Computational Studies of Protein Dynamics and Dynamic Similarity

Márton Münz

Structural Bioinformatics and Computational Biochemistry Unit,
Department of Biochemistry

Systems Biology Doctoral Training Centre

St. Cross College, University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity Term 2012

Abstract

At the time of writing this thesis, the complete genomes of more than 180 organisms have been sequenced and more than 80000 biological macromolecular structures are available in the Protein Data Bank (PDB). While the number of sequenced genomes and solved three-dimensional structures are rapidly increasing, the functional annotation of protein sequences and structures is a much slower process, mostly because the experimental determination of protein function is expensive and time-consuming. A major class of *in silico* methods used for protein function prediction aim to transfer annotations between proteins based on *sequence or structural similarities*. These approaches rely on the assumption that homologous proteins of similar primary sequences and three-dimensional structures also have similar functions. While in most cases this assumption appears to be valid, an increasing number of examples show that proteins of highly similar sequences and/or structures can have different biochemical functions. Thus the relationship between the divergence of protein sequence, structure and function is more complex than previously anticipated.

On the other hand, there is mounting evidence suggesting that minor changes of the sequences and structures of proteins can cause large differences in their conformational dynamics. As the intrinsic fluctuations of many proteins are key to their biochemical functions, the fact that very similar (almost identical) sequences or structures can have entirely different dynamics might be important for understanding the link between sequence, structure and function. In other words, the *dynamic similarity* of proteins could often serve as a better indicator of functional similarity than the similarity of their sequences or structures alone. Currently, little is known about how proteins are distributed in the 'dynamics space' and how protein motions depend on structure and sequence. These problems are relevant in the field of protein design, studying protein evolution and to better understand the functional differences of proteins. To address these questions, one needs a precise definition of *dynamic similarity*, which is not trivial given the complexity of protein motions.

This thesis is intended to explore the possibilities of describing the similarity of proteins in the 'dynamics space'. To this end, novel methods of characterizing and comparing protein motions based on molecular dynamics simulation data were introduced. The generally applicable approach was tested on the family of PDZ domains; these small protein-protein interaction domains play key roles in many signalling pathways. The methodology was successfully used to characterize the dynamic dissimilarities of PDZ domains and helped to explain differences of their functional properties (e.g. binding promiscuity) also relevant for drug design studies. The software tools developed to implement the analysis are also introduced in the thesis. Finally, a network analysis study is presented to reveal dynamics-mediated intramolecular signalling pathways in an allosteric PDZ domain.

Acknowledgements

Despite my best efforts, I would not have gotten far without the help and support of many people. First and foremost, I would like to thank my supervisors, Dr Phil Biggin and Prof Jotun Hein for their guidance, support and patience. I have learned a lot from both of them. Our consultations were very enjoyable, immensely helpful and many times put new life into my work.

I am especially thankful to Phil for working with me hard from the beginning to the end of my project, and for his assistance in writing up papers and ultimately my thesis. I would also like to thank Jotun for guiding me to this opportunity to study at the University of Oxford.

I am grateful to Prof Mark Sansom and Prof Martin Noble for their advices and encouragement. A special thanks goes to Dr Rune Lyngsø and Joseph Herman for their massive contribution to my work and the countless brainstorming sessions. I also have to thank Dr Ranjit Vijayan, Dr Oliver Beckstein, Dr Benjamin Hall, Ádám Novák, Sumeet Agarwal, Dr István Miklós, William Duncan, Shahd Horaib and Dr. Aziz Mithani for the helpful and invaluable discussions.

Being a student of the Systems Biology Doctoral Training Centre was a fantastic experience, and in particular, I would like to thank Dr Charlotte Deane and Prof David Gavaghan for their assistance and support.

In addition, I was lucky enough to belong to both the Structural Bioinformatics & Computational Biochemistry Unit (Department of Biochemistry) and the Genome Analysis Bioinformatics Group (Department of Statistics) and in both groups I met many inspiring people. I am especially grateful to the Doctoral Training Centre, the Department of Biochemistry, the Department of Statistics and Prof Thomas Darden (NIEHS, North Carolina, USA) to provide generous funding for my research.

However, most importantly, nothing would have been possible without the love of my family. They believed in me and supported me in all possible means from the beginning to the end. I am very lucky to have such wonderful parents who always stand by me and

have given me the opportunity to find my own way in life. Although they were 900 miles away, I felt them very close in these past years. It has always been my dream to make Anyapa proud. This thesis is for them. But also for my grandparents, Vera, Nagymama, Pali and Manó, all of whom I take with myself wherever I go. I am also very thankful to my mother-in-law, Marika néni and father-in-law, Zoli bácsi, who created the most loving environment for us when we spent some time in Hungary.

Finally, but above all, I thank my wonderful wife, Brigi, without whom this adventure surely would have been impossible. She is special. No words can describe how much she has helped me from day to day to keep going with my work and never give up. With all the hard work she has put into these four years, she would well deserve a DPhil degree herself. I cannot tell how grateful I am to her.

It is to Brigi, the love of my life, this thesis is dedicated.

Contents

List of Publications

1. Münz, M. and Biggin, P.C. (2012). JGromacs: a Java package for analyzing protein simulations. *J Chem Inf Model*, 52(1):255-9

2. Münz, M., Lyngsø R., Hein, J. and Biggin, P.C. (2010). Dynamics based alignment of proteins: an alternative approach to quantify dynamic similarity. *BMC Bioinformatics*, 11:188

3. Münz, M., Hein, J. and Biggin, P.C. (2012). The role of flexibility and conformational selection in the binding promiscuity of PDZ domains. *PLoS Comp Biol*, 8(11):e1002749

List of Figures

List of Tables

Chapter I

Introduction

From all we have learnt about the structure of living matter, we must be prepared to find it working in a manner that cannot be reduced to the ordinary laws of physics. And that not on the ground that there is any 'new force' or what not, directing the behaviour of the single atoms within a living organism, but because the construction is different from anything we have yet tested in the physical laboratory.

- Erwin Schrödinger, *What is Life?* (1956)

1.1 A single molecule as a complex system

Life has no secret ingredient. Living organisms are composed of the same chemical elements (mostly Hydrogen, Oxygen, Carbon, Nitrogen, Sulphur and Phosphorus) that also make up non-living material. What distinguishes a living system from its inanimate environment is its extraordinary complexity. In other words, the key to understanding a biological system is not really to learn *what* building blocks it is composed of, instead to puzzle out *how* these components are organized to function together as a system.

Although a living cell is indeed nothing more than a collection of molecules (as is a *non-living* object), the interaction networks of these biomolecules form a complex machinery that is incomparably more complicated and more precise than the most complex man-made machines. From this point of view, biology, now becoming increasingly interdisciplinary[1], shares many challenges with engineering sciences. While the objective in engineering is to design a system that can perform a desired function, the goal in modern biology is the opposite: to reverse-engineer complex biological systems in order to

understand how they are constructed to perform the observed functions.[2,3]

Complex Systems Theory, a relatively new and rapidly emerging field in mathematical, physical, social and life sciences, deals with systems which consist of a large number of mutually interacting components.[4] In such systems, characteristic features can emerge that cannot be understood by studying the components of the system in isolation. This property (i.e. "the whole is more than the sum of its parts") is often referred to as the irreducibility of the complex systems and is particularly true in the case of biological systems that are made up of a very large number of interacting components.[4,5] Emergent functions arising from complexity are observed on every levels of organization, from molecular to cellular to ecological systems.(see Figure 1.1).

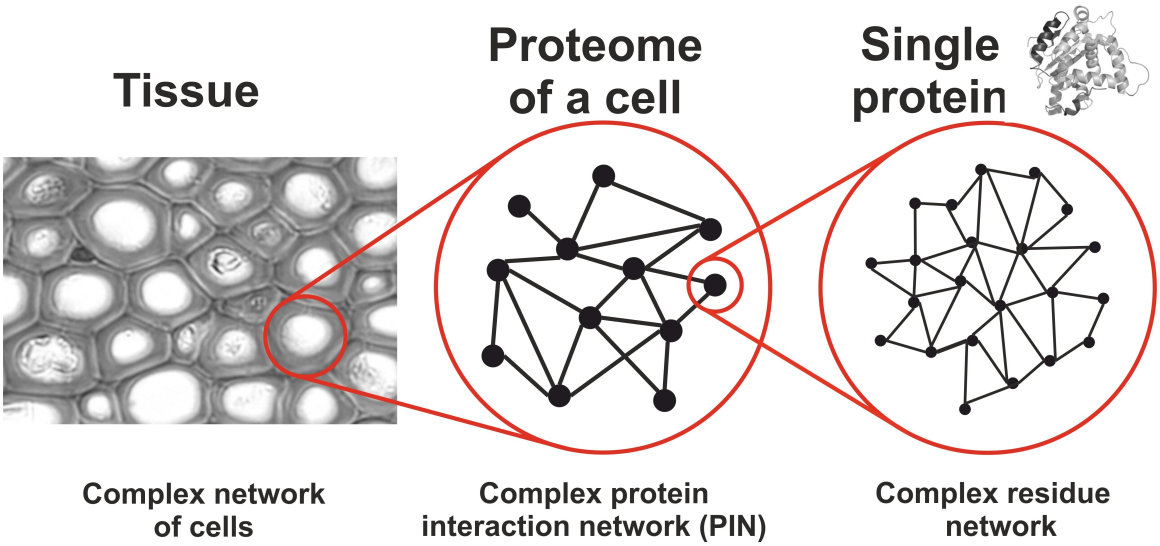

Figure 1.1: It is their complexity that distinguishes living systems from their inanimate environment. Complexity can be observed at every level of biological systems. A tissue is a a complex network of cells. The proteome within a single cell is a complex interaction network of proteins. An individual protein is made up of a complex network of amino acid residues. Therefore to fully understand how a protein performs its biochemical function, one needs a system-level approach.

