent of class A and B mixing [171]. A computational analysis with Brownian dynamics docking simulations was performed to understand the differences between the pro- and eukaryotic J–proteins and highlight differences in the JD–CTD interactions.

Experiments with DNAJA2 and DNAJB1 revealed a set of amino acid residues that performed intermolecular cross–links [169]. This list of residues in the JDs and CTDs is shown in Table 4.3. These cross–links are used to, beside other criteria, evaluate the simulation results and test their consistency. Additional intramolecular cross–links were found in DNAJB1 and the residues in the CTD are visualized in the SDA and clustering results section.

4.4.1 Methods

A list of all performed docking simulations is shown in Table 4.4 where the first part represents simulations with human J–protein domains and the last two sim-

Table 4.3: List of experimentally determined cross–links between JDs and CTDs. This data was used to evaluate the simulation results. The sequence numbering is taken from the respective UniProt sequences (see Table 4.1 and 4.2). Data taken from [169].

Interaction	Cross–links
$DNAJA2^{JD}$ – $DNAJB1^{CTD}$	K46–K209
$DNAJB1^{JD}$ – $DNAJA2^{CTD}$	K21–K226
$DNAJA2^{CTD}$ – $DNAJB1^{CTD}$	K223–K159
$DNAJA2^{CTD}$ – $DNAJB1^{CTD}$	K223–K188
$DNAJA2^{CTD}$ – $DNAJB1^{CTD}$	K223–K195

ulations between human and *E.coli* J–proteins. All protein–protein docking simulations were performed with a rigid–body treatment of the protein structures using the Simulation of Diffusional Association (SDA) program (version 7, http://mcm.h-its.org/sda7) [64, 194]. SDA uses Brownian dynamics (BD) simulation to perform the sampling of protein configurations subject to inter–protein forces and torques due to electrostatic and non–polar interactions. The electrostatic potentials of the structures were computed as described in Section 4.3. Additional preparation steps are described in the following section.

The effective charges were derived with ECM [51] and for each protein they were fit to reproduce the electrostatic potential in a 3–Å-thick layer extending outwards from the protein's solvent–accessible surface computed as defined by a probe of radius 4 Å. The effective charges for proteins were placed on the carboxylate oxygen atoms of Asp and Glu amino acid residues and the C–terminus, and the amine nitrogen atoms of Lys and Arg amino acid residues and the N–terminus. For the Zn^{2+} ion, an effective charge site with a formal charge of -2*e* was placed on the ion, corresponding to the summed charge of the ion and its four coordinating cysteine side–chains.

The desolvation penalty of each effective charge was computed as the sum of desolvation penalties due to the low dielectric cavity of each atom of the other protein [195], which was precomputed on a grid. The grid dimensions were set to 150^3 grid points with a spacing of 1 Å. Ionic strength and dielectric constants were assigned as for the electrostatic potential calculations. The ion radius was assigned as 1.5 Å.

The non–polar desolvation forces were computed using precomputed grids [196]. The distance parameters *(a)* and *(b)* were assigned values of 3.10 Å and 4.35 Å, respectively. The parameter *(c)* was assigned as 1.0 and the conversion

Table 4.4: List of CTD–JD docking simulations between class A and B J-protein domains. The corresponding classes are mentioned in the last column, the first defines the CTD and the second the JD.

CTD	JD	Classes
DNAJA1	DNAJA1	A – A
DNAJA2	DNAJA1	A – A
DNAJA1	DNAJA2	A – A
DNAJA2	DNAJA2	A – A
DNAJB1	DNAJA1	B – A
DNAJB4	DNAJA1	B – A
DNAJB1	DNAJA2	B – A
DNAJB4	DNAJA2	B – A
DNAJA1	DNAJB1	A – B
DNAJA2	DNAJB1	A – B
DNAJB1	DNAJB1	B – B
DNAJB4	DNAJB1	B – B
DNAJA1	DNAJB4	A – B
DNAJA2	DNAJB4	A – B
DNAJB1	DNAJB4	B – B
DNAJB4	DNAJB4	B – B
$CbpA_{E.coli}$	$DnaJ_{E.coli}$	B – A
DNAJB1	$DnaJ_{E.coli}$	B – A

factor to β = –0.0065 kcal mol^{-1} Å^{-2}. The grid dimensions were set to 150^3 grid points with a spacing of 1 Å.

The calculation of the excluded volume grids was done by describing the protein shape by a grid with a 0.25 Å spacing. A probe of radius of 1.77 Å was used to determine the protein shape and the radius of the solvent probe to determine the surface atoms was set to 1.4 Å.

The SDA docking protocol itself requires the following steps: For each protein pair docking simulation, 10,000 trajectories were generated with SDA. Trajectories were started with the proteins at a separation distance of 100 Å and a random relative orientation. A trajectory was terminated if the protein separation exceeded 300 Å or a simulation time of 500 ns was reached. The protein–protein separation was calculated as the distance between their centers of geometry (COG). Up to 3,000 configurations sampled during the BD trajectories with a separation of less than 105 Å were saved. During the BD simulations, if a new docking pose is considered similar to a previously saved pose, that is, has an approximate RMSD

less than 2 Å, then the configuration with the lower intermolecular energy is saved and the counter of this docking pose, the occupation, is incremented. The relative translational diffusion coefficient was set to 0.027 $Å^2$ ps^{-1}. The rotational diffusion coefficient for both proteins was set to 3.92 * 10^4 radian 2 ps^{-1}. The time step was 1 ps at separations less than 120 Å and increased linearly beyond this threshold with a slope of 2 ps $Å^{-1}$.

For the docking simulations with the CTD of $DnaJ_{E.\ coli}$, the cluster representatives were used to calculate the average Euclidean distance between their center of geometry (COG) and the COG of the two cluster representatives of the docking with the JD of DNAJB1 and CTD of DNAJA2. For this purpose, the CTD of $DnaJ_{E.\ coli}$ was superimposed on the CTD of DNAJA2 before carrying out the docking simulations. Because of the dimeric structure of the CTDs, the distances to both COG of the cluster representatives (cluster 1 and 2, see Figure 4.6) was calculated and the smaller distance was used for calculating the average distance of all cluster representatives.

4.4.2 Clustering of Docking Complexes

The detailed analysis of the SDA docking results is discussed in the next section, but in summary the simulated protein complexes show a high degree of variation, resulting in a huge variety of slightly different orientations of the docked protein domain, and also different interaction interfaces. Additionally, the dimeric structure of the CTDs led to protein complexes on each monomer. Therefore, an adaption of the standard docking procedure of SDA results was necessary to ensure an appropriate data analysis that also detect interaction interfaces on each monomer.

Figure 4.4 visualizes the updated clustering workflow. Generated SDA results are post–processed to define the final cluster representatives. The saved configurations for each docked protein pair are clustered with a bottom–up hierarchical clustering algorithm. The backbone RMSD between each docked protein configuration was calculated to produce an inter–configuration distance matrix. Initially, each docked structure was assigned to a separate cluster. The closest clusters were merged and the distance matrix updated. This process was repeated until all docked protein structures were in one cluster. The distance between clusters is defined as the average backbone RMSD between docked protein structures in one cluster relative to structures in another cluster. The representative of a cluster is

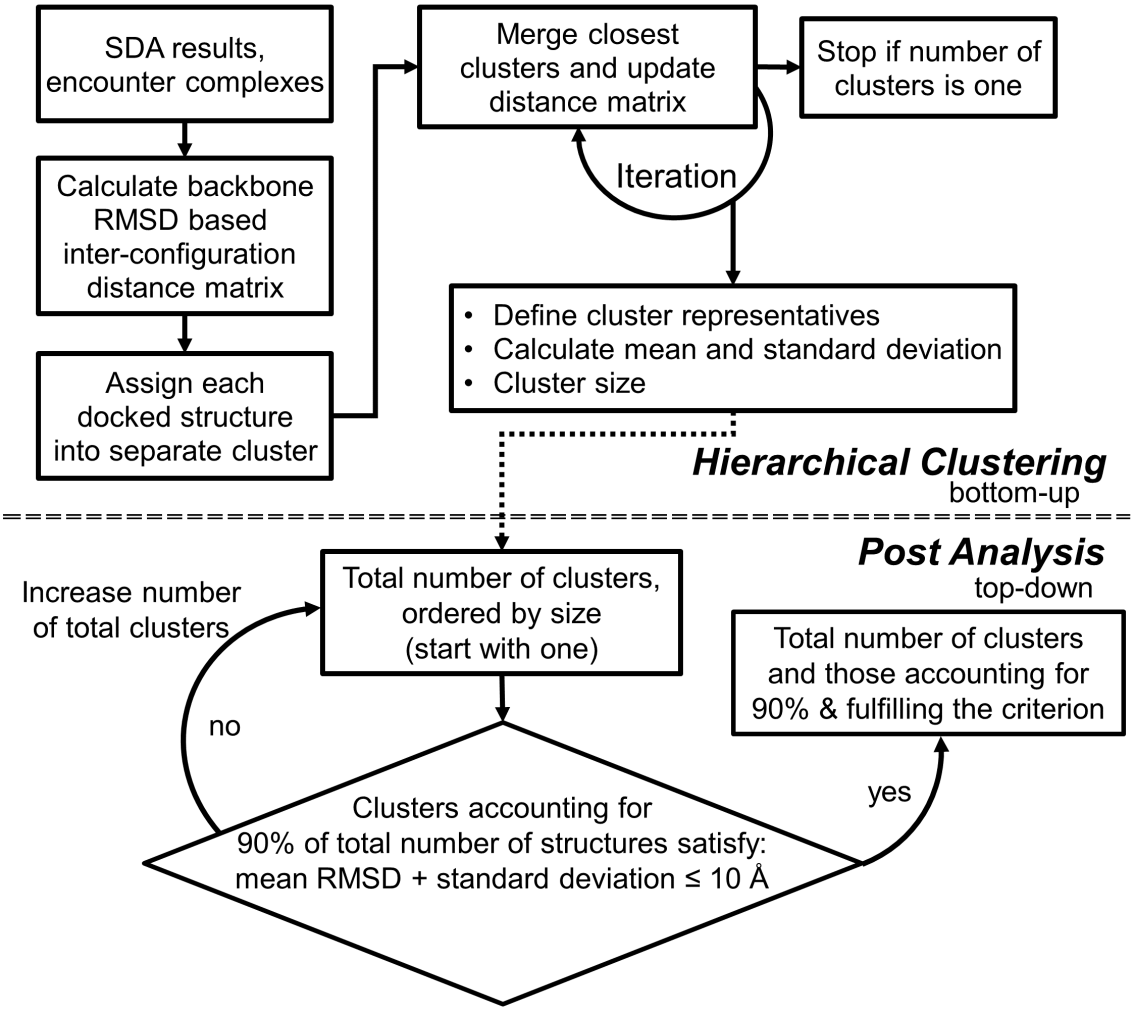

Figure 4.4: Workflow of the adapted clustering procedure for SDA results. The diamond represents a decision step that depends on the satisfaction of the shown criterion. See text for detailed description.

the protein configuration with the smallest RMSD to every other member of the cluster. In each clustering cycle, the cluster sizes, the means, and standard deviations of the RMSD of all clusters to the corresponding cluster representative are calculated and stored.

In the following post analysis, the generated clustering data was used to perform a top–down analysis. The number of configurations in each cluster in each clustering cycle was determined, taking the cluster occupation during the BD simulations into account (for explanation, see previous Section 4.4), and the clusters in each clustering cycle were ranked by size. The number of generated clusters was chosen using the following criteria. Starting with the last clustering cycle

(one total cluster) and the largest cluster, the minimum number of clusters accounting for 90% of the total number of configurations docked and satisfying the criterion that the mean RMSD plus the standard deviation of the clusters is less than 10 Å, was determined. This threshold resulted in docking configurations with similar COG of the J–domains, but orientations differing by about 90° being assigned to different clusters. This post analysis reduced the number of very small clusters and created compact clusters containing 90% of the generated clusters. The number of total and considered clusters already provide an initial overview of the simulations. When the total number of clusters was high, the generated data was probably non–specific, whereas if a few clusters were enough to produce compact clusters, the data used was more specific.

Finally, this cluster analysis workflow provides the total number of generated clusters of which only those are considered that account together for 90% of all data and each cluster fulfills the criterion shown in Figure 4.4. In the following, these clustering results are notated as follows: 2/4 for a total of 4 generated clusters and 2 considered clusters.

4.4.3 Discussion of SDA and Clustering Results

To define the interaction interface of J–domains and CTD dimers and compare them with the experimental cross–linking results, unbiased docking simulations were performed. The results of the applied SDA simulations are visualized in Figure 4.5, showing the preferred positions of the center of geometry of the docked JDs around the CTDs represented as meshes or contours. For each listed combination of JD and CTD docking simulation in Table 4.4, this type of visualization is shown in Figure 4.5. The corresponding clustering results are listed in Table 4.5.

Experiments revealed specific intermolecular cross–links [169], listed in Table 4.3. Corresponding Lys residues on the CTDs are highlighted in an orange space–filling representation in Figure 4.5, Figure 4.6, and Figure 4.7. For the intramolecular cross–links found in DNAJB1, the following residues are also highlighted in an orange space–filling representation in Figure 4.5 (bottom, second from the right): K158, K159, K188; K195, K202, K242, K306.

The results show a preferred binding interface of both classes of JDs on the lower CTD–II domain of class A CTDs. However, the clustering results in Table 4.5 reveal that class B JDs (DNAJB1, DNAJB4) bind more specifically at the lower CTD domain of class A. In general, the clustering procedure generated fewer clusters

for class B JD docking data than for class A JDs. The experimental cross–linking data corroborate these preferred interaction sites on the CTD–II of class A and CTD–I of class B CTDs.