Systems Biology is a fast developing integrative discipline that aims to gain a system-level understanding of biological systems and processes learning from Complex Systems Theory.[5–7] In contrast with the traditional reductionist approach of life sciences that aimed to decompose living systems into its basic components (cells, proteins, genes etc.) and understand the functions of these individual parts, the objective of systems biology is to

explain how these components work together as a complex dynamical system.[5,6]

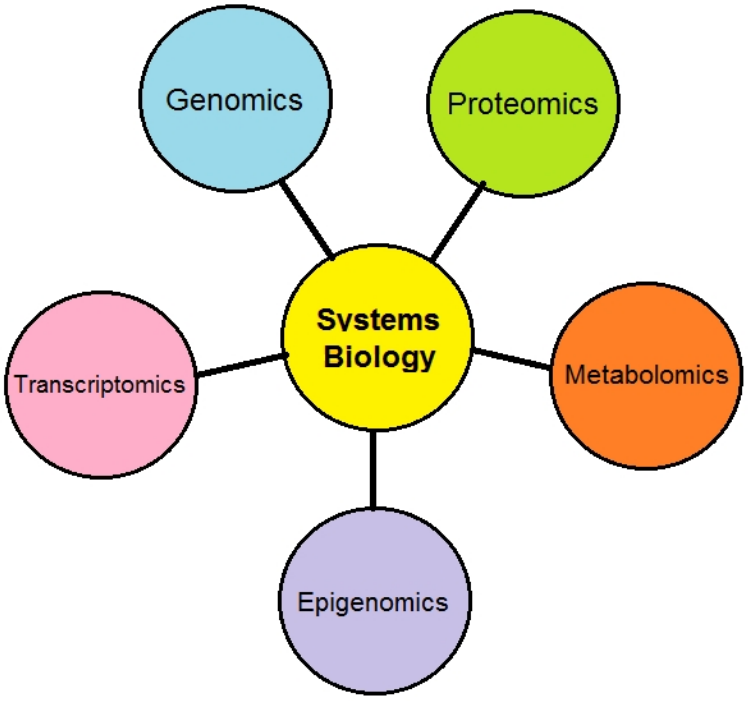

Figure 1.2: Solving the molecular puzzle of life: systems biology views a living cell as a complex system and aims to understand its emergent functional properties by the integration and system-level modelling of genomics, proteomics, transcriptomics, metabolomics and epigenomics data.

With the completion of genome programs and emergence of high-throughput experimental technologies (e.g. microarray, next-generation sequencing and mass spectrometry), the amount of available data about the building blocks of cells, their interactions and dynamical behaviour is rapidly increasing.[8,9] The challenge of systems biology is to integrate the genomics, transcriptomics, proteomics, metabolomics and epigenomics data (Figure 1.2), focusing on the interactions between the components, to reverse-engineer emergent cellular functions.[10] This challenge, however, is especially difficult due to the complexity of living systems: for example, it has been estimated that a single typical eukaryotic cell contains approximately one billion protein molecules.[11]

Although the major focus of Systems Biology is the analysis of protein interaction, gene regulatory and metabolic networks, even understanding a single protein molecule imposes

a very difficult challenge. Composed of a large number of interacting components (i.e. amino acid molecules), an individual protein molecule itself is a complex system (Figure 1.1). Therefore similarly to the approach of studying complex protein interaction networks, in order to understand how a particular protein carries out its specific function, one should think of it as an irreducible system which is more than the sum of its parts. In other words, to elucidate the biochemical function of a protein, one would need to gain a system-level understanding of the molecule.[12]

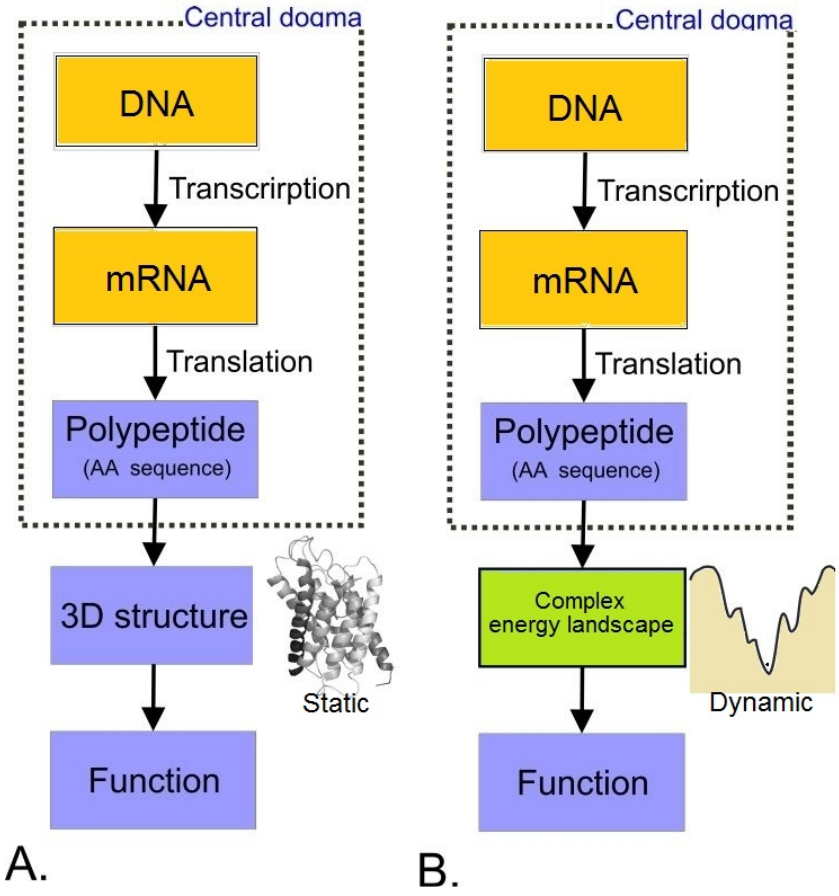

Figure 1.3: Information flow from DNA sequence to biochemical function. According to the central dogma of molecular biology, for protein-coding genes, information is transferred from DNA to RNA to the polypeptide chain (primary sequence of protein). In the outdated view (**A.**), the amino acid sequence encodes the native tertiary structure of the protein which determines its biochemical function. In reality (**B.**), the amino acid sequence encodes a complex free-energy landscape that determines the dynamic nature of the protein. Instead of a single native structure, proteins exist in an ensemble of conformations and their flexibility is key to their biochemical functions.

Figure 1.4: Being dynamic molecules, proteins cannot be described as a single structure, but exist in conformational ensembles. As an example, the solution NMR ensemble of apo calmodulin from Xenopus laevis (PDB: 1cfc) is presented. All 25 conformers in the ensemble are superposed using a weighted superposition method (weights indicated in different colors). The apparent structural variation of calmodulin is due to the intrinsic dynamics of the protein. (Image courtesy: Mechelke and Habeck 2010[13])

Although proteins are complex systems because they are composed of a network of interacting amino acid residues, their real complexity arises from their dynamic nature. According to the traditional view (Figure 1.3A), the primary amino acid sequence of a protein encodes a well-defined three-dimensional (tertiary) structure, and this native structure determines the function of the protein.[14] However, this oversimplified model of sequence-structure-function relationship provides a static picture of the protein. In reality, proteins are not rigid bodies but flexible molecules and their mobility is key to their functions.[15]

What is really encoded in the primary sequence is not a single conformation, but a complex free-energy landscape that determines the conformational dynamics of the protein (Figure 1.3B). Consequently, instead of adopting a single native structure, proteins exist in an ensemble of conformations (illustrated with the example of calmodulin in Figure 1.4). [16] As discussed in Section 1.2.1, the topography of the free-energy landscape determines the complex hierarchy of the functionally relevant substates of the protein.

A generally observed feature of complex systems is the emergence of collective, self-organized behaviour that cannot be inferred from the interactions between the individual components. Such collective behaviour were described in many complex biological phenomena such as the motion of bacterial colonies, nest construction of social insects, flocking of birds, migration of cancer cells, activation of neurons and function of the immune system. [17–22] Being a complex dynamical system, an individual protein molecule also exhibits collective behaviour: i.e. the collective fluctuation of its atoms. As discussed in Section 1.2.3, large-scale concerted motions of atoms in a protein are often crucially implicated in its biochemical function and are therefore under evolutionary selection pressure. [23]

1.2 The dynamic nature of proteins

1.2.1 The topography of the energy landscape

In contrast to many textbooks that depict them as static molecules of a well-defined three-dimensional structure, proteins are not rigid bodies, but flexible systems that constantly change their shape under physiological temperature. [16] Protein motions (ranging from small atomic fluctuations to collective movements of entire domains or subunits) are crucially implicated in the function of many proteins. For instance, the ability to change conformation has been found to be key in the function of several enzymes, transport and signalling proteins and proteins involved in the immune system (see a few examples below). [24] As discussed in the previous section, it is the energy landscape of a protein that determines its native conformational dynamics.

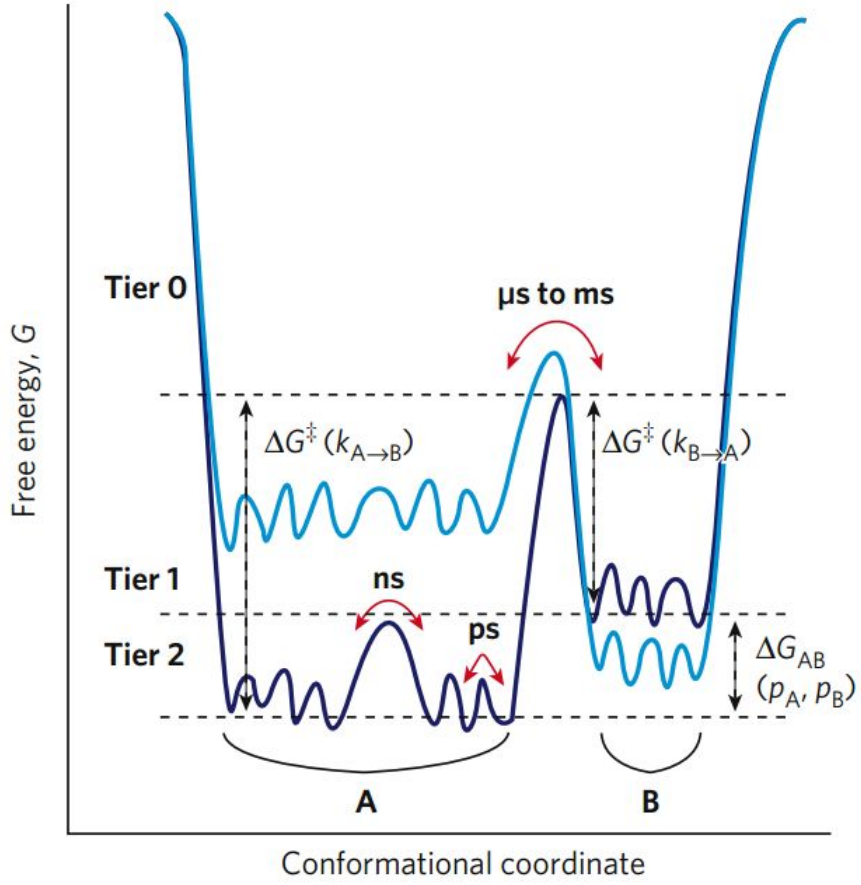

Figure 1.5: Simplified illustration (one-dimensional cross-section) of the high-dimensional energy landscape that determines the organization of substates and the hierarchy of timescales in protein dynamics. The relative populations of the two tier-0 states A and B (p_A and p_B) depend on their free energy difference (ΔG_{AB}). Furthermore, the rate of interconversion between the two states ($k_{A\to B}$ and $k_{B\to A}$) depend on the energy barriers separating them ($\Delta G^{\ddagger}(k_{A\to B})$ and $\Delta G^{\ddagger}(k_{B\to A})$). In contrast to the interconversion between tier-0 states that occur on the μs to ms timescale, the transition between tier-1 or tier-2 substates correspond to faster fluctuations described on the ps to ns timescales. Perturbations (e.g. ligand binding or mutation) can change the free-energy landscape (e,g. from the dark blue to the light blue curve). (Image courtesy: Henzler-Wildman and Kern 2007 [16])

The complex topography of the high-dimensional energy landscape is an intrinsic property of proteins[16] which encodes their folding, stability and function-related conformational changes[25]. The energy landscape is known to have a hierarchical structure and contains multiple local minima corresponding to different substates of the protein.[26] The relative populations of these conformational substates depend on their energies, while the rates of interconversion between them are determined by the heights of energy barriers

(see Figure 1.5).[16,26]

The hierarchical structure of the energy landscape is also directly related to the hierarchy of timescales of protein dynamics. One can distinguish between protein motions happening on the 'slow' timescale (also referred to as tier-0 dynamics) and motions happening on the 'fast' timescale (referred to as tier-1 and tier-2 dynamics). The former are the consequences of fluctuations between states separated by several k_BT energy barriers (i.e. tier-0 states), while the latter are the results of fluctuations between structurally similar states separated by energy barriers of less than k_BT located within the well of a particular tier-0 state.(Figure 1.5) The rare transitions between different tier-0 states correspond to large-amplitude, collective motions on the microsecond-to-millisecond timescale. By contrast, tier-2 and tier-1 transitions correspond to frequent, small-amplitude, local motions observed on the picosecond-to-nanosecond timescale, respectively.[16,27]

Since the topography of the energy landscape (and the resulting conformational dynamics) is a unique, inherent characteristics, Henzler-Wildman and Kern called this property the 'dynamic personality' of the protein.[16] Importantly, the functional properties of most proteins are closely related to their 'dynamic personalities', as discussed in the next subsections.

1.2.2 Collective motions of atoms

Collective motions (i.e. concerted fluctuation of groups of atoms) have been found to play crucial roles in the biochemical functions of many proteins.[28] Recurring movements were classified according to their types such as 'hinge', 'shear', 'twist' or 'screw' motions[29,30] which appear on multiple levels of the structure. For example, the Database of Macromolecular Movements (molmovdb)[29] provides a multi-level classification system that discriminates between loop, domain and subunit motions. However, since these motions are very complex and occur in a correlated fashion, instead of classifying them into discrete types, they were suggested to be better described by overlapping, "fuzzy" flexibility classes.[31]

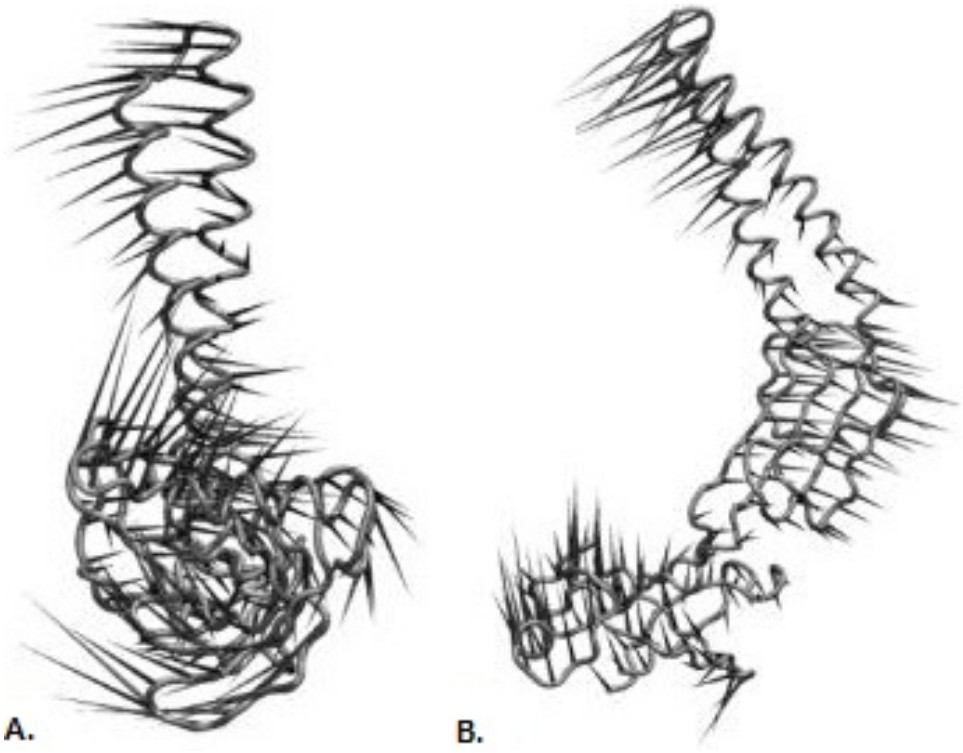

Figure 1.6: Most dominant modes of motions of the multidrug resistance protein mexA calculated using Principal Component Analysis (PCA) based on a 25 ns molecular dynamics simulation.[32] The first two eigenvectors (principal components) are visualized as "porcupine" representation with the Dynamite web server[33]. In this plot, a cone is assigned to each α-carbon atom pointing in the direction of the eigenvector for that atom, while its length is proportional to the amplitude of motion. **A:** Visualization of the first eigenvector (which describes 53% of the total motion) shows rotation of the α-helix at the C-terminal end of the β-barrel subdomain. **B:** The plot of the second eigenvector (which describes 20% of the total motion) reveals a hinge-bending motion between the β-domain and the α-helical hairpin. (Image courtesy: Vaccaro et al 2006[32])

In order to derive collective motions of atoms from conformational ensembles or Elastic Network Models (ENM), data mining methods such as Principal Component Analysis (PCA) or Normal Mode Analysis (NMA)[34,35] are often applied. (These mathematical techniques are described in Chapter 2.) A series of studies that used PCA to reduce the dimensionality of molecular dynamics trajectories have concluded that only a small set of principal components account for most of the variation in protein motion[36–39]. In other words, internal protein motions can usually be described appropriately by a few eigenvectors (of the covariance matrix), that are referred to as the *essential dynamics* modes (see an example in Figure 1.6). As discussed above, these collective motions correspond to

transitions between major conformational substates.

The challenge of relating large-scale collective motions to functional properties of proteins has been discussed in a number of papers and reviews[40]. In many studies the analysis of collective motions helped to understand the function of proteins[41–44].

1.2.3 Flexibility and function: examples

Several examples are known in which conformational flexibility is essential for the biochemical functions of proteins. Without trying to give a comprehensive list, only a few of these examples are mentioned here in order to illustrate the diversity of functional roles conformational dynamics may play in proteins.

First of all, protein motions are crucially implicated in various aspects of enzymatic functions. For instance, in dihydrofolate reductase (DHFR) which catalyzes the NADPH-dependent reduction of 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF), conformational changes that arise from fast thermal motions were suggested to facilitate the hydride transfer reaction. The intrinsic motions of the enzyme, substrate and cofactor are involved in reducing the donor-acceptor distance, creating a favourable electrostatic environment for the reaction and finding the correct orientation of the substrate and cofactor. [45] Another example is streptococcal pyrogenic exotoxin B (SpeB), a cysteine protease, in which large amplitude loop movements were shown to play key role in enzyme activation.[15] Upon removal of the pro-domain from the protein, a loop (called the 'latency loop') undergoes large fluctuation (resulting in >25 Å displacement) followed by the motion of a second loop (called the 'switch loop') that is now allowed to move away from the active site, thereby activating of the enzyme.

Furthermore, collective atomic fluctuations in enzymes often play important roles in substrate binding and product release. For instance, in Triosephosphate Isomerase (TIM) which catalyzes the isomerization of dihydroxyacetone phosphate (DHAP) to D-glyceraldehyde 3-phosphate (GAP), access to the active site of the enzyme is controlled by the constant fluctuation of a flexible loop (see Figure 1.7). This loop acts as a rigid lid that performs opening and closing motions which can also be observed in the absence of ligand.