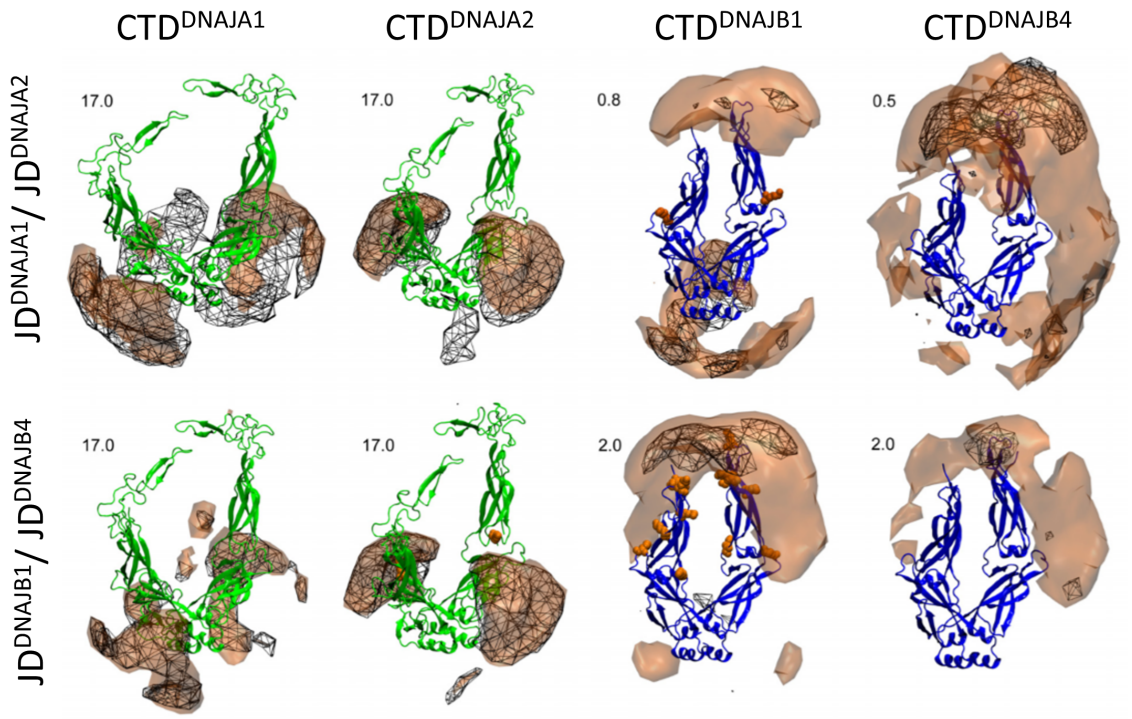

Figure 4.5: Preferred positions of the centers of geometry (COG) of J-domains (*y* axis, DNAJA1, DNAJA2, DNAJB1 and DNAJB4) around CTD dimers (*x* axis, class A, green, class B, blue) obtained from molecular docking simulations. $JD^{DNAJA1/DNAJB1}$, wireframe meshes; JD(DNAJA2/DNAJB4), brown contours, each contoured at the isovalue given in the top left of each image. The higher scores for class A CTDs indicate greater specificity of the complexes formed with J–domains; the lower scores for class B CTDs indicate much less specific interactions. Lysines in inter– and intra–J–protein DNAJA2–DNAJB1 cross–links, orange spheres. Figure elements are based on those in reference [169] and are rearranged.

The docking of the JD of DNAJA2 to the CTD of DNAJB1 is much weaker and less specific than the JD of DNAJB1 docking to the CTD of DNAJA2, but docking arrangements compatible with the cross–linking results were still obtained, see Figure 4.7.

The preferred binding arrangement of the JD of DNAJB1 on the CTD of DNAJA2 is analyzed in more detail, as specific cross–linking data is available. Figure 4.6 shows the cartoon representation of the two considered cluster representatives (blue) out of four generated total clusters. The cross–link found between K21 (JD, DNAJB1) and K226 (CTD, DNAJA2) is below a 30 Å threshold and therefore in agreement with the experiments. In addition, both cluster representatives are

Table 4.5: *In silico* prediction of JD–CTD interactions between class A and B J–proteins and *in vitro* evidence that physical interactions between J–proteins do not overlap with J–protein substrate binding sites. Properties of the docking arrangements obtained after clustering. Total number of clusters per simulation, denominator; number of selected clusters (corresponding to 90% of all docked complexes), numerator, bold. In parentheses, the range of average energy values (in units of kT) for the selected clusters. Lower energy values indicate more favourable binding; fewer clusters indicate a more defined binding mode. The results marked with a represent cluster values obtained for selected docking complexes that fall within the cross–linking range of 30 Å (see text for details). The distance specification for the $CTD^{DnaJ_{E.coli}}$ (last two rows) describes the average distance of center of geometry (COG) of the cluster representatives to the COG of the cluster representatives of the DNAJB1 JD to the DNAJA2 CTD (see text for details). The docked site for the CTD of $DnaJ_{E.\ coli}$ differs from that for human DNAJA2.

	CTD^{DNAJA1}	CTD^{DNAJA2}
JD^{DNAJA1}	**11**/25, (-23.6 to -20.8)	**6**/11, (-25.9 to -24.2)
JD^{DNAJA2}	**4**/4, (-23.9 to -22.4)	**4**/6, (-26.3 to -25.7)
JD^{DNAJB1}	**4**/9, (-22.9 to -20.8)	**2**/4, (-24.3 to -24.0)
JD^{DNAJB4}	**3**/9, (-22.1 to -21.5)	**6**/11, (-25.4 to -23.2)
	CTD^{DNAJB1}	CTD^{DNAJB4}
JD^{DNAJA1}	**42**/108, (-14.6 to -12.4)	**40**/109, (-14.9 to -11.5)
JD^{DNAJA2}	**26**/65, (-16.0 to -13.9) **94**/206, (-11.9 to -7.1)*	**23**/58, (-17.2 to -13.5)
JD^{DNAJB1}	**21**/60, (-16.8 to -14.4)	**19**/60, (-17.3 to -14.2)
JD^{DNAJB4}	**11**/34, (-18.0 to -15.7)	**1**/8, (-24.5)
	$CTD^{DnaJ_{E.coli}}$	∅ distance [Å]
$JD^{CbpA_{E.coli}}$	**6**/22, (-23.8 to -21.8)	44.1
JD^{DNAJB1}	**5**/26, (-21.0 to -20.2)	41.4

located on each monomer of the CTD homo dimer in the same binding conformation. Another criterion that corroborates the reliability of these simulation data is the accessibility of the HPD motif. As mentioned before, this motif is important for the HSP70 binding and this interaction is still possible in the overall molecular complex formed.

The clustering procedure for the simulation data of the JD of DNAJA2 to the CTD of DNAJB1 docking produced 26 out of 65 clusters. These clusters were very diverse and distributed around the whole DNAJB1 CTD. Further analysis of the interaction site around the defined cross-linking residues is not possible and therefore the simulation was repeated and only docked complexes that fall within the cross–linking range of 30 Å between the Lys209 in the CTD of DNAJB1 and Lys46 in the JD of DNAJA2 were considered for the final data set. This is still an

unbiased simulation, only the focus on complexes withing this 30 Å has changed. This data set is again clustered and 94 out of 206 clusters are generated. Even if the number of total and considered clusters is higher, the results allow an analysis of the complex formation around the interaction site that was detected experimentally via cross–links [169]. In Figure 4.7, the DNAJA2 JD representatives of cluster 1 and 2 are visualized in complex with the DNAJB1 CTD. This time, the two representatives show slightly different binding orientations (compare Figure 4.6 and 4.7). However, the accessibility of the HPD motif is still retained for both binding conformations. The JDs of DNAJA2 bind at the upper CTD–I of DNAJB1, which showed a positively charged patch in the electrostatic potential analysis (see Figure 4.3).

In summary, class B CTD dimers mostly showed a higher number of selected and total clusters with less favorable interaction energies than class A CTD dimers. Also a higher diversity in the docking poses of J–domains to class B CTD dimers (see clustering at N-termini of class B CTD dimers, Figure 4.5) can be observed.

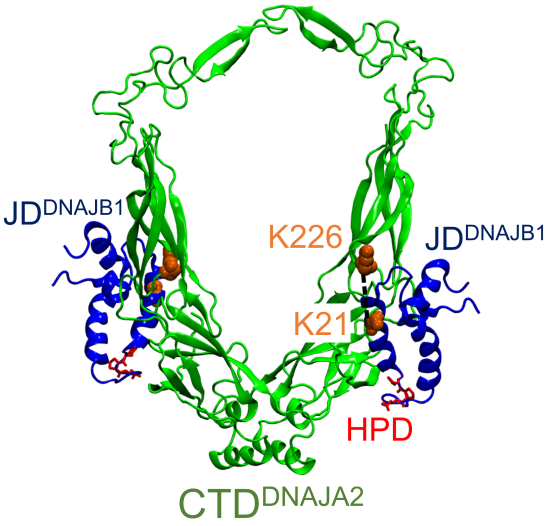

Figure 4.6: Ribbon diagrams showing representative positions of DNAJB1 JDs on the DNAJA2 CTD dimer from docking simulations; cross–linked Lys residues (space filling, orange, connected with black dashed lines). HPD motif (stick representation, red). Figure elements are based on those in reference [169] and are rearranged.

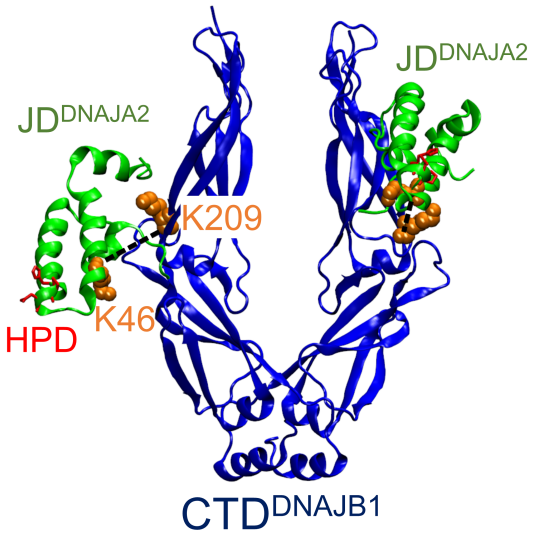

Figure 4.7: Ribbon diagrams showing representative positions of DNAJA2 JDs on the DNAJB1 CTD dimer from docking simulations; cross–linked Lys residues (space filling, orange, connected with black dashed lines). HPD motif (stick representation, red). Figure elements are based on those in reference [169] and are rearranged.

The SDA docking simulations with the CTD of $DnaJ_{E.coli}$ are visualized in Figure 4.8 with JD of $CbpA_{E.coli}$ (left) and DNAJB1 (right). The results of the clustering

procedure are listed in Table 4.5. Both simulation results visualized in 4.8 already show that the COG of both JDs are not located at the lower CTD of the $DnaJ_{E.coli}$, which is the case for the human DNAJA2 CTD (compare with Figure 4.5 and 4.6). Instead, both simulations have an overlapping interaction region at the upper CTD–I. The calculated average distances between the COGs for the six (JD of $CbpA_{E.coli}$) and five (JD of DNAJB1) cluster representatives are both above 40 Å, which is in agreement with the visual inspection of the COG mesh and contour representations.

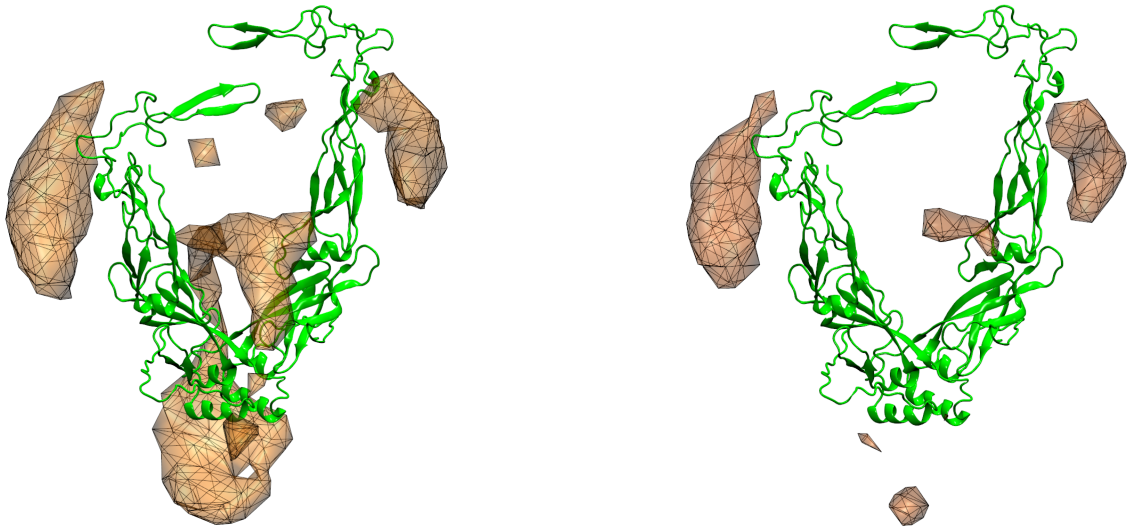

Figure 4.8: SDA docking results with $DnaJ_{E.coli}$ CTD and JDs of $CbpA_{E.coli}$ (left) and DNAJB1 (right). Overlaid mesh and contour representations (at isovalue 17 kcal/mol/e) show preferred positions of the center of geometry of the respective J–domains around the CTD dimer obtained from molecular docking simulations.

The SDA simulations highlight a preferred interaction site of human canonical class B JDs at the lower CTD of human class A CTDs. This interaction site is dominated by an electrostatic negatively charged region, which shows at the CTD–I/II hinge region small positively charged patches (see Figure 4.2). The interaction site on the upper CTD–I of class B CTDs provides a positively charged region for the JDs of class A to interact with. The simulation with the CTD of $DnaJ_{E. coli}$ revealed that canonical class B JDs, from *E. coli* and human, do not bind at the lower CTD–II, but instead prefer a region at the upper CTD–I/ZFLR.

4.5 PIPSA Analysis

This section is taken and adapted from reference [169].