[46,47] Similarly, in Adenylate Kinases (ADKs) that are responsible for catalyzing the interconversion of adenine nucleotides, collective domain motions which play essential roles in opening the nucleotide binding lids are required for product release. In addition, there appears to be a strong link between the described collective motions of the enzyme and the catalytic turnover of different hyperthermophilic and mesophilic ADK homologs.[48] However, protein dynamics are important for ligand binding and release not only in enzymes. For example, large-scale protein fluctuations on the nanosecond-to-microsecond timescale were found to be crucial for ligand escape in myoglobin.[49]

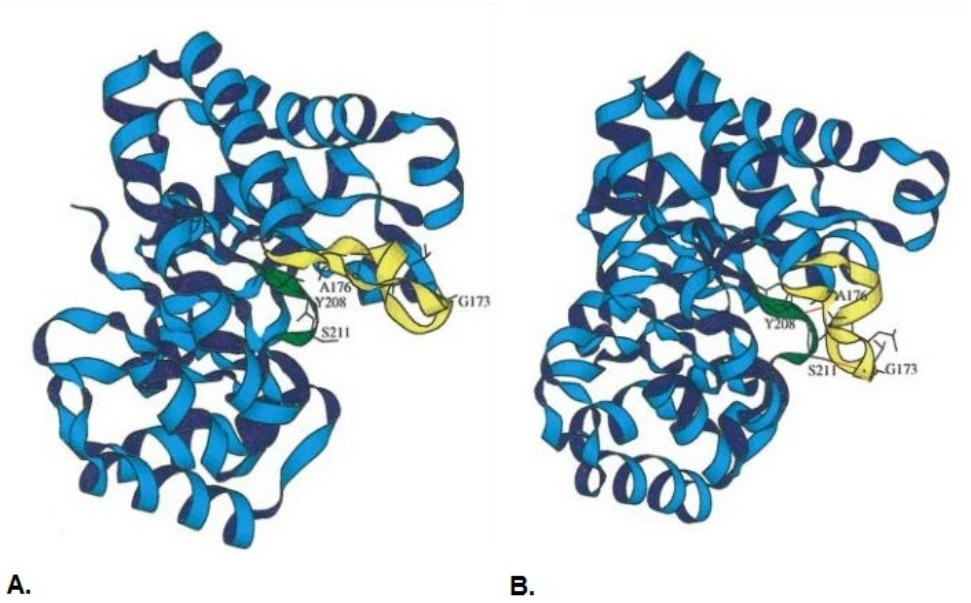

Figure 1.7: Functionally important loop motion in Triosephosphate Isomerase (TIM) is one of many examples where conformational dynamics play crucial role in biochemical functions. Comparison of the open (**A.**) and closed (**B.**) conformation of the subunit 1 of the enzyme shows that the flexible 11-residue loop (in yellow) acts as lid at the active site. Residues also found to be crucial for proper loop opening and closure are highlighted in green. (Image courtesy: Derreumaux and Schlick 1998 [46])

For many proteins, the conformational selection mechanism was found to be key to molecular recognition of single or multiple interaction partner(s). Such proteins actually visit the different binding conformations required for forming complexes with their ligands. Therefore, in these cases, the intrinsic fluctuations of proteins are necessary for accessing the required conformations and participating in the binding interactions. Exam-

ples in which conformational selection was shown to play crucial role include the highly flexible ubiquitin[50] and the SPE7 antibody[51]. (Both proteins and the mechanism of conformational selection are discussed in details in Chapter 5.)

Protein motions are also involved in ion channel gating as exemplified by K^+ channels, acid-sensing ion channels (ASICs) and nicotinic acetylcholine receptors (nAChRs). For instance, in KcsA potassium channel, millisecond-timescale conformational dynamics were found to play a key role in gating.[52,53] In particular, KcsA fluctuates between the open and closed state of the pore and these cooperative motions involved in gating of the channel appear to be also coupled to the selectivity filter region which controls selective permeation [52,53]. Similarly, in the acid-sensing ion channel ASIC1, gating was found to be governed by collective motions of the protein: the rotation of the extracellular domain and the collective dynamics of the thumb and finger domains result in a "twist-to-open" motion of the channel pore.[54] Finally, in case of the nicotinic acetylcholine receptor nAChR, collective motions of different domains (i.e. extracellular domain, transmembrane domain, intracellular domain and ligand-binding-sites) were found to be responsible for the opening-closing mechanism of the channel.[55]

In addition to ion channels, large-scale domain motions are important for the function of many proteins. For instance, in Alcohol Dehydrogenase (ADH) that catalyzes the interconversion between ethanol and acetaldehyde, correlated interdomain motions were shown to play crucial role in binding and releasing of a necessary cofactor (NAD^+).[56] Another example is the thermostable DNA-polymerase I from Thermus aquaticus (Taq-polymerase) in which coupled domain motions were found to be involved in its complex function of nucleotide synthesis and cleavage.[57]

As reviewed by Smock et al.[58], intrinsic protein fluctuations are also of key importance in intercellular and intracellular signalling. The mechanisms by which signalling proteins transmit information between one another and the role of conformational dynamics is increasingly well understood. In most cases, upstream signals (e.g. interactions with partner proteins, peptides and small ligands, or covalent modifications such as phosphorylation) remodel the free-energy landscape of the signalling protein altering its dynamics.[58]

The change of dynamics may alone be responsible for the propagation of the signal throughout the protein without any apparent conformational change to happen. An example is the phosphotyrosine-binding domain of insulin receptor substrate-1 (IRS-1) that is crucial for transmitting signals from hormone-activated insulin receptors to intracellular pathways. While IRS-1 undergoes only slight conformational changes upon binding to a phosphotyrosine-containing peptide, a pathway of dynamically altered residues connecting the peptide-binding site and a distal surface was identified.[59] Interestingly, while only minimal differences were found between the apo and peptide-bound IRS-1 structures, comparison of NMR relaxation data showed significant changes between the dynamics of backbone amide groups and side-chain methyl groups for a number of residues, some of which are located remotely from the binding site. Furthermore, these dynamically altered residues were found to form a pathway connecting the binding site with a distal (and probably functionally important) site of the protein. As the authors suggest, this dynamics-mediated pathway may play crucial role in intramolecular communication by transmitting signals between the binding site and the distal site without prominent conformational change to happen. Similarly, in the dimeric catabolite activator protein (CAP), allosteric communication between the two subunits was found to be transmitted by changes of dynamics.[60] (See Section 6.2.1 for more details of catabolite activator protein and dynamically driven allostery.)

1.2.4 Experimental and computational methods

The amount of data available about protein dynamics is rapidly growing due to the advances of experimental and computational methods that allow to study the conformational flexibility of biological macromolecules.[61] One of the most important evidences of conformational dynamics of proteins often comes from the comparison of snapshots representing their different substates (i.e. tier-0 states). A number of experimental methods can provide such atomic (or near-atomic) resolution snapshots including X-ray crystallography, NMR (nuclear magnetic resonance) spectroscopy, cryo-electron microscopy and small-angle X-ray scattering.[16]

In addition to structural information, X-ray crystallography experiments provide B-factor profiles (also referred to as the Debye-Waller factor) which characterize the thermal mobility of atoms.[62,63] However, since the B-factors depend on both real atomic fluctuations and lattice disorders, their interpretation as a measure of intrinsic mobility of the protein is difficult.[64]

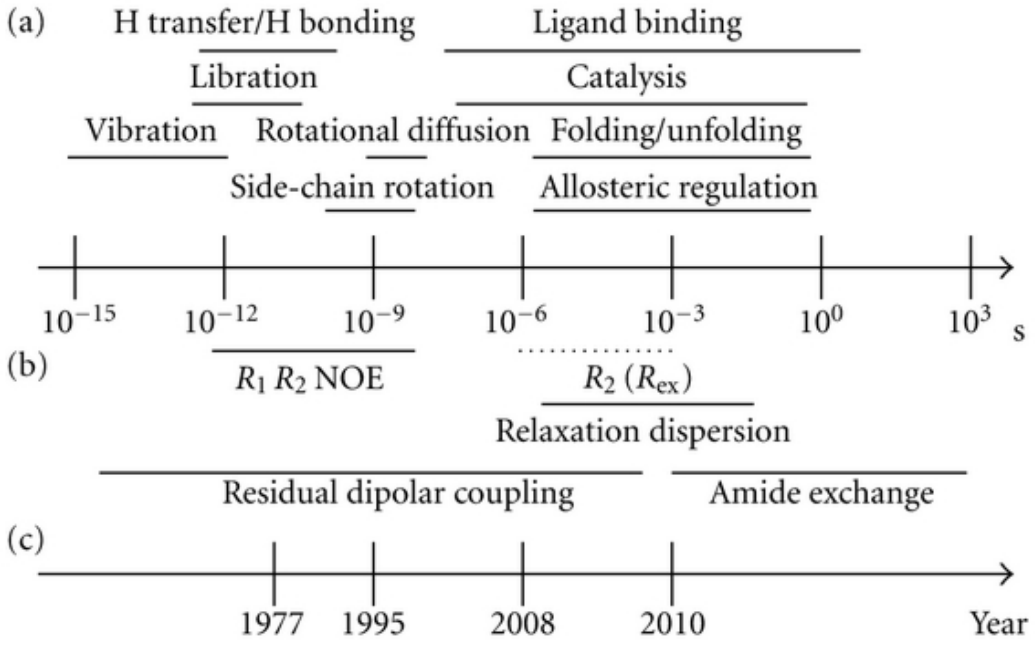

Figure 1.8: A wide range of timescales of protein dynamics (**A.**) can be captured using different NMR techniques (**B.**). The approximate years from which the same timescales can be studied with molecular dynamics simulations.(**C.**) (Image courtesy: Fisette et al. 2012[65])

On the other hand, NMR spectroscopy also serves as a powerful tool that allows to study protein dynamics covering a broad range of timescales (see Figure 1.8A and B). For example, nuclear spin-relaxation measurements can capture both fast fluctuations happening on the ps to ns timescales and slow fluctuations observed on the μs to ms timescales with atomic resolution.[66] Although due the technical limitations, NMR experiments were originally feasible only for small (<30 kDa), soluble proteins, thanks the advances of modern spectrometer technology, molecules as large as 100 kDa can be investigated today with this method.[67] Furthermore, in addition to solution NMR techniques, solid-state NMR (ss-

NMR) spectroscopy can be used to measure insoluble proteins and this technique does not have protein size limitations.[68]

In the same time, fluorescence methods applied at the single-molecule level are opening promising new possibilities to study protein dynamics by allowing to monitor the real-time behaviour of individual molecules.[69] Single-molecule FRET (fluorescence resonance energy transfer) is a sensitive technique which can measure the change of the relative distance between two fluorophores linked to specific sites of the protein.[70] Although the method can provide information only about a single characteristic distance, smFRET was successfully used to follow the conformational dynamics of proteins (e.g. the proton-powered subunit rotation of F_0F_1-ATP synthase[71]) and protein folding (e.g. the folding and unfolding dynamics of adenylate kinase[72]).