To quantitatively assess the differences in electrostatic potential among prokaryotic and eukaryotic CTDs, a Protein Interaction Property Similarity Analysis (PIPSA) [56] was performed around the previously discovered JD interaction interface located at the CTD hinge regions of DNAJA2 and DNAJB1 (black dotted circles, Figure 4.9). The 25 Å radius spheres encompass residues implicated in opposite class JD interaction from crosslinking and Förster resonance energy transfer (FRET) experiments [169] and the JD docking simulations in the previous section.

Experiments with a RRR mutant of DNAJB1 JD (see Table 4.2) revealed a strong decrease of the synergistic effect of mixing class A and B J-proteins [169]. To analyze the impact of these three mutations, located on helices I and IV of the JD, on the electrostatic potential and on the interaction with class A J-proteins, a PIPSA analysis with JDs (including canonical and non–canonical JDs) was done.

4.5.1 Methods

For the electrostatic potential comparison with PIPSA it is necessary to superimpose the protein structures. Therefore, the class A CTDs were superimposed using the lower CTD–II domain of DNAJA2 (UniProt sequence residue numbers: 257 – 339) and the class B CTDs are superimposed using the upper CTD–I domain of DNAJB1 (UniProt sequence residue numbers: 156 – 245). All structures were superimposed with the alignment tool of the PyMOL software, which performs a sequence alignment followed by a structural superposition and refinement process. In addition to the CTDs of both classes, the JDs of selected class B J–proteins were also analyzed regarding their electrostatic potential similarity (Section 4.5.4). For this purpose, a global structural alignment was performed for selected class B J–domain structures on the DNAJB1 JD.

The similarity of the calculated electrostatic potentials of the superimposed structures was computed using the PIPSA software [54, 55, 56]. The resulting distance matrix was used for a Ward's clustering. Only for Type A CTDs was an average–clustering used, but this yielded similar results to the Ward's clustering. For the local PIPSA analysis, a center and a radius were defined as follows. For the local PIPSA analysis of the class A CTD, the midpoint between the residue

K226 in the DNAJA2 CTD and K21 in the DNAJB1 J–domain was chosen. This pair of residues was found in a lysine–specific cross–linking experiment, see Table 4.3 [169]. The previously performed docking simulation of the DNAJA2 CTD and the DNAJB1 J–domain supports the domain interaction and the coordinate of the midpoint was taken from the representative complexed structure (see Section 4.4 for more information). The radius of the sphere was set to 25 Å to include the whole predicted interaction site. The same procedure was applied for the DNAJB1 CTD and the DNAJA2 J–domain, for which two cross–linking residues, K209 in the DNAJB1 CTD and the K46 in the DNAJA2 J–domain, were identified, see Table 4.3 [169]. The radius of the sphere was also set to 25 Å. For the PIPSA analysis of the metazoan JDs, average–clustering was applied. All metazoan JD structures were superimposed on the DNAJB1 JD and a sphere with a radius of 25 Å was set to cover the region around helix I and IV and the RRR mutation site of DNAJB1^{RRR}.

4.5.2 Results

The electrostatic potential comparison with the PIPSA tool provides a quantitative analysis of the interaction sites on the CTDs. Figure 4.9 shows the results for the class A (top) and class B (bottom) local PIPSA analysis around the interaction sites revealed in the previously performed SDA docking simulations. The labels for the heat maps are based on the corresponding UniProt accession number and are colored black for eukaryotes and orange for prokaryotes.

The local PIPSA analysis result of the class A CTD region (interaction site with class B JDs) shows a clear separation of eukaryotic (black) and prokaryotic (orange) proteins. The two class A representative J–protein structures for pro– and eukaryotes (DnaJ$_{E.coli}$ and DNAJA2, respectively) in Figure 4.9 have the spherical region for comparison highlighted with a black dotted line. The dominant negatively charged patch on the DNAJA2 CTD–II region can not be observed in the comparable region of DnaJ$_{E.coli}$, but instead there is a general increase in exposed positive charges at the JD interaction region. These changes in electrostatic potentials therefore lead to a clear separation of pro– and eukaryotic class A CTDs.

The class B CTDs can be grouped into three clusters. The prokaryotes are grouped together in two separate clusters, where the CTD of *S. cerevisiae* (Sis1, P25294) represents an outlier of the eukaryotes as the structure has a higher similarity with some bacterial representatives (see bottom left in Figure 4.9). The class B CTDs show in general a less distinct electrostatic potential pattern at the inter-

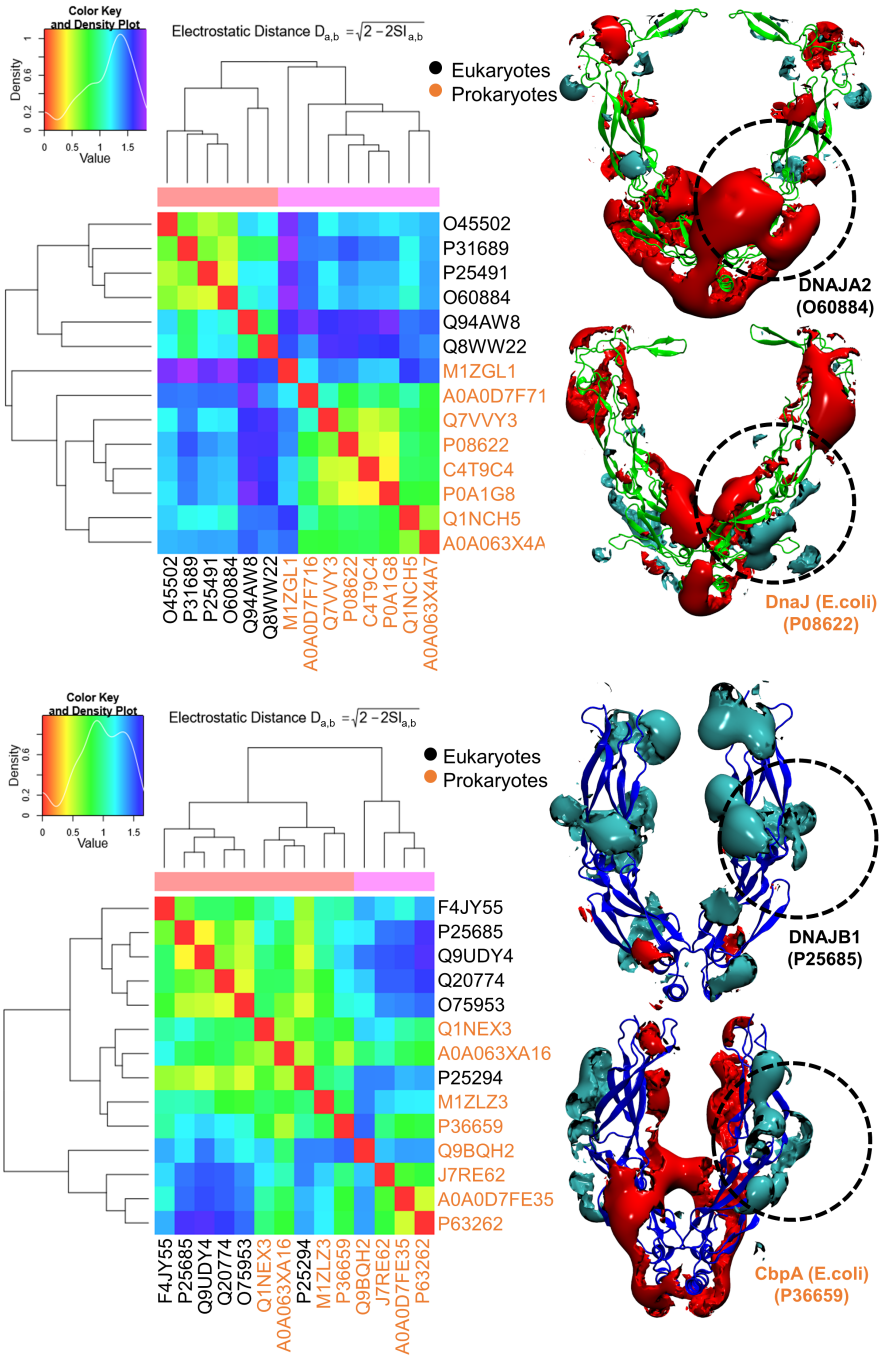

Figure 4.9: Evaluation of JD interaction sites on CTDs of the opposite class J–proteins. Local PIPSA analysis of electrostatic potentials at class B JD interaction sites on CTDs of class A J–proteins (top) and class A JD interaction sites on CTDs of class B J-proteins (bottom). Eukaryotic sequences are coloured in black and prokaryotic ones in orange. The electrostatic potentials analyzed in the spherical region (radius 25 Å) are indicated by the dashed black circles on the representative eukaryotic (DNAJA2 and DNAJB1) and prokaryotic (DnaJ and CbpA of *E.coli*) structures. These regions are clustered by electrostatic distance using Average (class A, top) and Ward's (class B, bottom) clustering. The heat map shows clustering of J-proteins according to electrostatic distance. The color key represents the elctrostatic distance that considers the similarity index of the compared potentials (red, left: high similarity – blue,right: low similarity). The color key and density plots are depicted on the top left. Figure elements are based on those in reference [171] and are rearranged.

action site with class A JDs. Nevertheless, most of the eukaryotes are noticeably grouped together. Visual inspection of the electrostatic potentials of the two representative J–protein structures for pro– and eukaryotes ($CbpA_{E.coli}$ and DNAJB1, respectively) in Figure 4.9 highlights that both structures have positively charged patches at the hinge region of CTD–I and CTD–II. This patch is slightly differently distributed and only small negatively charged patches might influence the PIPSA analysis. This explains, why the clustering of class B CTDs is less distinct than for class A CTDs in separating pro– and eukaryotic J-protein structures.

4.5.3 Analysis of J-protein representatives

Parts of this section are taken and adapted from reference [171].

In addition to the PIPSA analysis with the complete data set, an analysis of representative J–proteins from different organisms, including human, fungi, nematodes and bacteria, was performed. The reduction of data helps to identify core features and differences between the selected organisms that lead to a separation of prokaryotic and eukaryotic functionality of J–protein interactions. The following organisms are included in this analysis: *H.sapiens* (DNAJA2, DNAJB1), *C. elegans* (DNJ–12, DNJ–13), *S. cerevisiae* (Ydj1, Sis1), *A. thaliana* (Atj3, At5g25530), *P. oryzihabitans*, *B. pertussis* and *E. coli* (DnaJ, CbpA). The electrostatic potentials are shown in Figure 4.10.

Based on the electrostatic similarities around the hinge regions, the PIPSA analysis shows clustering of the CTD of J–proteins into two groups separating the prokaryotes from eukaryotes (Figure 4.11). The J–proteins ATJ3 and At5g25530 from *A. thaliana* show electrostatic potential patterns that are more eukaryotic–like (see Figure 4.10). The clustered groups of CTDs of both class A and class B J–proteins of yeast, nematode and human reflect highly conserved charge distributions at the JD interaction interface (see Figure 4.11). The same regions in prokaryotic CTDs show distinct clustering for both classes, but indicate a different charge distribution from the eukaryotic CTDs (see Figure 4.11). Therefore, we conclude that the electrostatically complementary opposite class JD interaction interface is highly conserved among human, worm and yeast J-proteins, but not in bacterial counterparts.

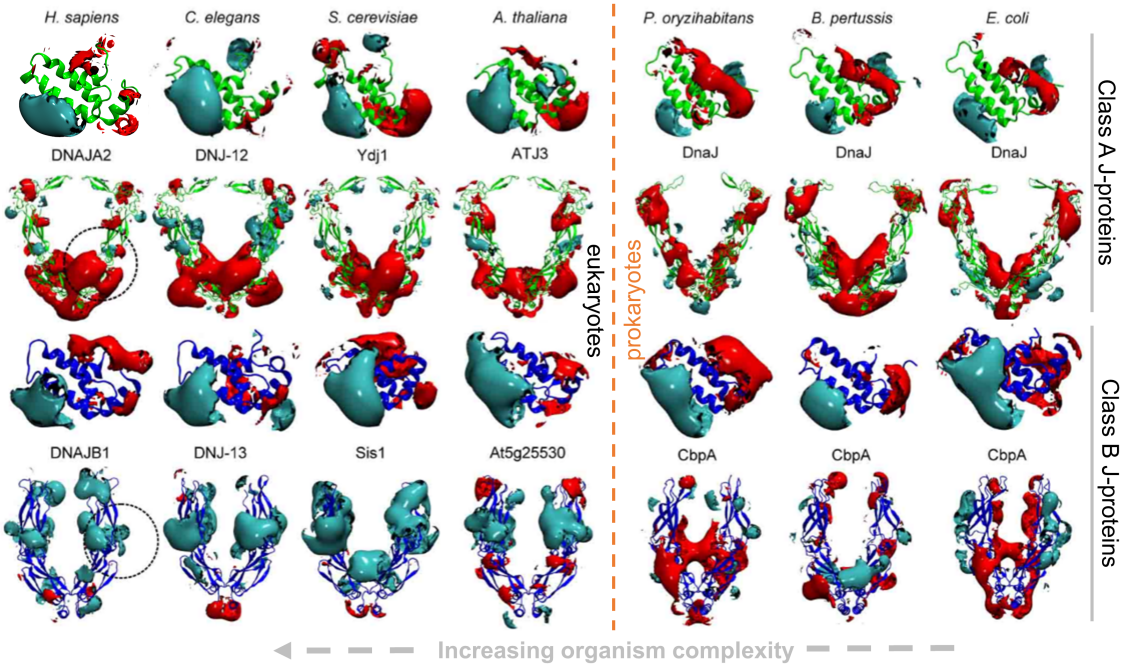

Figure 4.10: Overview of the electrostatic potentials of class A and B JDs and CTDs for representative pro– and eukaryote J–proteins. The organism complexity increases from right to left and includes J–proteins from: bacteria, plants, yeast and metazoan (human, nematode). Figure elements are based on those in reference [171] and are rearranged.