In addition, computational methods allow the *in silico* exploration of the conformational space accessible to proteins of interest. Molecular Dynamics (MD) simulations serve as a powerful tool commonly used to study the intrinsic motions of proteins.[73] MD simulations provide continuous trajectories of the conformational dynamics of biological macromolecules, giving an insight into the dynamics of the system at an atomic level of detail and with high (femtosecond) time-resolution (see a more detailed description of the technique in Chapter 2). However, due to the high computational cost of these calculations, the time-scales explored by MD simulations were traditionally limited to $\sim$100 ns to $\sim$1 μs and were therefore mostly suitable to study the tier-1 and tier-2 dynamics of proteins.[16] Recently, the accessed timescales were extended to 1 millisecond (Figure 1.8C).[74] In order to achieve better sampling of the conformational space, several accelerated MD variants have been developed[75] including High-Temperature Molecular Dynamics (HTMD)[76], Replica Exchange Molecular Dynamics (REMD)[77], Umbrella Sampling[78], Multiple Time-Step (MTS) methods[79] and Digitally Filtered Molecular Dynamics (DFMD)[80].

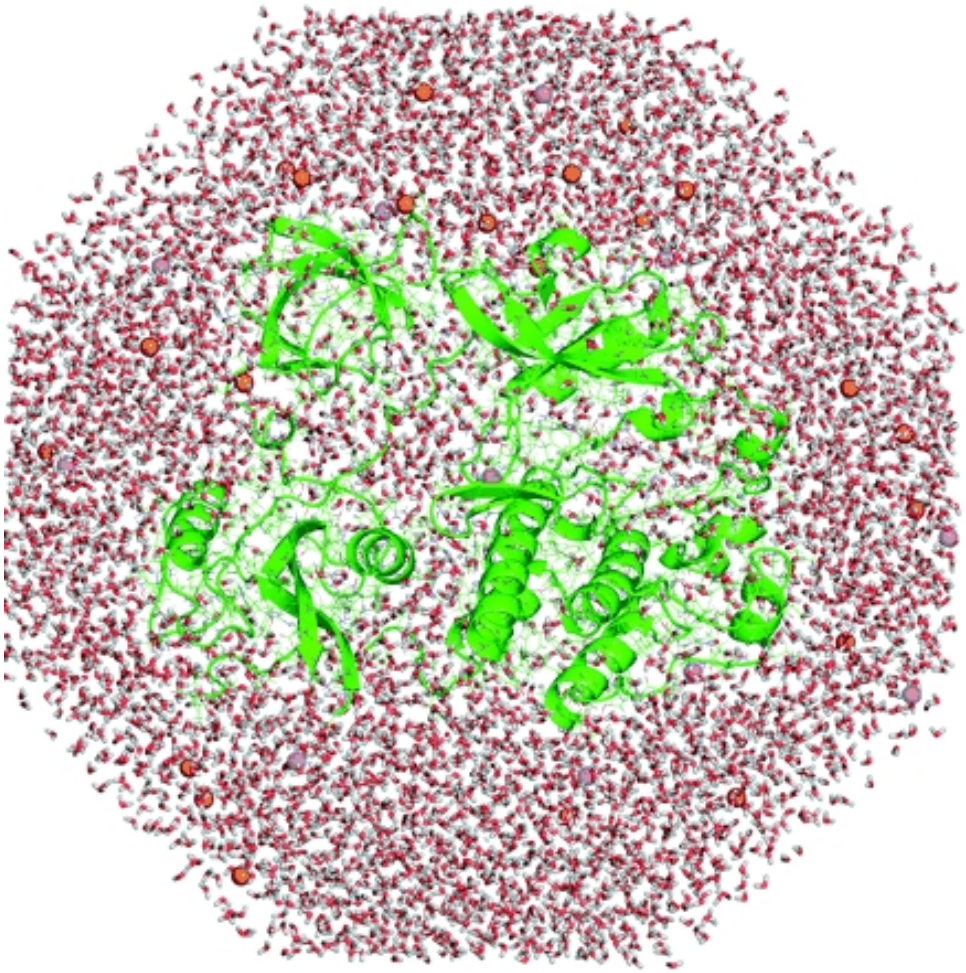

Figure 1.9: Snapshot of an all-atom molecular dynamics (MD) simulation of Src kinase protein (shown in green). The protein is solvated in a box of ∼15,000 water molecules (oxygen and hydrogen atoms shown in red and white, respectively). Potassium and chloride ions (shown as purple and orange spheres) are also added to the box. The whole simulation system consists of a total of ∼50,000 atoms. (Image courtesy: Karplus and Kuriyan. 2005[73])

Furthermore, in addition to all-atom MD (AT-MD) simulations that explicitly represent every atom of the biomolecular system, coarse-grained molecular dynamics (CG-MD) simulations that use simplified models of the system were shown to be effective.[81,82] CG-MD simulations represent groups of atoms as single particles ("pseudo-atoms") thereby significantly reducing the number of interactions to be calculated.[82] Although the predictions of CG-MD simulations are not as reliable as those of atomistic simulations, CG-MD models have the great advantage of enabling to simulate macromolecular systems of large

(i.e. submicrometric) sizes and on biologically relevant timescales (i.e. μs to ms).[81] Additionally, multiresolution approaches that combine low-resolution CG models with high-resolution all-atom models were developed that aim to unite the advantages of both levels of description.[83]

Alternatively, non-dynamic methods such as CONCOORD[84] can also be used for sampling the conformational space of proteins. The CONCOORD algorithm does not rely on any potential function, instead it can randomly generate a conformational ensemble around a known structure to satisfy a set of geometric constraints calculated from the strengths of interatomic interaction.

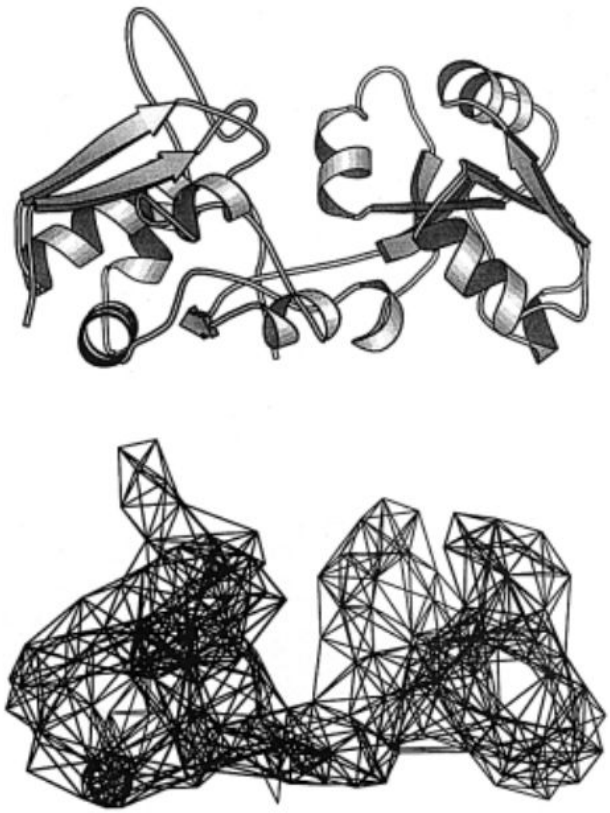

Figure 1.10: Elastic Network Model (ENM) of the lysine-arginine-ornithine (LAO) binding protein. The top figure shows the standard cartoon representation of the protein, while the bottom figure presents the residue network in which pairs of Cα-atoms closer than 8 Å are connected by harmonic springs. (Image courtesy: Tama and Sanejouand 2001[85])

Finally, Elastic Network Models (ENM) that are based on a simplified (coarse-grained)

representation of the molecules have also emerged as a widely used tool for studying protein fluctuations.[86] The protein structure is represented in an ENM as a network of particles which are connected by elastic springs (see Figure 1.10).[86] In Gaussian Network Models (GNM), the particles correspond to the α-Carbon atoms of the protein. However, depending on how much details the model incorporates, the particles may represent groups of atoms or representative points of amino acids or side chains (e.g. their center of mass). ENMs rely on a harmonic approximation of the free-energy surface around an equilibrium structure. Normal Mode Analysis (NMA) applied to ENM models (ENM-NMA) was effectively used in many studies to identify functionally important collective motions of proteins (see a more detailed description of the technique in Chapter 2).[87,88]

1.3 Comparison of protein dynamics

Since conformational dynamics play important roles in the biochemical functions of many proteins, the comparative analysis of protein motions may help us to better understand the similarities and differences of various functional properties. Several studies aimed to compare the dynamics of a given collection of proteins; only few examples are mentioned below (some of them are also discussed in Section 4.2).

Spiwok et al. have studied five cold-active enzymes (i.e. enzymes from cold-adapted organisms that are evolved to have high catalytic activity at low temperatures) to understand the functional differences between these proteins and their meso- or thermophilic counterparts[89] To this end, they have run molecular dynamics simulations for each pair of cold-active and meso/thermophilic enzymes and have performed comparative analysis of their residue fluctuation profiles and collective motions. The comparison of conformational dynamics has revealed key differences in the rate and extent of the opening/closing mechanism of active sites in these enzymes resulted from the adaptation to low temperatures.

In another example, Barreca et al. have performed a comparative molecular dynamics study of the wild-type and double mutant HIV-1 integrase (IN) in complex with a small-

molecule inhibitor (5CITEP).[90] HIV-1 integrase is one of the three essential enzymes (besides reverse transcriptase and protease) of the human immunodeficiency virus (type 1) and is consequently an emerging target of antiviral drugs.[91] Barreca et al. have compared the conformational dynamics of the wild-type integrase and the double mutant enzyme which contained mutations known to cause drug resistance. Their comparative method has identified significant dynamic differences between the wild-type and mutant proteins, in particular, regarding a loop located next to the active site. These results helped Barreca et al. to create a hypothetical model of HIV-1 IN inhibition and drug resistance.

Another example of comparative analysis of protein motions was an NMR study of two members of the lipid binding protein (LBP) family, H-FABP (heart fatty acid binding protein) and ILBP (ileal lipid binding protein).[92] In this study, comparison of ^{15}N NMR relaxation data has revealed that the backbone dynamics of the two structurally similar proteins are surprisingly different: while H-FABP is relatively rigid, ILBP was found to be highly flexible. This varying flexibility of different members of the LBP family was then suggested to be related to differences in their ligand-binding affinities.[93]

Finally, Tai et al. have compared the conformational dynamics of the active sites of four distantly related enzymes[94] including three hydrolases (acetylcholinesterase, outer-membrane phospholipase A and outer-membrane protease T) and a transferase (PagP). The comparative analysis was carried out based on molecular dynamics simulation data using the BioSimGrid framework[95]. Although these four enzymes do not share a common three-dimensional fold, their active sites are structurally similar. Tai et al. have found striking differences between the flexibility of the active sites in the four enzymes and have linked these distinguishing dynamic characteristics to catalytic mechanisms.