4.5.4 Class B J–domain comparison

Experimental results of a charge reversal mutant of the JD of DNAJB1 revealed an insufficient interclass J–protein cooperation and disaggregation efficiency [169]. This RRR mutation leads to a loss of the dipole character and a positively charged patch on the DNAJB1 around helices I and IV (see DNAJB1^{RRR} in Figure 4.12).

Analyses with non–canonical J–proteins, such as DNAJB2 and DNAJB4, revealed an inability to reactivate (disaggregation and refolding) aggregated luciferase [171]. Additional FRET assays with a DNAJB1 chimera containing either the JD of DNAJB8 or CbpA$_{E.coli}$ were also done [171]. These FRET competition assays with additional unlabeled chimeras showed a similar reduction in donor quenching, that was much lower than the wild type DNAJB1 J-protein. This means that the chimera proteins, containing the JD of the non–canonical DNAJB8 or the prokaryotic JD of CbpA$_{E.coli}$ are not able to compete with DNAJB1. This gives a hint that the JD of class B canonical eukaryotes are more specialized to perform complex formation with class A CTDs.

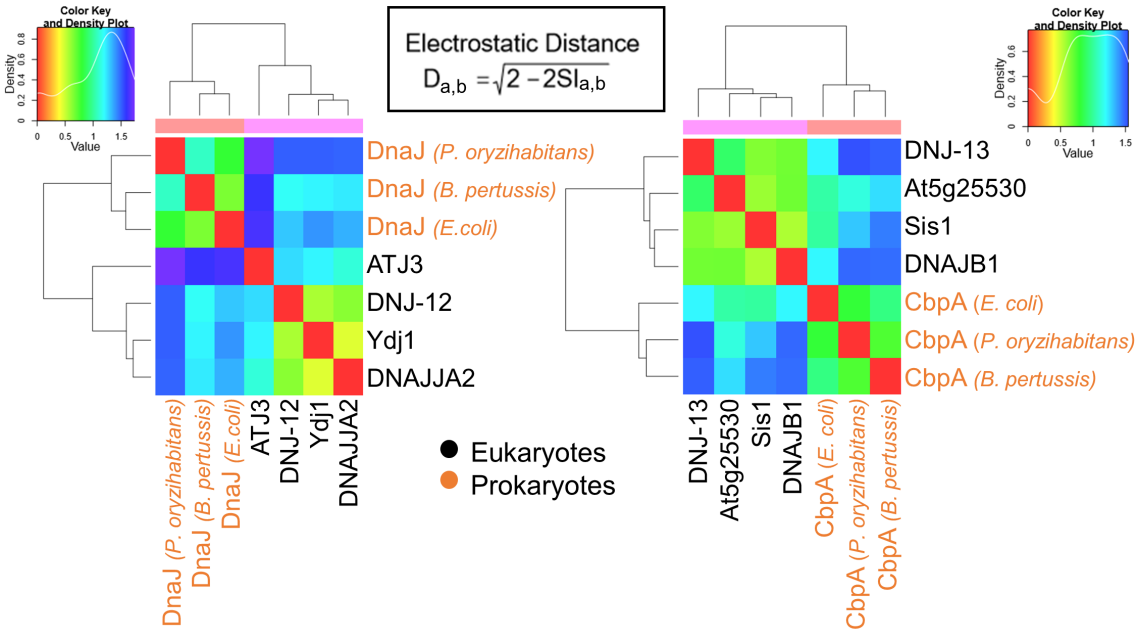

Figure 4.11: Local PIPSA analysis of class A (left) and B (right) CTD representatives around interaction interfaces. The electrostatic potentials in the spherical region (shown in Figure 4.9) are clustered by similarity using Ward's clustering.The heat maps show clustering of J–proteins by similarity (red: high similarity, blue: low similarity). Eukaryotic sequences are coloured in black and prokaryotic ones in orange. Figure elements are based on those in reference [171] and are rearranged.

Electrostatic potential calculations of the JDs of DNAJB2 and DNAJB8 showed a lack of dipole character similar to the DNAJB1^{RRR} mutant (see Figure 4.12). This dipole character is a quite conserved feature in canonical JDs of class A and B.

This difference in electrostatic potentials between canonical and non–canonical JDs is also shown quantitatively with a PIPSA analysis (Figure 4.13). Based on the spherical region around helices I and IV, the JDs of canonical J–proteins (DnaJ–13, DNAJB1, DNAJB4, At5g25530, CbpA$^{E.coli}$, Sis1) are grouped together, whereas the non–canonical JDs and the RRR mutant of DNAJB1 are separated. However, the canonical JDs are grouped into two subgroups that contains metazoan (DNJ–13, DNAJB1, DNAJB4) in one, and non–metazoan (At5g25530, CbpA$^{E.coli}$, Sis1) JDs in the other one. The experiments showed that CbpA$_{E.coli}$ could not perform complex formation with class A J-proteins. However, the electrostatic potential of the JD of CbpA$_{E.coli}$ has a dipole character and is grouped together with the other canonical JDs in the PIPSA analysis (see Figure 4.12 and 4.13). This reveals that the dipole character of the JDs is relevant for the complex formation, but is not the only criterion for the interaction and synergistic effect of mixed class J–proteins.

These results give a hint at why non–canonical J–proteins, such as DNAJB2 and DNAJB8 evade nonspecific interactions with other J–proteins. The naturally occurring charged reversion (negative to positive) in the JDs leads to a loss of the dipole character and prevents interaction with the opposite class J–proteins.

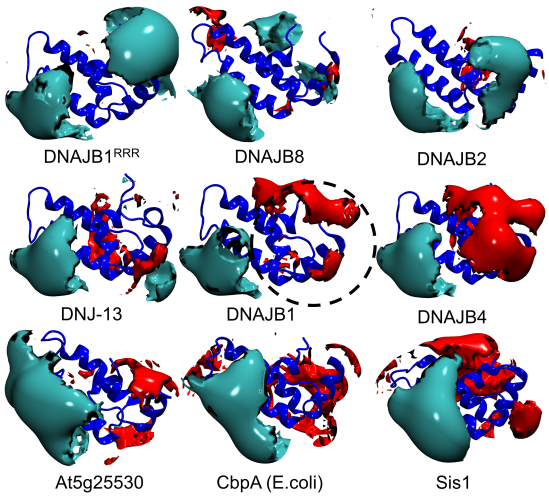

Figure 4.12: Eukaryotic class B JDs. Black dotted circle highlights the region analyzed in the PIPSA analysis. Figure elements are based on those in reference [171] and are re-arranged.

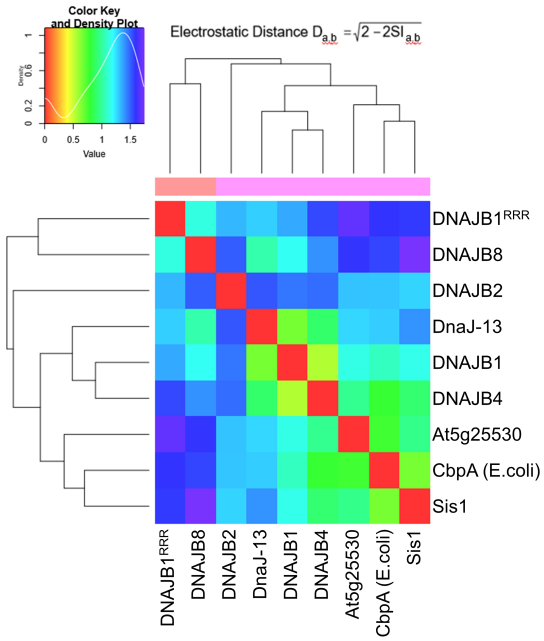

Figure 4.13: PIPSA Analysis of eukaryotic class B JDs around helices I and IV (see black dotted line in Figure 4.12). Figure elements are based on those in reference [171] and are re-arranged.

4.6 Summary and Discussion

In this chapter the questions of how the different J–protein classes interact with each other, and which features in the J–proteins lead to interclass synergy and highly efficient protein disaggregation activity were investigated. For this purpose, the three–dimensional structures of J–proteins from classes A and B from different organisms were modelled. The highly flexible and disordered G/F–rich linker region was neglected in this analysis because of missing structural data and because its high flexibility can not be easily treated in protein docking simulations. The modelled structures of the JDs and CTDs were used for the calculation of their electrostatic potentials. Visual inspection revealed clear, class specific, differences in the electrostatic potentials of the CTDs. However, the JDs showed a generally

conserved bipolar character in both classes, with this feature more apparent in class B.

In protein docking simulations with human class A and B J–proteins agreement with experimental results was obtained, which helped to confirm the interclass complexation of canonical J-proteins. Docking with the prokaryotic J–proteins from *E.coli* showed a distant interaction site that had no overlap with the identified binding site in human. Further experiments, for example, cross–links could help to further analyze if this interaction site really exists and the oligomeric structure is not able to perform efficiently, or if this interaction site is an artifact of the simulations, perhaps due to the missing G/F–rich region or the neglect of other proteins (e.g. HSP70). The simulations with the *E.coli* J–proteins revealed that the interaction of class B JDs and class A CTDs needs to occur in specific CTD regions to allow the whole oligomeric chaperone complex to be functionally efficient.

It could be shown that SDA provided a suitable tool for the detection of the multiple J–domain interaction sites. Together with the adapted clustering procedure, with its bottom–up post–analysis, appropriate cluster representatives could be selected. Based on the dimeric structure it was expected to find at least two cluster representatives, one at each monomer. The results showed that comparable JD–CTD interaction sites were detected at the opposite monomer, see Figures 4.6 and 4.7. In addition, the number of clusters also represent the diversity of the JD binding. The SDA tool was suitable for the docking process with these V–shaped protein structures, and the dimeric structure of the CTDs. The diffusion of the JDs could be simulated all around the CTD dimers, and allowed JD to dock on both monomers.

Even with the neglect of internal protein dynamics, due to the rigid protein structures, and despite omitting the G/F–rich linker region, the experimental results could be confirmed. Importantly, the simulation and experimental data are independent of each other, but give results in agreement. The modeling of the highly flexible and dimeric J–protein structures using low sequence identity was able to produce partial models with low quality. However, these structures were sufficient to perform the molecular docking simulations, and to analyze their electrostatic potentials, which were both consistent with experimental observations. Nevertheless, a detailed analysis of the possible impact of the G/F–rich linker on the inter– and intramolecular interaction of JDs and CTDs would be of value.

Finally, using PIPSA to provide a quantitative analysis of the electrostatic potentials around the identified interaction sites on the CTDs showed a clear pro– and eukaryotic separation. The additional PIPSA analysis of canonical and non–canonical J-domains of class B showed, together with experimental results, that the bipolar character of JDs, especially the negatively charged region around helix I and IV, might play a role in the J–protein selection and complexation. These results also show that electrostatics play a role in the J–protein interaction network. Changes in the electrostatic potentials can lead to a loss of the synergic cooperation of class A and B proteins, as the RRR mutant of DNAJB1 showed.

To summarize, the synergic cooperation between complexed J–protein co–-chaperone proteins of class A and B provides a highly efficient protein disaggregation activity in human and other metazoan HSP70–dependent systems. The complex formation allows J–proteins to initiate transient higher order chaperone structures involving HSP70 and interacting nucleotide exchange factors. The transient interaction of opposite class JDs and CTDs provides a powerful, flexible, and finely regulatable disaggregase activity for different kinds and sizes of protein aggregates and a further level of regulation crucial for cellular protein quality control.

It could be shown that a eukaryote–specific signature for interclass complexation of canonical J–proteins exists. Consistently, complexes exist in yeast and human cells, but not in bacteria, and correlate with cooperative action in disaggregation *in vitro*. Alterations in the signature exclude some J–proteins from networking, which ensures correct J–protein pairing, functional network integrity and J–protein specialization. These results suggest a fundamental change in J–protein biology during the prokaryote–to–eukaryote transition that allowed for increased fine-tuning and broadening of HSP70 function in eukaryotes.

5

Concluding Discussion

The analysis of proteins, considering their structure, dynamics, functions and interactions with binding partners is a crucial task in many research projects. The application of computational methods can assist the researcher and provide new insights, for example, into specific interactions of proteins or the protein dynamics. However, the use of computational methods often requires expert knowledge about the used data, or experience with the many available computational methods and software.

The aim of this work was to develop new computational methods that are accessible for a broad range of researchers and assist in the analysis of proteins regarding their structure and functionality. An application case demonstrated the benefits of combining different computational methods and experimental data to gain new insights into protein features and interactions, and detected previously unknown functionalities of specific proteins.

The webservers presented in this thesis provide useful tools for analyzing protein structures and sequences, their evolutionary relations and the impact of the dynamics of proteins with a focus on their binding pockets. The ProSAT^{+} webserver provides a tool for mapping protein sequence data and associated annotations on the three–dimensional structure and therefore assists researchers to better investigate and understand the overall function of a protein. Additional features, for example the integrated BLAST search to find similar sequences with additional sequence annotations, simplify common tasks encountered while analyzing proteins and assist with the sequence mapping. The improvement of the protein structure analysis by automatically mapping sequence annotations on the three–dimensional structures provides researchers with an easy–to–use tool for

the initial protein analysis. In addition, with the integrated URL generator, the communication and collaboration between research groups is now simplified. The easy transfer of specific annotation information together with the visualization of the protein structure is a helpful tool to prevent misunderstandings or mistakes because of wrong residue number mappings.

The consideration of protein pocket dynamics, especially in the drug design process, is important. The TRAPP, ProSAT$^+$, and LigDig webservers provide different functionalities to analyze protein binding pockets regarding their similarity, dynamics, conservation and known sequence annotations. The classification introduced here of protein binding pocket dynamics into five classes – subpocket, adjacent pocket, breathing motion, channel/tunnel, and allosteric pocket – assists in distinguishing different types of pockets. The TRAPP webserver enables the analysis of the structural plasticity of a binding pocket, and the identification of transient pocket regions and residues important for selectively targeting a specific protein. Together with the sequence–based conservation analysis of on–target and off–target groups, it enables the identification of residues that distinguish the two groups, providing a basis for developing more selective drugs. Together with the sequence annotations provided by ProSAT$^+$ that contain information about the functional impact of certain residues, this combined, synthesized and filtered data facilitates research on the specific functionality of proteins. The discovery of new, more specific drugs with fewer side effects requires a detailed analysis of the function, behavior, and dynamics of the target protein. The presented webservers aid the initial, more detailed analysis of all these features. However, the computational methods have limitations, such as computing power and the level of detail that can be simulated. Furthermore, computational methods often require biological data and the quality and amount of such data is highly critical for the computed results. Therefore, additional experimental verification of computational results is often required.

The three webservers show today's possibility of connecting and processing diverse biological data from different sources to provide the user with new information. The current limitations include the access to useful data and the presentation of the information to the user in a usable, clear, and easy to understand way. This also means a quick access to critical information filtered from a huge data set. The future tools need to combine data, process it and present the user critical and possible interesting points, but of course give access to the complete data if neces-

sary. In the case of the TRAPP webserver, for example, the next logical step will be to highlight subpockets that seem to be relevant for targeting the specific protein with a drug. In the best case this additional method will also consider the information provided by the on– and off–target feature. In principle it is even possible to perform an automatic screening of drugs, considering the discovered druggable subpockets and the on– and off–target features, and present only the high scoring drug candidates that can then be tested experimentally.

The application of several combined computational tools assisted in the discovery of the J–protein network that provides an explanation for the high efficiency of metazoan disaggregase function. The combined use of experimental, computational and simulation data was very successful and allowed the cross confirmation of obtained results. The applied methodologies were suitable for studying the interaction of J–proteins. Experiments showed the relevance of electrostatics for the J–protein interactions, therefore methodologies where the electrostatic potential features of proteins were calculated and compared could be applied and provided meaningful results. Furthermore, with the analysis of the mutated JD of DNAJB1, the methodology proved its sensitivity to minor changes in sequence. The functional change of the J–protein networking could be highlighted by visualizing and comparing the electrostatic potentials of J–proteins from different, and evolutionarily distant organisms. Together with the Brownian dynamics simulations, the interaction sites were localized and confirmed by the experiment data.

The example of the J–proteins showed the success of comparing electrostatic potentials and combining the results with Brownian dynamics simulation. It is expected that this combination of computational methods together with the adapted clustering procedure can also be applied to other proteins, to study their interaction sites and the influence of electrostatics. The additional performed PIPSA analysis is a very useful tool to quantitatively compare the proteins based on their electrostatics, and perform research on their evolutionary relation ships. Of course, not every change in protein function and interaction in the evolutionary process is related to the electrostatics, but the comparison of it is an relatively straight forward procedure and can therefore offer a first comparison of evolutionary related proteins.

The integration of the ProSAT$^+$ functionality into other webservers showed the relevance of this tool for other aspects of protein analysis. The topics and analyses of protein structure, sequence, evolution, dynamics and protein interaction are

highly dependent on each other. Therefore, it is recommended to investigate all these properties when analyzing one of them. Evolution has an impact on several levels, starting from the gene level it can change on a higher level the molecular features of proteins, such as the electrostatic potential, which in turn is relevant for molecular interactions. To understand this overall molecular network it is necessary to appropriately analyze and apply methodologies on all levels. The webservers presented here already provide a useful set of methodologies to perform analyses on most of these levels.

This work contributed to the improvement of computational protein analysis and, with the new webservers, provides access for many researchers to methodologies for studying protein dynamics, their functions, and interactions. The discovery of the J–protein network shows the necessity of combining methodologies and data, and was important for understanding the highly complex protein disaggregation process. The analysis of protein interactions, their molecular features and the combination with experimental results are critical nowadays to understand molecular functions and pathways. Experts from different fields, from biology to computer science, need to cooperate to handle the huge amount of data, to have the required expert knowledge, and have discussions about the problems from different perspectives. Improved experimental techniques, such as cryoelectron microscopy, will provide further details in the future about the structure of proteins, their dynamics and interactions. This will allow, together with the parallel improvement of computational power and methodologies, both research fields to address more complex biological questions.

In summary, the overall future goal for computational methods will be to provide expert experience in the form of implemented tools that allow non–experts from different research fields to get access to such knowledge. This requires robust tools that are easy–to–use without any prior knowledge about the data or methodology. The produced results should be automatically processed or filtered, which could have been done only with expert knowledge. The finally presented results can help other researches to get computational analysis data that can be combined with their, for example, experimental results and might provide critical information to understand complex biological processes.

Bibliography

[1] A. Tramontano, *The ten most wanted solutions in protein bioinformatics*. Chapman & Hall/CRC, 2005.

[2] A. M. Lesk, *Introduction to protein science: architecture, function, and genomics*. Oxford University Press, 2010.

[3] F. H. Crick, "On protein synthesis," in *Symposia of the Society for Experimental Biology*, vol. 12, pp. 138–163, 1958.

[4] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, and S. H. Bryant, "CDD: NCBI's conserved domain database," *Nucleic Acids Research*, vol. 43, no. D1, pp. D222–D226, 2015.

[5] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of molecular biology*, vol. 292, no. 2, pp. 195–202, 1999.

[6] J. A. Cuff, M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, "JPred: A consensus secondary structure prediction server," *Bioinformatics*, vol. 14, no. 10, pp. 892–893, 1998.

[7] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton, "JPred4: A protein secondary structure prediction server," *Nucleic Acids Research*, vol. 43, no. W1, pp. W389–W394, 2015.

[8] B. Rost, "Review: protein secondary structure prediction continues to rise," *Journal of structural biology*, vol. 134, no. 2-3, pp. 204–18, 2001.

[9] W. Pirovano and J. Heringa, "Protein Secondary Structure Prediction," in *Data Mining Techniques for the Life Sciences*, vol. 609, pp. 327–348, Humana Press, 2010.

[10] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[11] T. E. Creighton, *Biophysical Chemistry of Nucleic Acids and Proteins*. Helvetian Press, 2010.

[12] P. G. Higgs and T. K. Attwood, *Bioinformatics and molecular evolution*. Blackwell Pub, 2005.

[13] J. Zhang and J.-R. Yang, "Determinants of the rate of protein sequence evolution," *Nature reviews. Genetics*, vol. 16, no. 7, pp. 409–20, 2015.

[14] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similiarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[15] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[16] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–10, 1990.

[17] A. Pavlopoulou and I. Michalopoulos, "State-of-the-art bioinformatics protein structure prediction tools," *International Journal of Molecular Medicine*, vol. 28, no. 3, pp. 295–310, 2011.

[18] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. G. Cassarino, M. Bertoni, L. Bordoli, and T. Schwede, "SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information," *Nucleic Acids Research*, vol. 42, no. W1, pp. 252–258, 2014.

[19] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, "The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling," *Bioinformatics*, vol. 22, no. 2, pp. 195–201, 2006.

[20] L. Bordoli, F. Kiefer, K. Arnold, P. Benkert, J. Battey, and T. Schwede, "Protein structure homology modeling using SWISS-MODEL workspace," *Nature protocols*, vol. 4, no. 1, pp. 1–13, 2009.

[21] A. Stank, D. B. Kokh, J. C. Fuller, and R. C. Wade, "Protein Binding Pocket Dynamics," *Accounts of chemical research*, vol. 49, no. 5, pp. 809–815, 2016.

[22] E. Fischer, "Einfluss der Configuration auf die Wirkung der Enzyme," *Ber. Dtsch. Chem. Ges.*, vol. 27, pp. 2985–2993, 1894.

[23] D. E. Koshland, "Application of a Theory of Enzyme Specificity to Protein Synthesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 44, pp. 98–104, 1958.

[24] J. Monod, J. Wyman, and J. P. Changeux, "On the Nature of Allosteric Transitions: a Plausible model," *Journal of molecular biology*, vol. 12, pp. 88–118, 1965.

[25] J.-P. Changeux and S. Edelstein, "Conformational selection or induced fit? 50 years of debate resolved," *F1000 biology reports*, vol. 3, p. 19, 2011.

[26] A. D. Vogt and E. Di Cera, "Conformational selection or induced fit? A critical appraisal of the kinetic mechanism," *Biochemistry*, vol. 51, pp. 5894–902, 2012.

[27] M. Watanabe and O. Becker, "Dynamics Methods," in *Computational Biochemistry and Biophysics*, ch. 3, pp. 39–68, CRC Press, 2001.

[28] H.-X. Zhou, "From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions," *Biophysical journal*, vol. 98, pp. L15–7, 2010.

[29] G. G. Hammes, Y.-C. Chang, and T. G. Oas, "Conformational selection or induced fit: a flux description of reaction mechanism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 13737–41, 2009.

[30] S. Gianni, J. Dogan, and P. Jemth, "Distinguishing induced fit from conformational selection," *Biophysical chemistry*, vol. 189, pp. 33–9, 2014.

[31] Y. B. Kim, S. S. Kalinowski, and J. Marcinkeviciene, "A pre-steady state analysis of ligand binding to human glucokinase: Evidence for a preexisting equilibrium," *Biochemistry*, vol. 46, no. 5, pp. 1423–1431, 2007.

[32] L. T. Gooljarsingh, C. Fernandes, K. Yan, H. Zhang, M. Grooms, K. Johanson, R. H. Sinnamon, R. B. Kirkpatrick, J. Kerrigan, T. Lewis, M. Arnone, A. J. King, Z. Lai, R. A. Copeland, and P. J. Tummino, "A biochemical rationale for the anticancer effects of Hsp90 inhibitors: slow, tight binding inhibition by geldanamycin and its analogues," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 20, pp. 7625–7630, 2006.

[33] R. A. Copeland, "Conformational adaptation in drug-target interactions and residence time," *Future medicinal chemistry*, vol. 3, no. 12, pp. 1491–1501, 2011.

[34] S. Henrich, O. M. H. Salo-Ahen, B. Huang, F. Rippmann, G. Cruciani, and R. C. Wade, "Computational approaches to identifying and characterizing protein binding sites for ligand design," *Journal of Molecular Recognition*, vol. 23, pp. 209–219, 2010.

[35] A. Schreyer and T. Blundell, "CREDO: A protein-ligand interaction database for drug discovery," *Chemical Biology and Drug Design*, vol. 73, pp. 157–167, 2009.

[36] A. L. Hopkins and C. R. Groom, "The druggable genome," *Nature Reviews Drug Discovery*, vol. 1, no. 9, pp. 727–730, 2002.

[37] B. Huang, "MetaPocket: a meta approach to improve protein ligand binding site prediction," *Omics : a journal of integrative biology*, vol. 13, no. 4, pp. 325–330, 2009.

[38] O. Carugo and P. Argos, "Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors," *Proteins: Structure, Function and Genetics*, vol. 31, no. 2, pp. 201–213, 1998.

[39] S. K. Lüdemann, O. Carugo, and R. C. Wade, "Substrate Access to Cytochrome P450cam: A Comparison of a Thermal Motion Pathway Analysis with Molecular Dynamics Simulation Data," *Journal of Molecular Modeling*, vol. 3, no. 8, pp. 369–374, 1997.

[40] E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek, L. Biedermannova, J. Sochor, and J. Damborsky, "CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures," *PLoS computational biology*, vol. 8, p. e1002708, 2012.

[41] J. Brezovsky, E. Chovancova, A. Gora, A. Pavelka, L. Biedermannova, and J. Damborsky, "Software tools for identification, visualization and analysis of protein tunnels and channels," *Biotechnology Advances*, vol. 31, pp. 38–49, 2013.

[42] S. Bietz and M. Rarey, "ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations," *Journal of Chemical Information and Modeling*, vol. 55, no. 8, pp. 1747–1756, 2015.

[43] J. Konc and D. Janezic, "ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment," *Bioinformatics*, vol. 26, no. 9, pp. 1160–1168, 2010.

[44] M. L. Sierk and G. J. Kleywegt, "Déjà vu all over again: Finding and analyzing protein structure similarities," *Structure*, vol. 12, no. 12, pp. 2103–2111, 2004.

[45] B. Hoffmann, M. Zaslavskiy, J.-P. Vert, and V. Stoven, "A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction," *BMC bioinformatics*, vol. 11, p. 99, 2010.

[46] T. Krotzky, C. Grunwald, U. Egerland, and G. Klebe, "Large-Scale Mining for Similar Protein Binding Pockets: With RAPMAD Retrieval on the Fly Becomes Real," *Journal of Chemical Information and Modeling*, vol. 55, no. 1, pp. 165–179, 2015.

[47] S. Schmitt, D. Kuhn, and G. Klebe, "A new method to detect related function among proteins independent of sequence and fold homology," *Journal of Molecular Biology*, vol. 323, no. 2, pp. 387–406, 2002.

[48] E. Kellenberger, C. Schalon, and D. Rognan, "How to Measure the Similarity Between Protein Ligand-Binding Sites?," *Current Computer - Aided Drug Design*, vol. 4, no. 3, pp. 209–220, 2008.

[49] M. Bartolowits and V. J. Davisson, "Considerations of Protein Subpockets in Fragment-Based Drug Design.," *Chemical biology & drug design*, vol. 87, no. 1, pp. 5–20, 2015.

[50] R. Nussinov and G. Schreiber, *Computational protein-protein interactions*. CRC Press, 2009.

[51] R. R. Gabdoulline and R. C. Wade, "Effective Charges for Macromolecules in Solvent," *The Journal of Physical Chemistry*, vol. 100, no. 9, pp. 3868–3878, 1996.

[52] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, "Electrostatics of nanosystems: application to microtubules and the ribosome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 18, pp. 10037–41, 2001.

[53] J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Luty, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott, and J. A. McCammon, "Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program," *Computer Physics Communications*, vol. 91, no. 1-3, pp. 57–95, 1995.

[54] N. Blomberg, R. R. Gabdoulline, M. Nilges, and R. C. Wade, "Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity," *Proteins: Structure, Function and Genetics*, vol. 37, no. 3, pp. 379–387, 1999.