The examples discussed above illustrate that comparative analysis of dynamics can be an efficient approach to identify functionally important protein motions. Comparison of conformational flexibility of homologous, functionally-related proteins can lead us to better understanding of the role of dynamics in various functional properties. On the other hand, comparative analysis of dynamics can help to figure out how exactly protein motions are encoded in primary sequence and tertiary structure (see Section 1.4 for more

detailed discussion about the mapping between protein sequence, structure, dynamics and function).

1.4 Continuity and discontinuity of the protein universe

1.4.1 Three distinct layers of description

The exploration of the 'protein universe' (defined as the "collection of all proteins of every biological species that lives or has lived on earth"[96]) is one of the central problems of molecular biology and bioinformatics. Understanding the diversity and distribution of naturally occurring proteins has implications for example in protein evolution studies, protein engineering, function prediction and classification.[97–99] The 'protein universe' is most often described by multiple distinct layers called the sequence, structure and function space of proteins.[100] The first layer, the 'protein sequence space' is an abstract space representing all possible amino acid sequences where each point symbolize a particular protein sequence. While the theoretical size of protein sequence space (the number of all possible sequences) is practically infinite, only a tiny fraction of it corresponds to existing sequences.[101]

Several methods were developed for quantifying the similarity of protein sequences (i.e. measuring the distance between points of the sequence space). There are important differences between the alternative measures. For example, while the Needleman-Wunsch global sequence alignment algorithm[102] provides an overall sequence similarity score, the Smith-Waterman algorithm[103] quantifies local similarity between sequences (see Section 2.5). Some sequence alignment methods also use scoring matrices such as BLOSUM[104] or PAM[105] to incorporate evolutionary information. In contrast to the dynamic programming approach used by the Needleman-Wunsch and Smith-Waterman algorithms, tools such as BLAST (Basic Local Alignment Search Tool)[106], PSI-BLAST (Position-Specific Iterative BLAST)[107] and FASTA[108] rely on rapid heuristics to quantify the local similarity of protein sequences. A number of clustering and automatic classification methods based on these sequence similarity measures were proposed including CluSTr[109], Tribe-MCL[110]

and Systers[111]. Interestingly, it has been found that naturally occurring proteins are not evenly distributed in the sequence space, instead are organized in distinct clusters that correspond to sets of functionally and structurally similar proteins.[112–115]

The second layer of description is the 'protein structure space' which is an abstraction of the collection of all three-dimensional structures adopted by all sequences of the 'sequence space'. Thus each point of the 'structure space' represents a possible tertiary structure. Many approaches were developed for quantifying the similarity of protein structures (i.e. the distance between points in the structure space).[116] The alignment scores provided by structural alignment algorithms such as DALI[117], CE[118], VAST[119] and SSAP[120] serve as alternative measures of structural similarity. In addition, other algorithms such as PINTS [121] and ProBiS[122] are able to detect and measure local structural similarities of proteins (see Section 2.5). Several studies addressed the question of mapping the protein structure space using various similarity measures[99,123,124](see Figure 1.11) and have found that proteins of similar biochemical functions tend to be located adjacent in the structure space.

In order to understand the structure of the protein structure space, a number of structural classification systems have been created such as the manually curated SCOP (Structural Classification of Proteins)[125], the semi-automatic CATH (Class Architecture Topology Homology)[126] and the purely automatically generated FSSP (Families of Structurally Similar Proteins)[127]. Both the SCOP and CATH databases provide hierarchical classification and both assume a discretized picture of the protein structure space. In contrast, recent studies have pointed out that the protein structure space has a continuous rather than discrete nature[128] or at least has a complex discrete-continuous duality[129,130].

Finally, the third layer of description is the 'protein function space' that represents the whole set of biochemical functions carried out by proteins.[131] However, unlike protein sequence and structure, protein function is much more difficult to define, and therefore the first challenge is the standardization of terminology used to describe functional properties of proteins.[132,133] Several comprehensive classification systems of biochemical functions has been developed such as the Gene Ontology (GO) database[134] generally applicable to annotate molecular functions and the Enzyme Commission number (EC number) nomen-

clature[135] specifically used to classify enzyme-catalyzed reactions.

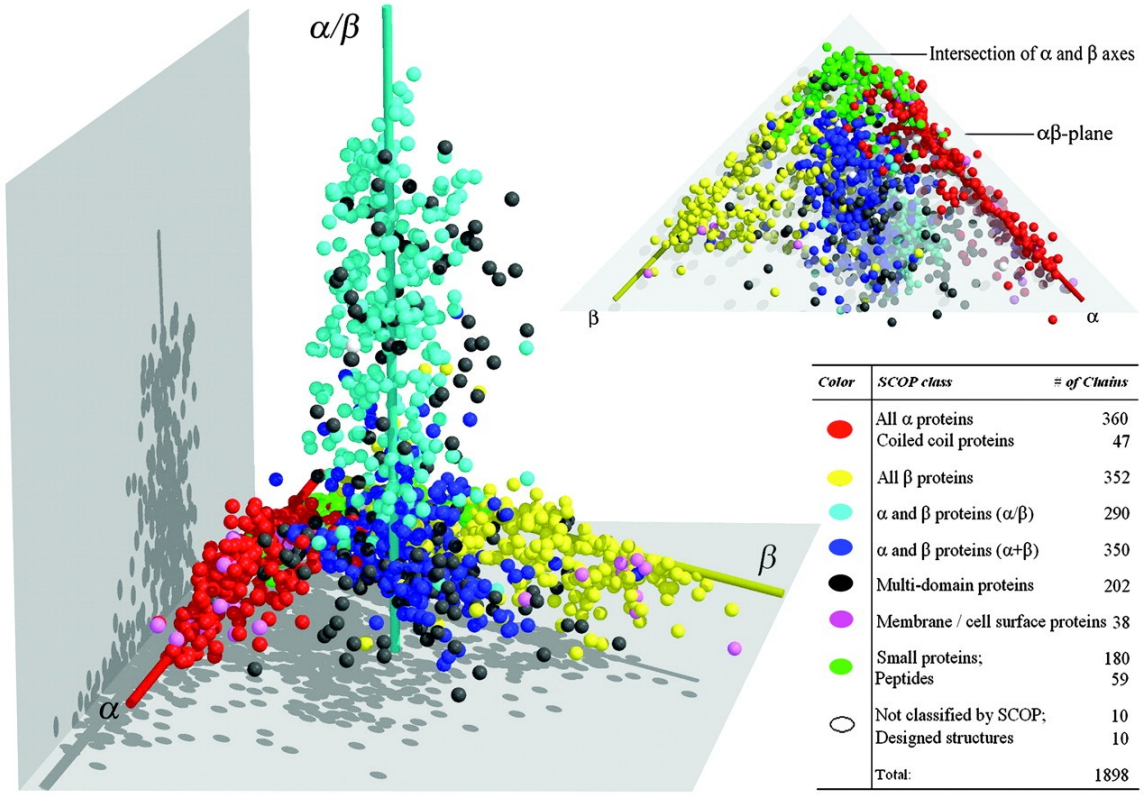

Figure 1.11: Several studies have aimed to analyse the topology of the protein structure space. For example, the above 3D map was created by Hou et al. using the multidimensional scaling (MDS) approach. The total of 1898 structures mapped (represented by spheres) are coloured according to the SCOP classes they belong to. In this map, proteins belonging to the α, β and α/β SCOP classes form separate, elongated clusters along the three axes. (Image courtesy: Hou et al. 2005[99])

Furthermore, it is even more difficult to define an applicable quantitative measure of the *similarity* of protein functions. However, this is especially important in order to study the topology of 'function space' in relation to the topology of sequence space and structure space. Alternative similarity scores were reviewed by Chagoyen et al.[136] and Pesquita et al.[137]. Several proposed measures of functional similarity rely on the GO annotation system[133,138–141], while some other similarity scores are based on the comparison of the EC numbers of enzymes[142].

1.4.2 Mapping between sequence, structure and function

One of the most important challenges in bioinformatics is to describe how the above discussed protein sequence, structure and function space map on each other.[123] This question is of central relevance for the problem of computational function prediction (see below), protein design and understanding the evolutionary processes that have shaped proteins. Putting it another way: studying the elementary steps of protein evolution can lead us to a better understanding of the current organization of protein universe.[143]

For example, following a gene duplication event, the resulting two copies of the protein-coding gene begin to accumulate mutations independently and the two proteins start to diverge.[144] Importantly, however, the rate of divergence of the two copies may be different within the protein sequence, structure and function space. In other words, two points that are close in one space, may map on two points that are distant in another space (based on some similarity measures). This is exactly what was found regarding the relationship of the sequence and structure space of proteins: sequence tends to diverge faster than structure. Many examples are known for proteins that share a highly conserved tertiary structure despite the absence of detectable sequence similarity.[145–147] For example, structural alignment and superposition of the GTPase domains of Ras p21 oncogene protein and elongation factor Tu (EF-Tu) show that the two proteins have very similar 3D-structures (αC-RMSD of 1.36 Å), despite their very low sequence identity of only 17%.[145]

Detecting remote evolutionary relationships (homology) based on sequence similarity of proteins becomes especially difficult if the proteins belong to the so-called "twilight zone", meaning that their sequence identity is not larger than the average sequence identity of random proteins (i.e. below a threshold of $\sim$20%).[148] In this case, structural similarity serves as a better indicator of homology because even if protein sequences diverge beyond recognition, the tertiary structures may remain similar.[149] As a result of the different rate of divergence in the sequence and structure space, despite the large number of possible protein sequences, the number of distinct structural folds is relatively small.[150] For example, in the latest release (1.75) of the SCOP database, the 38221 categorized PDB entries have been classified into only 1195 distinct structural folds.