[55] F. De Rienzo, R. R. Gabdoulline, M. C. Menziani, R. C. Wade, F. D. E. Rienzo, R. R. Gabdoulline, M. C. Menziani, and R. C. Wade, "Blue copper proteins: a comparative analysis of their molecular interaction properties," *Protein science*, vol. 9, pp. 1439–1454, 2000.

[56] R. C. Wade, R. R. Gabdoulline, and F. De Rienzo, "Protein interaction property similarity analysis," *International Journal of Quantum Chemistry*, vol. 83, no. 3-4, pp. 122–127, 2001.

[57] S. Richter, A. Wenzel, M. Stein, R. R. Gabdoulline, and R. C. Wade, "webPIPSA: a web server for the comparison of protein interaction properties," *Nucleic acids research*, vol. 36, no. Web Server issue, 2008.

[58] A. M. Buckle, G. Schreiber, and A. R. Fersht, "Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution," *Biochemistry*, vol. 33, no. 30, pp. 8878–8889, 1994.

[59] R. Brown, "A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies," *Philosophical Magazine Series 2*, vol. 4, no. 21, pp. 161–173, 1828.

[60] A. Einstein, *Investigations on the Theory of the Brownian Movement*. Courier Corporation, 1956.

[61] M. von Smoluchowski, "Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen," *Annalen der Physik*, vol. 326, no. 14, pp. 756–780, 1906.

[62] R. R. Gabdoulline and R. C. Wade, "Simulation of the diffusional association of barnase and barstar," *Biophysical journal*, vol. 72, no. 5, pp. 1917–29, 1997.

[63] R. R. Gabdoulline and R. C. Wade, "Brownian dynamics simulation of protein-protein diffusional encounter," *Methods*, vol. 14, no. 3, pp. 329–341, 1998.

[64] M. Martinez, N. J. Bruce, J. Romanowska, D. B. Kokh, M. Ozboyaci, X. Yu, M. A. Öztürk, S. Richter, and R. C. Wade, "SDA 7: A modular and parallel implementation of the simulation of diffusional association software," *Journal of Computational Chemistry*, vol. 36, no. 21, pp. 1631–1645, 2015.

[65] J. D. MacCuish and N. E. MacCuish, *Clustering in bioinformatics and drug discovery*. CRC Press, 2011.

[66] R. R. Sokal and C. D. Michener, "A Statistical Method for Evaluating Systematic Relationships," *The University of Kansas Science Bulletin*, vol. 38, pp. 1409–1438, 1958.

[67] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[68] F. Murtagh and P. Legendre, "Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm," *arXiv:1111.6285*, 2011.

[69] F. Murtagh and P. Legendre, "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?," *Journal of Classification*, vol. 31, no. 3, pp. 274–295, 2014.

[70] R. R. Gabdoulline, R. Hoffmann, F. Leitner, and R. C. Wade, "ProSAT: functional annotation of protein 3D structures," *Bioinformatics*, vol. 19, no. 13, pp. 1723–1725, 2003.

[71] R. R. Gabdoulline, S. Ulbrich, S. Richter, and R. C. Wade, "ProSAT2 - Protein Structure Annotation Server," *Nucleic Acids Research*, vol. 34, no. Web Server issue, 2006.

[72] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Research*, vol. 28, no. 1, pp. 45–48, 2000.

[73] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. A. Sigrist, K. Hofmann, and A. Bairoch, "The PROSITE database, its status in 2002," *Nucleic acids research*, vol. 30, no. 1, pp. 235–238, 2002.

[74] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic acids research*, vol. 34, no. Database issue, pp. D187–91, 2006.

[75] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg, "BRENDA, the enzyme database: updates and major

new developments," *Nucleic acids research*, vol. 32, no. Database issue, pp. D431–3, 2004.

[76] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[77] C. J. Mielke, L. J. Mandarino, and V. Dinu, "AMASS: A database for investigating protein structures," *Bioinformatics*, vol. 30, no. 11, pp. 1595–1600, 2014.

[78] S. I. O'Donoghue, K. S. Sabir, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, M. Roos, N. Perdigao, F. A. Buske, J. Heinrich, B. Rost, and A. Schafferhans, "Aquaria: simplifying discovery and insight from protein structures," *Nature methods*, vol. 12, no. 2, pp. 98–99, 2015.

[79] A. S. Rose and P. W. Hildebrand, "NGL Viewer: A web application for molecular visualization," *Nucleic Acids Research*, vol. 43, no. W1, pp. W576–W579, 2015.

[80] U. Omasits, C. H. Ahrens, S. Müller, and B. Wollscheid, "Protter: Interactive protein feature visualization and integration with experimental proteomic data," *Bioinformatics*, vol. 30, no. 6, pp. 884–886, 2014.

[81] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera - A visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.

[82] A. Stank, S. Richter, and R. C. Wade, "ProSAT+: visualizing sequence annotations on 3D structure," *Protein Engineering Design and Selection*, vol. 29, no. 8, pp. 281–284, 2016.

[83] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," *BMC bioinformatics*, vol. 10, p. 421, 2009.

[84] S. Velankar, J. M. Dana, J. Jacobsen, G. Van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin, and G. J. Kleywegt, "SIFTS:

Structure Integration with Function, Taxonomy and Sequences resource," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D483–D489, 2013.

[85] The UniProt Consortium, "UniProt: a hub for protein information," *Nucleic Acids Research*, vol. 43, no. Database issue, pp. D204–D212, 2015.

[86] R. Apweiler, A. Bairoch, and C. H. Wu, "Protein sequence databases," *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 76–80, 2004.

[87] The UniProt Consortium, "UniProt: a hub for protein information.," *Nucleic acids research*, vol. 43, no. Database issue, pp. D204–12, 2015.

[88] A. Fleischmann, M. Darsow, K. Degtyarenko, W. Fleischmann, S. Boyce, K. B. Axelsen, A. Bairoch, D. Schomburg, K. F. Tipton, and R. Apweiler, "IntEnz, the integrated relational enzyme database," *Nucleic Acids Research*, vol. 32, no. Database issue, pp. D434–D437, 2004.

[89] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. M. Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M. C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler, "The UniProt-GO Annotation database in 2011," *Nucleic Acids Research*, vol. 40, no. Database issue, pp. D565–D570, 2012.

[90] M. Punta, P. Coggill, R. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. Sonnhammer, S. Eddy, A. Bateman, and R. Finn, "The Pfam protein families database," *Nucleic Acids Res*, vol. 40, no. Databse issue, pp. D290–D301, 2012.

[91] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. De Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera,

J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. A. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats, and S. Y. Yong, "InterPro in 2011: New developments in the family and domain prediction database," *Nucleic Acids Research*, vol. 40, no. Database issue, pp. D306–D312, 2012.

[92] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "Data growth and its impact on the SCOP database: New developments," *Nucleic Acids Research*, vol. 36, no. Database issue, pp. D419–D425, 2008.

[93] A. L. Cuff, I. Sillitoe, T. Lewis, A. B. Clegg, R. Rentzsch, N. Furnham, M. Pellegrini-Calace, D. Jones, J. Thornton, and C. A. Orengo, "Extending CATH: Increasing coverage of the protein structure universe and linking structure with function," *Nucleic Acids Research*, vol. 39, no. Database issue, pp. D420–D426, 2011.

[94] J. O. Ebbert, D. M. Dupras, and P. J. Erwin, "Searching the Medical Literature Using PubMed: A Tutorial," *Mayo Clinic proceedings*, vol. 78, no. 1, pp. 87–91, 2003.

[95] A. Prlić, A. Yates, S. E. Bliven, P. W. Rose, J. Jacobsen, P. V. Troshin, M. Chapman, J. Gao, C. H. Koh, S. Foisy, R. Holland, G. Rimsa, M. L. Heuer, H. Brandstätter-Müller, P. E. Bourne, and S. Willis, "BioJava: An open-source framework for bioinformatics in 2012," *Bioinformatics*, vol. 28, no. 20, pp. 2693–2695, 2012.

[96] K. Henrick and J. M. Thornton, "PQS: A protein quaternary structure file server," *Trends in Biochemical Sciences*, vol. 23, no. 9, pp. 358–361, 1998.

[97] A. E. Miele, F. Draghi, A. Arcovito, A. Bellelli, M. Brunori, C. Travaglini-Allocatelli, and B. Vallone, "Control of heme reactivity by diffusion: Structural basis and functional characterization in hemoglobin mutants," *Biochemistry*, vol. 40, no. 48, pp. 14449–14458, 2001.

[98] A. Mottaz, F. P. A. David, A. L. Veuthey, and Y. L. Yip, "Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar," *Bioinformatics*, vol. 26, no. 6, pp. 851–852, 2010.

[99] J. C. Fuller, M. Martinez, S. Henrich, A. Stank, S. Richter, and R. C. Wade, "LigDig: a web server for querying ligand-protein interactions," *Bioinformatics*, vol. 31, no. 7, pp. 1147–1149, 2015.

[100] A. Stank, D. B. Kokh, M. Horn, E. Sizikova, R. Neil, J. Panecka, S. Richter, and R. C. Wade, "TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets," *Submitted*, 2017.

[101] Z. Lu and D. M. Cyr, "Protein folding activity of Hsp70 is modified differentially by the Hsp40 co-chaperones Sis1 and Ydj1," *Journal of Biological Chemistry*, vol. 273, no. 43, pp. 27824–27830, 1998.

[102] J. Yu, W. Wei, E. Danner, R. K. Ashley, J. N. Israelachvili, and J. H. Waite, "Mussel protein adhesion depends on interprotein thiol-mediated redox modulation," *Nature chemical biology*, vol. 7, no. 9, pp. 588–590, 2011.

[103] O. Cohavi, S. Corni, F. de Rienzo, R. di Felice, K. E. Gottschalk, M. Hoefling, D. Kokh, E. Molinari, G. Schreiber, A. Vaskevich, and R. C. Wade, "Protein-surface interactions: Challenging experiments and computations," *Journal of Molecular Recognition*, vol. 23, no. 3, pp. 259–262, 2010.

[104] M. Ozboyaci, D. B. Kokh, S. Corni, and R. C. Wade, "Modeling and simulation of protein-surface interactions: achievements and challenges," *Quarterly Reviews of Biophysics*, vol. 49, no. 16, p. e4, 2016.

[105] X. Yu, V. Cojocaru, and R. C. Wade, "Conformational diversity and ligand tunnels of mammalian cytochrome P450s," *Biotechnology and Applied Biochemistry*, vol. 60, no. 1, pp. 134–145, 2013.

[106] J. Konc, M. Hodoscek, and Dusanka Janezic, "Molecular Surface Walk," *Croatia Chemica Acta*, vol. 79, no. 2, pp. 237–241, 2006.

[107] J. Konc and D. Janezic, "Protein-protein binding-sites prediction by protein surface structure conservation," *Journal of Chemical Information and Modeling*, vol. 47, no. 3, pp. 940–944, 2007.

[108] M. R. Arkin, M. Randal, W. L. DeLano, J. Hyde, T. N. Luong, J. D. Oslob, D. R. Raphael, L. Taylor, J. Wang, R. S. McDowell, J. A. Wells, and A. C. Braisted, "Binding of small molecules to an adaptive protein-protein interface," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 1603–1608, 2003.

[109] J. Hyde, A. C. Braisted, M. Randal, and M. R. Arkin, "Discovery and characterization of cooperative ligand binding in the adaptive region of interleukin-2," *Biochemistry*, vol. 42, no. 21, pp. 6475–83, 2003.

[110] S. Tassin-Moindrot, A. Caille, J. P. Douliez, D. Marion, and F. Vovelle, "The wide binding properties of a wheat nonspecific lipid transfer protein. Solution structure of a complex with prostaglandin B2," *European journal of biochemistry / FEBS*, vol. 267, pp. 1117–1124, 2000.

[111] C. Pargellis, L. Tong, L. Churchill, P. F. Cirillo, T. Gilmore, A. G. Graham, P. M. Grob, E. R. Hickey, N. Moss, S. Pav, and J. Regan, "Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site," *Nature structural biology*, vol. 9, pp. 268–272, 2002.

[112] M. P. Jacobson, R. A. Friesner, Z. Xiang, and B. Honig, "On the role of the crystal environment in determining protein side-chain conformations," *Journal of Molecular Biology*, vol. 320, no. 3, pp. 597–608, 2002.

[113] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw, and R. A. Friesner, "A Hierarchical Approach to All-Atom Protein Loop Prediction," *Proteins: Structure, Function and Genetics*, vol. 55, pp. 351–367, 2004.

[114] C. E. Stebbins, A. A. Russo, C. Schneider, N. Rosen, F. U. Hartl, and N. P. Pavletich, "Crystal structure of an Hsp90-geldanamycin complex: targeting of a protein chaperone by an antitumor agent," *Cell*, vol. 89, no. 2, pp. 239–250, 1997.

[115] L. Wright, X. Barril, B. Dymock, L. Sheridan, A. Surgenor, M. Beswick, M. Drysdale, A. Collier, A. Massey, N. Davies, A. Fink, C. Fromont, W. Aherne, K. Boxall, S. Sharp, P. Workman, and R. E. Hubbard, "Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms," *Chemistry and Biology*, vol. 11, no. 6, pp. 775–785, 2004.

[116] C. D. Thanos, M. Randal, and J. A. Wells, "Potent Small-Molecule Binding to a Dynamic Hot Spot on IL-2," *Journal of the American Chemical Society*, vol. 125, no. 50, pp. 15280–15281, 2003.

[117] G. Lessene, P. E. Czabotar, B. E. Sleebs, K. Zobel, K. N. Lowes, J. M. Adams, J. B. Baell, P. M. Colman, K. Deshayes, W. J. Fairbrother, J. A. Flygare, P. Gibbons, W. J. A. Kersten, S. Kulasegaram, R. M. Moss, J. P. Parisot, B. J. Smith, I. P. Street, H. Yang, D. C. S. Huang, and K. G. Watson, "Structure-guided design of a selective BCL-X(L) inhibitor," *Nature chemical biology*, vol. 9, no. 6, pp. 390–7, 2013.

[118] B. E. Sleebs, P. E. Czabotar, W. J. Fairbrother, W. D. Fairlie, J. A. Flygare, D. C. S. Huang, W. J. A. Kersten, M. F. T. Koehler, G. Lessene, K. Lowes, J. P. Parisot, B. J. Smith, M. L. Smith, A. J. Souers, I. P. Street, H. Yang, and J. B. Baell, "Quinazoline sulfonamides as dual binders of the proteins B-cell lymphoma 2 and B-cell lymphoma extra long with potent proapoptotic cell-based activity," *Journal of Medicinal Chemistry*, vol. 54, no. 6, pp. 1914–1926, 2011.

[119] K. W. Underwood, K. D. Parris, E. Federico, L. Mosyak, R. M. Czerwinski, T. Shane, M. Taylor, K. Svenson, Y. Liu, C. L. Hsiao, S. Wolfrom, M. Maguire, K. Malakian, J. B. Telliez, L. L. Lin, R. W. Kriz, J. Seehra, W. S. Somers, and M. L. Stahl, "Catalytically active MAP KAP kinase 2 structures in complex with staurosporine and ADP reveal differences with the autoinhibited enzyme," *Structure*, vol. 11, no. 6, pp. 627–636, 2003.

[120] H. Pan, J. C. Lee, and V. J. Hilser, "Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 22, pp. 12020–5, 2000.

[121] J. Romanowska, D. B. Kokh, J. C. Fuller, and R. C. Wade, "Computational Approaches for Studying Drug Binding Kinetics," in *Thermodynamics and Kinetics of Drug Binding*, pp. 211–235, Wiley-VCH Verlag GmbH & Co. KGaA, 2015.

[122] G. Klebe, "The use of thermodynamic and kinetic data in drug discovery: Decisive insight or increasing the puzzlement?," *ChemMedChem*, vol. 10, no. 2, pp. 229–231, 2015.

[123] J. A. McCammon and S. H. Northrup, "Gated binding of ligands to proteins," *Nature*, vol. 293, no. 5830, pp. 316–317, 1981.

[124] E. P. Garvey, "Structural Mechanisms of Slow-Onset, Two-Step Enzyme Inhibition," *Current Chemical Biology*, vol. 4, no. 1, pp. 64–73, 2010.

[125] H. Li, K. S. Leung, P. J. Ballester, and M. H. Wong, "Istar: A web platform for large-scale protein-ligand docking," *PLoS ONE*, vol. 9, no. 1, 2014.

[126] C. Wilson, R. V. Agafonov, M. Hoemberger, S. Kutter, A. Zorba, J. Halpin, V. Buosi, R. Otten, D. Waterman, D. L. Theobald, and D. Kern, "Using ancient protein kinases to unravel a modern cancer drug's mechanism," *Science*, vol. 347, no. 6224, pp. 882–886, 2015.

[127] S. Eyrisch and V. Helms, "Transient pockets on protein surfaces involved in protein-protein interaction," *Journal of medicinal chemistry*, vol. 50, pp. 3457–64, 2007.

[128] A. Ahmed, F. Rippmann, G. Barnickel, and H. Gohlke, "A normal mode-based geometric simulation approach for exploring biologically relevant conformational transitions in proteins," *Journal of Chemical Information and Modeling*, vol. 51, no. 7, pp. 1604–1622, 2011.

[129] D. Seeliger, J. Haas, and B. L. de Groot, "Geometry-based sampling of conformational transitions in proteins.," *Structure*, vol. 15, no. 11, pp. 1482–92, 2007.

[130] P. Ashford, D. S. Moss, A. Alex, S. K. Yeap, A. Povia, I. Nobeli, and M. A. Williams, "Visualisation of variable binding pockets on protein surfaces by probabilistic analysis of related structure sets," *BMC Bioinformatics*, vol. 13, p. 39, 2012.

[131] D. Seeliger and B. L. De Groot, "Conformational transitions upon ligand binding: Holo-structure prediction from apo conformations," *PLoS Computational Biology*, vol. 6, no. 1, p. e1000634, 2010.

[132] S. Wells, S. Menor, B. Hespenheide, and M. F. Thorpe, "Constrained geometric simulation of diffusive motion in proteins," *Physical biology*, vol. 2, pp. S127–S136, 2005.

[133] A. Metz, C. Pfleger, H. Kopitz, S. Pfeiffer-Marek, K.-H. Baringhaus, and H. Gohlke, "Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein-protein interface," *Journal of chemical information and modeling*, vol. 52, pp. 120–33, 2012.

[134] P. Schmidtke, A. Bidon-Chanal, F. J. Luque, and X. Barril, "MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories," *Bioinformatics*, vol. 27, pp. 3276–85, 2011.

[135] I. R. Craig, C. Pfleger, H. Gohlke, J. W. Essex, and K. Spiegel, "Pocket-space maps to identify novel binding-site conformations in proteins," *Journal of chemical information and modeling*, vol. 51, pp. 2666–79, 2011.

[136] D. B. Kokh, S. Richter, S. Henrich, P. Czodrowski, F. Rippmann, and R. C. Wade, "TRAPP: a tool for analysis of transient binding pockets in proteins," *Journal of chemical information and modeling*, vol. 53, pp. 1235–52, 2013.

[137] A. Goncearenco, S. Mitternacht, T. Yong, B. Eisenhaber, F. Eisenhaber, and I. N. Berezovsky, "SPACER: server for predicting allosteric communication and effects of regulation," *Nucleic acids research*, vol. 41, pp. W266–W272, 2013.

[138] C. Kaya, A. Armutlulu, S. Ekesan, and T. Haliloglu, "MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues," *Nucleic Acids Research*, vol. 41, pp. W249–W255, 2013.

[139] T. Paramo, A. East, D. Garzón, M. B. Ulmschneider, and P. J. Bond, "Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: trj_cavity," *Journal of Chemical Theory and Computation*, vol. 10, pp. 2151–2164, 2014.

[140] J. Parulek, C. Turkay, N. Reuter, and I. Viola, "Visual cavity analysis in molecular simulations," *BMC bioinformatics*, vol. 14, p. S4, 2013.

[141] M. Petšek, P. Košinová, J. Koča, and M. Otyepka, "MOLE: A Voronoi Diagram-Based Explorer of Molecular Channels, Pores, and Tunnels," *Structure*, vol. 15, pp. 1357–1363, 2007.

[142] J. Schames and R. Henchman, "Discovery of a novel binding trench in HIV integrase," *Journal of medicinal chemistry*, vol. 47, no. 8, pp. 1879–1881, 2004.

[143] V. Summa, A. Petrocchi, F. Bonelli, B. Crescenzi, M. Donghi, M. Ferrara, F. Fiore, C. Gardelli, O. G. Paz, D. J. Hazuda, P. Jones, O. Kinzel, R. Laufer, E. Monteagudo, E. Muraglia, E. Nizi, F. Orvieto, P. Pace, G. Pescatore, R. Scarpelli, K. Stillmock, M. V. Witmer, and M. Rowley, "Discovery of raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection," *Journal of Medicinal Chemistry*, vol. 51, pp. 5843–5855, 2008.

[144] F. Pietrucci, A. V. Vargiu, and A. Kranjc, "HIV-1 Protease Dimerization Dynamics Reveals a Transient Druggable Binding Pocket at the Interface," *Scientific Reports*, vol. 5, no. 18555, 2015.

[145] X. Zheng, L. Gan, E. Wang, and J. Wang, "Pocket-based drug design: exploring pocket space," *The AAPS journal*, vol. 15, no. 1, pp. 228–41, 2013.

[146] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: an open source platform for ligand pocket detection," *BMC bioinformatics*, vol. 10, p. 168, 2009.

[147] P. Schmidtke, V. Le Guilloux, J. Maupetit, and P. Tufféry, "fpocket: Online tools for protein ensemble pocket detection and tracking," *Nucleic Acids Research*, vol. 38, pp. 582–589, 2010.

[148] B. Laurent, M. Chavent, T. Cragnolini, A. C. E. Dahl, S. Pasquali, P. Derreumaux, M. S. P. Sansom, and M. Baaden, "Epock: Rapid analysis of protein pocket dynamics," *Bioinformatics*, vol. 31, no. 9, pp. 1478–1480, 2014.

[149] J. D. Durrant, L. Votapka, J. Sørensen, and R. E. Amaro, "POVME 2.0: An enhanced tool for determining pocket shape and volume characteristics," *Journal of Chemical Theory and Computation*, vol. 10, no. 11, pp. 5047–5056, 2014.

[150] D. B. Kokh, P. Czodrowski, F. Rippmann, and R. C. Wade, "Perturbation Approaches for Exploring Protein Binding Site Flexibility to Predict Transient Binding Pockets," *Journal of Chemical Theory and Computation*, vol. 12, no. 8, pp. 4100–4113, 2016.

[151] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1781–1802, 2005.

[152] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, no. 15, pp. 1875–1882, 2007.

[153] R. C. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[154] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC bioinformatics*, vol. 5, p. 113, 2004.

[155] G. Yachdav, S. Wilzbach, B. Rauscher, R. Sheridan, I. Sillitoe, J. Procter, S. E. Lewis, B. Rost, and T. Goldberg, "MSAViewer: interactive JavaScript visualization of multiple sequence alignments," *Bioinformatics*, vol. 32, no. 22, pp. 3501–3503, 2016.

[156] A. Cavazzuti, G. Paglietti, W. N. Hunter, F. Gamarro, S. Piras, M. Loriga, S. Allecca, P. Corona, K. McLuskey, L. Tulloch, F. Gibellini, S. Ferrari, and M. P. Costi, "Discovery of potent pteridine reductase inhibitors to guide antiparasite drug development," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 1448–1453, 2008.

[157] V. Cody, J. R. Luft, and W. Pangborn, "Understanding the role of Leu22 variants in methotrexate resistance: Comparison of wild-type and Leu22Arg variant mouse and human dihydrofolate reductase ternary crystal complexes with methotrexate and NADPH," *Acta Crystallographica Section D: Biological Crystallography*, vol. 61, no. 2, pp. 147–155, 2005.

[158] J. P. Volpato, B. J. Yachnin, J. Blanchet, V. Guerrero, L. Poulin, E. Fossati, A. M. Berghuis, and J. N. Pelletier, "Multiple conformers in active site of human dihydrofolate reductase F31R/Q35E double mutant suggest structural basis for methotrexate resistance," *Journal of Biological Chemistry*, vol. 284, no. 30, pp. 20079–20089, 2009.

[159] N. Schormann, S. E. Velu, S. Murugesan, O. Senkovich, K. Walker, B. C. Chenna, B. Shinkre, A. Desai, and D. Chattopadhyay, "Synthesis and characterization of potent inhibitors of Trypanosoma cruzi dihydrofolate reductase," *Bioorganic & medicinal chemistry*, vol. 18, no. 11, pp. 4056–4066, 2010.

[160] M. S. Hipp, S. H. Park, and U. U. Hartl, "Proteostasis impairment in protein-misfolding and -aggregation diseases," *Trends in Cell Biology*, vol. 24, no. 9, pp. 506–514, 2014.

[161] R. I. Morimoto, "Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging," *Genes and Development*, vol. 22, no. 11, pp. 1427–1438, 2008.

[162] J. Kirstein-Miles, A. Scior, E. Deuerling, and R. I. Morimoto, "The nascent polypeptide-associated complex is a key regulator of proteostasis," *The EMBO Journal*, vol. 32, pp. 1451–1468, 2013.

[163] H. Rampelt, J. Kirstein-Miles, N. B. Nillegoda, K. Chi, S. R. Scholz, R. I. Morimoto, and B. Bukau, "Metazoan Hsp70 machines use Hsp110 to power protein disaggregation," *The EMBO journal*, vol. 31, no. 21, pp. 4221–35, 2012.

[164] P. Goloubinoff, A. Mogk, A. P. Zvi, T. Tomoyasu, and B. Bukau, "Sequential mechanism of solubilization and refolding of stable protein aggregates by a bichaperone network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 24, pp. 13732–7, 1999.

[165] D. A. Parsell, A. S. Kowal, M. A. Singer, and S. Lindquist, "Protein disaggregation mediated by heat-shock protein Hsp104," *Nature*, vol. 372, no. 6505, pp. 475–478, 1994.

[166] J. Shorter, "The mammalian disaggregase machinery: Hsp110 synergizes with Hsp70 and Hsp40 to catalyze protein disaggregation and reactivation in a cell-free system," *PloS one*, vol. 6, no. 10, p. e26319, 2011.

[167] J. Yochem, H. Uchida, M. Sunshine, H. Saito, C. P. Georgopoulos, and M. Feiss, "Genetic analysis of two genes, dnaJ and dnaK, necessary for Escherichia coli and bacteriophage lambda DNA replication," *Molecular & general genetics : MGG*, vol. 164, no. 1, pp. 9–14, 1978.

[168] H. H. Kampinga and E. A. Craig, "The HSP70 chaperone machinery: J proteins as drivers of functional specificity," *Nature reviews. Molecular cell biology*, vol. 11, no. 8, pp. 579–92, 2010.

[169] N. B. Nillegoda, J. Kirstein, A. Szlachcic, M. Berynskyy, A. Stank, F. Stengel, K. Arnsburg, X. Gao, A. Scior, R. Aebersold, D. L. Guilbride, R. C. Wade, R. I. Morimoto, M. P. Mayer, and B. Bukau, "Crucial HSP70 co-chaperone complex unlocks metazoan protein disaggregation," *Nature*, vol. 524, no. 7564, pp. 247–251, 2015.

[170] N. B. Nillegoda and B. Bukau, "Metazoan Hsp70-based protein disaggregases: emergence and mechanisms," *Frontiers in molecular biosciences*, vol. 2, no. 57, 2015.

[171] N. B. Nillegoda, A. Stank, D. Malinverni, N. Alberts, A. Szlachcic, A. Barducci, P. D. L. Rios, R. C. Wade, Bukau, and Bernd, "Evolution of an intricate J-protein network driving protein disaggregation by the Hsp70 chaperone machinery," *Submitted*, 2017.