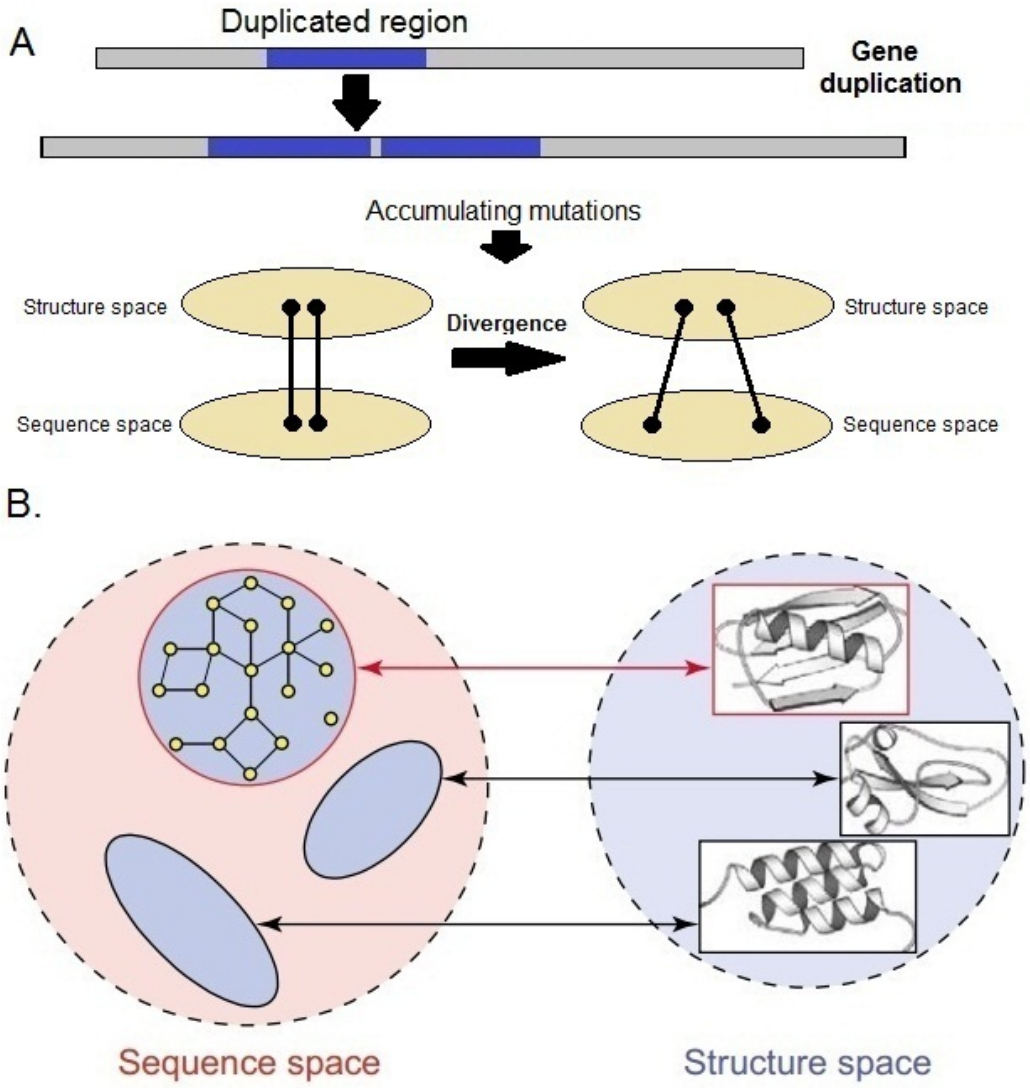

Figure 1.12: A. Following a gene duplication event, the two protein copies (which are identical in the beginning) start to diverge due to the accumulating mutations. However, the rate of their divergence is different in sequence space and structure space: i.e. their primary sequences tend to diverge faster than their tertiary structures. **B.** As a result, the same structural fold may be adopted by a set of closely related sequences which are organized into a network referred to as the neutral network. Individual sequences are represented by the nodes of the network and two sequences are connected if they differ in only a single point mutation. (Image courtesy: Xia and Levitt 2004[115])

In order to better understand the mapping between sequence and structure space and the dynamics of protein evolution, a number of studies[151,152] have used simplified protein models such as hydrophobic-polar (HP) 2D square lattice models.[153,154] The advantage of

these simplified models is that they allow to calculate the complete mapping between sequence and structure space. These studies have found that indeed many protein sequences may fold into the same structure and the sets of homologous sequences adopting identical structures are interconnected by single substitutions (point-mutations) forming complex networks referred to as the neutral network of the fold. These neutral networks were shown to have a "super-funnel" organization: i.e. each network is arranged around a "prototype sequence" which has the most neighbours in the network and the more we walk away from this sequence, the less stable the corresponding protein structures are.[151,152] It is important to note, that a single protein fold may map to multiple neutral networks in the sequence space.[115] The size of the neutral networks of a structure (i.e. the number of sequences that fold into the structure) is called the 'designability' of the fold[155,156] which is a property that varies across different structures.[155] In general, however, the results of lattice model studies suggest that naturally occurring proteins have high designability.[156]

Understanding the nature of the mapping from sequence space to structure space is particularly important for the field of protein structure prediction which develops *in silico* methods that aim to infer the three-dimensional structures of proteins from their amino acid sequences.[14] *De novo* (or *ab initio*) protein structure prediction try to solve this problem by using only sequence information.[158] By contrast, the homology modelling[159] and protein threading[160] approaches rely on templates (i.e. previously solved structures of homologous proteins). It is generally considered that for generating reliable homology models of a protein, at least 30% sequence identity with the template(s) is required.[159] However, increasing number of exceptions complicating the situation are known: for example, two proteins having 88% sequence identity were found to adopt significantly different 3D-structures (as discussed in more details below) (see Figure 1.13).[157,161]

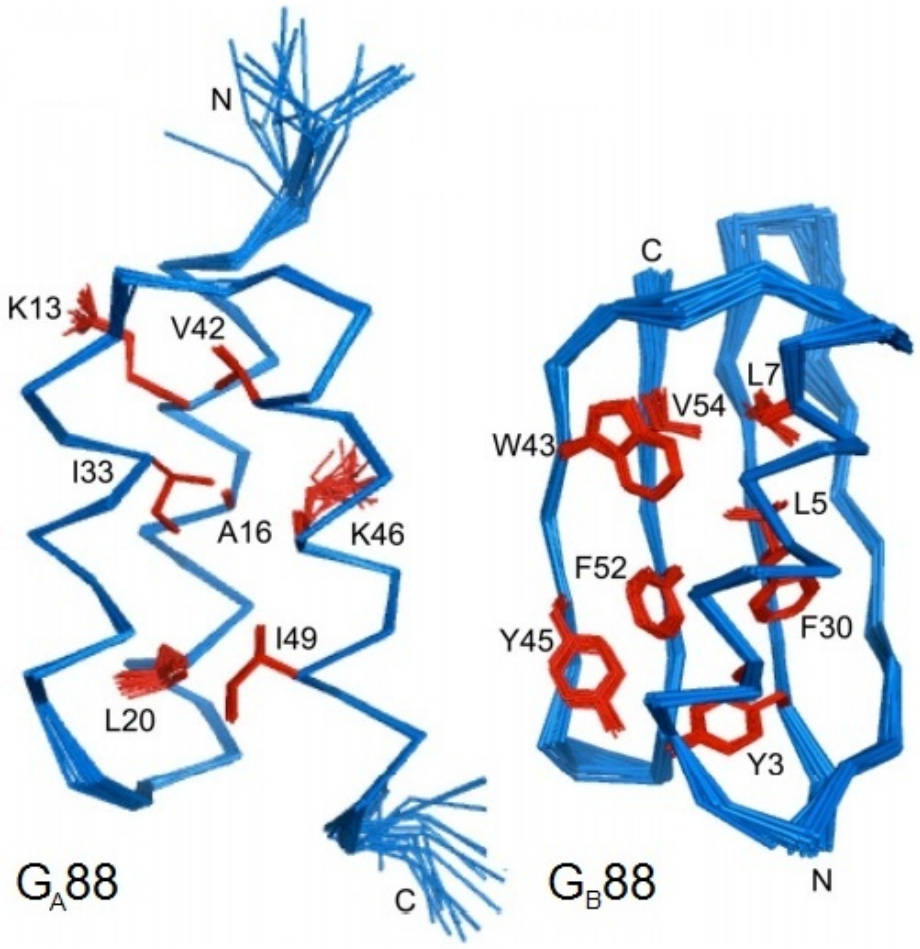

Figure 1.13: Comparison of the NMR ensembles (comprising 20 structures) of two engineered proteins (G_A88 and G_B88) that have 88% sequence identity, yet entirely different global structures and functions. Main chain atoms are shown in blue, hydrophobic core side-chains are highlighted in red. (Image courtesy: He et al. 2008[157])

The problem of describing the mapping from protein sequence and structure space to protein function space is even more complex. A series of studies aimed to understand the relationship between sequence similarity and function similarity.[162–165] For instance, Shah et al. have tried to discriminate between EC (Enzyme Commission) classes based on pairwise sequence similarity of the enzymes, and have found that most EC classes cannot be perfectly defined using any sequence similarity threshold. In other words, for most EC classes one can find at least one protein sequence outside the class that is more similar to a sequence inside the class than two sequences both belonging to the class.[162]

Although there is consistent correlation between sequence similarity and function similarity[163,164], in general, predicting protein functions based on solely sequence similarity (i.e. homology transfer) suffers from serious difficulties. In particular, the reliable transfer of functional annotations between similar protein sequences requires very high degree of sequence identity (> 70%)[165,166] and the problem becomes especially difficult if sequence identity is below 40%.[167] One of the reasons why the sequence-based function inference is even more challenging than the inference of structure is that certain residues are directly involved in function while the conservation of the fold is less sensitive to random mutations.[166] To summarize, sequence-based function prediction methods have limited accuracy pointing at the complex relationship between protein sequence space and function space.

Finally, the mapping between the structure space and function space has also been investigated in several studies both for theoretical and practical reasons (i.e.).[99,126,168,169] For example, Orengo et al. have analysed of the degree of functional similarity of protein structures with similar folds in the CATH database and have found that the vast majority (>90%) of homologous enzyme families comprise functionally similar proteins (i.e. those that share at least their first three EC identifiers).[126] As a consequence, functional annotations can be transferred between proteins that have significant global structural similarity [168]. On the other hand, however, the similarity of the global structure may not always correlate with functional similarity (see in the next subsection): for instance, proteins with different global folds can have similar local functional sites as a result of convergent evolution, or proteins with the same fold can have entirely different local functional sites as a result of divergent evolution.[55,170,171] Consequently, in many cases, local structural similarity may serve as a better indicator of functional similarity, as was demonstrated for example by Hwang et al. who have discovered the previously unknown nucleotide binding function of an archaeal protein based on local structural comparison.[172].