[172] J. Labbadia, S. S. Novoselov, J. S. Bett, A. Weiss, P. Paganetti, G. P. Bates, and M. E. Cheetham, "Suppression of protein aggregation by chaperone modification of high molecular weight complexes," *Brain*, vol. 135, no. 4, pp. 1180–1186, 2012.

[173] J. Hageman, M. A. Rujano, M. A. W. H. van Waarde, V. Kakkar, R. P. Dirks, N. Govorukhina, H. M. J. Oosterveld-Hut, N. H. Lubsen, and H. H. Kampinga, "A DNAJB Chaperone Subfamily with HDAC-Dependent Activities Suppresses Toxic Protein Aggregation," *Molecular Cell*, vol. 37, no. 3, pp. 355–369, 2010.

[174] J. Gillis, S. Schipper-Krom, K. Juenemann, A. Gruber, S. Coolen, R. Van Den Nieuwendijk, H. Van Veen, H. Overkleeft, J. Goedhart, H. H. Kampinga, and E. A. Reits, "The DNAJB6 and DNAJB8 protein chaperones prevent intracellular aggregation of polyglutamine peptides," *Journal of Biological Chemistry*, vol. 288, no. 24, pp. 17225–17237, 2013.

[175] J. Li, X. Qian, and B. Sha, "The Crystal Structure of the Yeast Hsp40 Ydj1 Complexed with Its Peptide Substrate," *Structure*, vol. 11, no. 12, pp. 1475–1483, 2003.

[176] Z. Lu and D. M. Cyr, "The conserved carboxyl terminus and zinc finger-like domain of the co- chaperone Ydj1 assist Hsp70 in protein folding," *Journal of Biological Chemistry*, vol. 273, no. 10, pp. 5970–5978, 1998.

[177] M. E. Cheetham and A. J. Caplan, "Structure, function and evolution of DnaJ: conservation and adaptation of chaperone function," *Cell stress & chaperones*, vol. 3, no. 1, pp. 28–36, 1998.

[178] N. Guex, M. C. Peitsch, and T. Schwede, "Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective," *Electrophoresis*, vol. 30, pp. S162–173, 2009.

[179] J. Kopp and T. Schwede, "The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models," *Nucleic acids research*, vol. 32, no. Database issue, pp. D230–4, 2004.

[180] F. Kiefer, K. Arnold, M. Künzli, L. Bordoli, and T. Schwede, "The SWISS-MODEL Repository and associated resources," *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D387–392, 2009.

[181] H. Suzuki, S. Noguchi, H. Arakawa, T. Tokida, M. Hashimoto, and Y. Satow, "Peptide-binding sites as revealed by the crystal structures of the human Hsp40 Hdj1 C-terminal domain in complex with the octapeptide from human Hsp70," *Biochemistry*, vol. 49, no. 39, pp. 8577–8584, 2010.

[182] Y. Q. Qian, D. Patel, F. U. Hartl, and D. J. McColl, "Nuclear magnetic resonance solution structure of the human Hsp40 (HDJ-1) J-domain," *Journal of molecular biology*, vol. 260, no. 2, pp. 224–235, 1996.

[183] Y. Wu, J. Li, Z. Jin, Z. Fu, and B. Sha, "The crystal structure of the C-terminal fragment of yeast Hsp40 Ydj1 reveals novel dimerization motif for Hsp40," *Journal of Molecular Biology*, vol. 346, no. 4, pp. 1005–1011, 2005.

[184] H. Y. Yu, T. Ziegelhoffer, J. Osipiuk, S. J. Ciesielski, M. Baranowski, M. Zhou, A. Joachimiak, and E. A. Craig, "Roles of intramolecular and intermolecular interactions in functional regulation of the Hsp70 J-protein co-chaperone Sis1," *Journal of Molecular Biology*, vol. 427, no. 7, pp. 1632–1643, 2015.

[185] M. Pellecchia, T. Szyperski, D. Wall, C. Georgopoulos, and K. Wüthrich, "NMR Structure of the J-domain and the Gly/Phe-rich Region of the Escherichia coli DnaJ Chaperone," *Journal of Molecular Biology*, vol. 260, no. 2, pp. 236–250, 1996.

[186] X.-C. Gao, C.-J. Zhou, Z.-R. Zhou, M. Wu, C.-Y. Cao, and H.-Y. Hu, "The C-terminal Helices of Heat Shock Protein 70 Are Essential for J-domain Binding and ATPase Activation," *Journal of Biological Chemistry*, vol. 287, no. 8, pp. 6044–6052, 2012.

[187] B. Sha, S. Lee, and D. M. Cyr, "The crystal structure of the peptide-binding fragment from the yeast Hsp40 protein Sis1," *Structure*, vol. 8, no. 8, pp. 799–807, 2000.

[188] T. R. M. Barends, R. W. W. Brosi, A. Steinmetz, A. Scherer, E. Hartmann, J. Eschenbach, T. Lorenz, R. Seidel, R. L. Shoeman, S. Zimmermann, R. Bittl, I. Schlichting, and J. Reinstein, "Combining crystallography and EPR: crystal and solution structures of the multidomain cochaperone DnaJ," *Acta crystallographica. Section D, Biological crystallography*, vol. 69, no. Pt 8, pp. 1540–1552, 2013.

[189] G. Vriend, "WHAT IF: A molecular modeling and drug design program," *Journal of Molecular Graphics*, vol. 8, no. 1, pp. 52–56, 1990.

[190] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids," *Journal of the American Chemical Society*, vol. 118, no. 15, pp. 11225–11236, 1996.

[191] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[192] J. Tsai and M. G. Douglas, "A conserved HPD sequence of the J-domain is necessary for YDJ1 stimulation of Hsp70 ATPase activity at a site distinct from substrate binding," *Journal of Biological Chemistry*, vol. 271, no. 16, pp. 9347–9354, 1996.

[193] W. C. Suh, C. Z. Lu, and C. A. Gross, "Structural features required for the interaction of the Hsp70 molecular chaperone DnaK with its cochaperone

DnaJ," *Journal of Biological Chemistry*, vol. 274, no. 43, pp. 30534–30539, 1999.

[194] R. R. Gabdoulline and R. C. Wade, "Simulation of the diffusional association of barnase and barstar," *Biophysical journal*, vol. 72, no. 5, pp. 1917–1929, 1997.

[195] A. H. Elcock, R. R. Gabdoulline, R. C. Wade, and J. McCammon, "Computer simulation of protein-protein association kinetics: acetylcholinesterase-fasciculin," *Journal of Molecular Biology*, vol. 291, no. 1, pp. 149–162, 1999.

[196] R. R. Gabdoulline and R. C. Wade, "On the Contributions of Diffusion and Thermal Activation to Electron Transfer between Phormidium laminosum Plastocyanin and Cytochrome f: Brownian Dynamics Simulations with Explicit Modeling of Nonpolar Desolvation Interactions and Electron Transfer Event," *Journal of the American Chemical Society*, vol. 131, no. 26, pp. 9230–9238, 2009.

List of Figures

List of Tables

6

Appendix

Table 6.1: List of used crystal or NMR structure and related information for JDs and CTDs.

Domain, Class	Name	PDB ID	Range	Resolution	UniProt Acc. Number
JD, A	DNAJA1	2lo1	1 - 70	(NMR)	P31689
JD, A	DNJ–12	2och	3 - 68	1.86 Å	O45502
JD, A	$DnaJ_{E.\ coli}$	1xbl	2 - 72	(NMR)	P08622
JD, B	DNAJB1	1hdj	1 - 70	(NMR)	P25685
JD, B	Sis1	2o37	3 - 70	1.25 Å	P25294
JD, B	DNAJB2	2lgw	1 - 71	(NMR)	P25686
JD, B	DNAJB8	2dmx	1 - 71	(NMR)	Q8NHS0
CTD, A	Ydj1	1nlt & 1xao	110 - 378	2.70 Å	P25491
CTD, B	DNAJB1	3agz	156 - 340	2.51 Å	P25685
CTD, B	Sis1	1c3g	179 - 349	2.70 Å	P25294

Table 6.2: List of homology modeled three-dimensional structures and related information. Those with the Ydj1* as the template structure used the modeled Ydj1 structure as template that was based on the PDB IDs 1nl1 and 1xao. The one marked with ** was modeled using the model of $DnaJ_{A.\ aceti}$ that was also modeled with the Ydj1 model.

Domain, Class	Name	Template [PDB ID]	UniProt Acc. Number, Seq. Range	Seq. id. [%], QM4
JD, A	DNAJA2	2lo1	O60884, 4–72	69.23%, -0.49
JD, A	Ydj1	2lo1	P25491, 1–71	66.67%, -2.38
JD, A	$DnaJ_{S.\ typhi}$	1xbl	P0A1G8, 2–72	95.77%, -1.02

Domain, Class	Name	Template [PDB ID]	UniProt Acc. Number, Seq. Range	Seq. id. [%], QM4
JD, A	DnaJ$_{K.\ pneumoniae}$	1xbl	C4T9C4, 2–72	95.77%, -0.87
JD, A	DnaJ$_{P.\ oryzihabitans}$	1xbl	A0A0D7F716, 2–72	70.42%, -2.09
JD, A	DnaJ$_{B.\ pertusis}$	1xbl	Q7VVY3, 2–72	71.83%, -1.2
JD, A	DnaJ$_{A.\ aceti}$	4j80	A0A063X4A7, 2–73	52.63%, -0.64
JD, A	DnaJ$_{Sphingomonas\ sp.}$	4rwu	Q1NCH5, 2–72	47.95%, -1.68
JD, A	DnaJ$_{C.\ ultunese}$	1xbl	M1ZGL1, 1–70	56.70%, -2.03
JD, A	DnaJ$_{A.\ thaliana}$	4rwu	Q94AW8, 13–80	61.76%, -0.38
JD, B	DNAJB4	1hdj	Q9UDY4, 1–70	73.13%, -0.24
JD, B	DNJ–13	1hdj	Q20774, 1–76	71.62%, 0.02
JD, B	CbpA$_{E.coli}$	3ucs	P36659, 2–70	100%, 1.26
JD, B	CbpA$_{S.\ typhi}$	3ucs	P63262, 2–71	98.57%, 1.23
JD, B	CbpA$_{K.\ pneumoniae}$	3ucs	W9BQH2, 2–71	88.57%, 1.02
JD, B	CbpA$_{P.\ oryzihabitans}$	3ucs	A0A0D7FE35, 2–68	71.64%, 0.69
JD, B	CbpA$_{B.\ pertusis}$	4j80	J7RE62, 4–70	51.43%, 0.3
JD, B	CbpA$_{A.\ aceti}$	4j7z	A0A063XA16, 3–72	45.71%, -0.51
JD, B	CbpA$_{Sphingomonas\ sp.}$	2dmx	Q1NEX3, 1–70	50.00%, -2.13
JD, B	CbpA$_{C.\ ultunese}$	2yua	M1ZLZ3, 1–72	38.46%, -1.91
JD, B	CbpA$_{A.\ thaliana}$	2m6y	F4JY55, 3–76	66.13%, -2.0
CTD, A	DNAJA1	Ydj1*	P31689, 102–369	44.53%, -2.73
CTD, A	DNAJA2	Ydj1*	O60884, 111–379	47.55%, -2.38
CTD, A	DNJ–12	Ydj1*	O45502, 101–368	39.69%, -2.86
CTD, A	DnaJ$_{E.coli}$	Ydj1*	P08622, 113–365	32.81%, -6.95
CTD, A	DnaJ$_{S.\ typhi}$	Ydj1*	P0A1G8, 116–375	33.33%, -7.85
CTD, A	DnaJ$_{K.\ pneumoniae}$	Ydj1*	C4T9C4, 114–373	32.94%, -7.2
CTD, A	DnaJ$_{P.\ oryzihabitans}$	Ydj1*	A0A0D7F716, 117–373	30.62%, -5.17
CTD, A	DnaJ$_{B.\ pertusis}$	Ydj1*	Q7VVY3, 127–385	29.30%, -6.41
CTD, A	DnaJ$_{A.\ aceti}$	Ydj1*	A0A063X4A7, 117–371	27.60%, -9.40
CTD, A	DnaJ$_{Sphingomonas\ sp.}$	Ydj1**	Q1NCH5, 119–372	54.00%, -7.48
CTD, A	DnaJ$_{C.\ ultunese}$	Ydj1*	M1ZGL1, 110–366	27.27%, -11.82
CTD, A	DnaJ$_{A.\ thaliana}$	Ydj1*	Q94AW8, 116–374	45.28%, -6.12
CTD, A	DNAJA4	Ydj1*	Q8WW22, 103–369	46.39%, -7.74
CTD, B	DNAJB4	3agz	Q9UDY4, 152–335	70.81%, 1.65
CTD, B	DNJ–13	3agz	Q20774, 154–327	55.44%, 1.33

Domain, Class	Name	Template [PDB ID]	UniProt Acc. Number, Seq. Range	Seq. id. [%], QM4
CTD, B	CbpA$_{E.coli}$	3lz8	P36659, 114–303	84.46%, -0.76
CTD, B	CbpA$_{S.\ typhi}$	3lz8	P63262, 114–303	81.31%,-0.86
CTD, B	CbpA$_{K.\ pneumoniae}$	3lz8	W9BQH2, 114–302	99.48%, -0.42
CTD, B	CbpA$_{P.\ oryzihabitans}$	3lz8	A0A0D7FE35, 123–311	55.96%, -1.42
CTD, B	CbpA$_{B.\ pertusis}$	3lz8	J7RE62, 125–310	43.62%, -3.25
CTD, B	CbpA$_{A.\ aceti}$	4j80	A0A063XA16, 132–299	32.10%, -2.67
CTD, B	CbpA$_{Sphingomonas\ sp.}$	3lz8	Q1NEX3, 139–310	32.75%, -4.75
CTD, B	CbpA$_{C.\ ultunese}$	3lz8	M1ZLZ3, 129–293	32.93%, -4.25
CTD, B	CbpA$_{A.\ thaliana}$	3agz	F4JY55, 166–347	48.90%, -0.7