Hegyi and Gerstein have come to similar conclusions when studying the relationship between the structure and function of a large set of enzymes by comparing their classification in the SCOP and EC databases. They have found considerable variation in the

functional diversity of different protein folds. For example, the five functionally most versatile folds identified were the TIM-barrel, the Rossmann fold, the alpha-beta hydrolase fold, the ferredoxin fold and the P-loop containing NTP hydrolase fold. On the other hand, the most versatile enzymatic functions (i.e. which had the most structural folds associated with them) were the glycosidases and carboxylases.[169] Examples of proteins with different global structures but identical (or similar) biochemical functions include bacterial subtilisin and mammalian chymotrypsin which are both serine proteases despite adopting different folds.[169] These results point out the difficulty of the structure-based function prediction problem and shed light on the complex relationship between the protein structure space and function space.

1.4.3 Transition points in sequence and structure space

Sequence- and structure-based function prediction (homology transfer) methods work reasonably well because in most regions, the mapping from sequence space to structure space and to function space is continuous in the sense that "small" changes in one space correspond to "small" changes in the other space. Consequently, for example, similar protein sequences tend to map to similar structures and functions. However, the continuous nature of sequence→structure, sequence→function and structure→function maps is not generally the case. Instead, the maps were found to contain discontinuities: i.e. points at which small change in one space is accompanied by large change in the other space. For example, minor changes of the primary sequence (e.g. single point mutations) may in some cases result in large changes of the 3D-structure, even switching between entirely different folds. Those points where mapping between sequence, structure and function has such discontinuity will be termed as "transition points" (referring to the phenomenon that, for example, single point mutations in the sequence can cause transition between distinct structural folds).

As discussed in the previous subsection, the results of lattice model studies suggest that protein sequences adopting a common fold can be represented as a neutral network in which links connect neighbouring sequences (i.e. those that differ in single point mu-

tations).[151,152] Furthermore, it was found that distinct neutral networks may also be connected through the above-mentioned transition points (or "bridge states") (i.e. sequences that form bridges between different structural folds) (see Figure 1.14).[173,174] Such transition points in the sequence space are likely to be crucial in protein evolution as they facilitate abrupt changes of the global structure via gradual changes of the amino acid sequence by single point mutations.

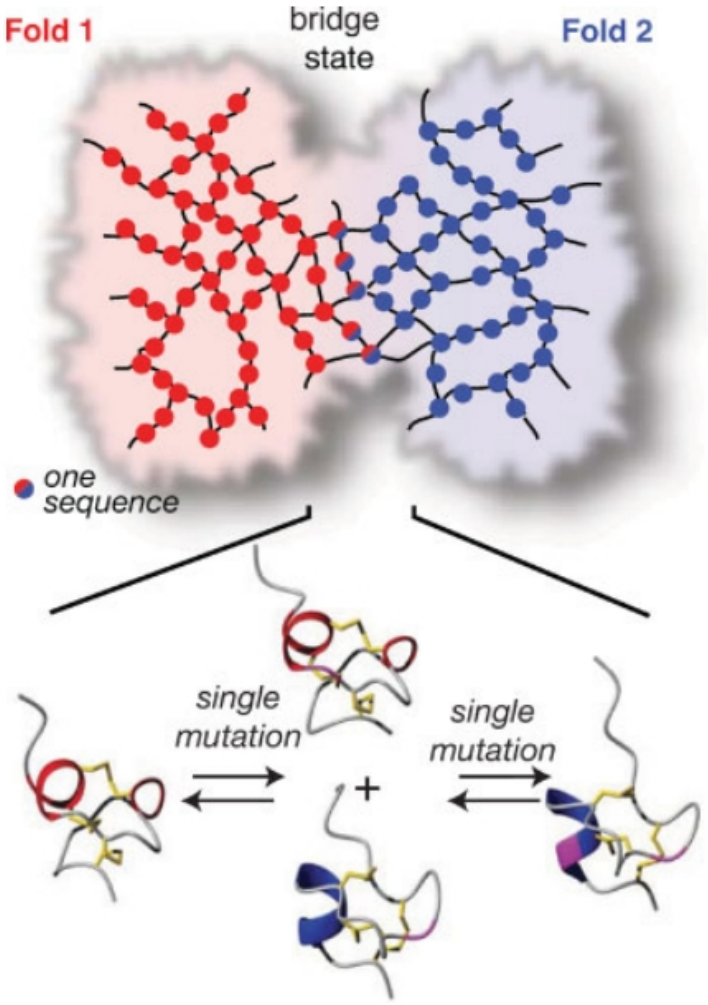

Figure 1.14: In sequence space, neutral networks of certain structural folds are connected by so-called "transition points" (also referred to as "bridge states") which enable evolutionary transitions between distinct folds by single mutational steps (top figure). Some transition points were found to correspond to "metamorphic proteins" which exist in a dynamic equilibrium between multiple folds as for example shown for cysteine-rich domains (bottom figure). Single point mutations can stabilize one of these alternative structures. (Image courtesy: Meier and Ozbeg 2007[174])

Although it will take further research to test the predictions of simplified lattice models on real proteins, several examples have already been reported of proteins that have highly similar sequences yet very different global three-dimensional structures.[161,175] For instance, Alexander et al. have shown that it was possible to redesign two small, naturally occurring proteins to create two sequences which have 88% sequence identity but have entirely different folds and different biochemical functions. One of the two engineered sequences adopts a 3-α helix fold capable of binding to albumin, while the other sequence adopts an α/β fold capable of binding to immunoglobulin G (IgG) (Figure 1.13). Since in these proteins only 12% of the amino acid positions (i.e. a total of 7 positions) encode the difference between the two structures, switching between the two folds requires only a few mutational steps.[157,161]

Moreover, a number of proteins referred to as "metamorphic proteins"[176] were shown to be able to adopt more than one native folds and can actually fluctuate between these very different global structures. Connecting the neutral networks of different folds, these metamorphic proteins were suggested to serve as transition points between dissimilar global structures and may therefore play key roles in protein evolution.[176] For example, in case of NW1, a naturally occurring cysteine-rich domain, a single point mutation was shown to switch to a protein that exists in an equilibrium between the original NW1 fold (with 22 % occupancy) and the distinct fold of another cysteine-rich domain, Mcol1C (with 78 % occupancy). Moreover, a further point mutation has lead to a domain which exclusively adopted the fold of Mcol1C (Figure 1.14).[174]

Additionally, as already discussed above, the fact that two proteins have highly similar global structures does not necessarily mean that they also have identical or similar biochemical functions. Numerous examples are known of proteins that share the same fold but have different active sites and therefore carry out diverse functions.[177,178] In this case, local structural similarity of the functional sites is often better indicator of functional correspondence than global similarity.[179] For example, the TIM barrel fold that consists of eight α-helices and eight parallel β-strands[180] serves as a highly versatile structural scaffold on which various different types of functional sites could evolve resulting in TIM barrel pro-

teins with a wide range of functions.[181] Moreover, even if two proteins have highly similar active sites in addition to adopting essentially identical folds, they may catalyze different chemical reactions, as is the case, for example, for mandelate racemase (MR) and muconate lactonizing enzyme (MLE)[182].

The sensitivity of protein functions to minor sequence and structure perturbations imposes limitations on homology-based function inference methods. On the other hand, the existence of transition points between sequence/structure and function also has important consequences for the evolution of new protein functions and protein engineering.

1.4.4 The fourth layer: protein dynamics space

The difficulty of inferring the tertiary protein structure from the primary sequence, or predicting the molecular function of a protein from its sequence and structure has been emphasized thus far. However, since the conformational dynamics of many protein are closely related to their biochemical functions (as discussed in 1.2.3), one can also use dynamic information about the protein to predict or understand its function. In other words, a fourth layer of description, here referred to as the 'protein dynamics space', can be introduced that may serve as a bridge between the sequence/structure space and function space of proteins. The concept of 'dynamics space' does not presently exist in the literature, mainly because, such like protein functions, the property of protein dynamics is very difficult to define. In addition, it is even more problematic to specify what one means on 'similarity' of protein motions and to develop quantitative measures of dynamic similarity. However, in order to compare the organization of the dynamic space with that of the sequence, structure and function space, exact definitions of protein dynamics and similarity measures are required. Some comparative protein dynamics studies (also discussed in Section 4.2) have experimented with different dynamic similarity measures[183–186], but the optimal way to quantify similarity of protein dynamics is yet to be established.

As discussed in Section 1.2.3, intrinsic protein motions often play important roles in biochemical functions, therefore studying the mapping from sequence and structure space to dynamics space is closely related to the problem of sequence and structure-based func-

tion prediction. However, little is known about how protein dynamics is encoded in sequence and structure.

The dependence of large-scale collective motions on the three-dimensional structures of proteins has been intensively studied. Keskin et al. have found that proteins with similar scaffolds (3D-folds) have similar cooperative global motions, however, individual structural segments may still exhibit distinct dynamic behaviour that allow a given fold to perform diverse functions.[187] Low-resolution coarse-grained ENM models, which represent the protein at the level of Cα-atoms, were shown to carry enough information to accurately predict the essential dynamics of proteins.[86,188,189]. The most important conclusion of these studies is that functionally important modes of global dynamics are in some extent independent from sequence, since coarse-grained GNM models can reproduce large-scale collective motions without using sequence information.

On the other hand, a series of studies have suggested that sequence has a more refined role in determining dynamics than just encoding the structural fold of the protein. Some studies have analysed the conservation of protein motions in relation to the conservation of amino acid sequence.[190–192] For example, by comparative analysis of Cα B-factor profiles for a large set of homologous proteins, Maguid et al. have found that backbone dynamics diverges slowly and can be conserved even across proteins that have highly diverged sequences.[190] These conclusions are in agreement with the two observations that proteins of similar folds tend to have similar large-scale motions and highly diverged sequences may adopt the same fold.

In a following study, Maguid et al. have analysed the evolutionary divergence of collective protein motions using Gaussian Network Models. Their results have shown that the most collective modes of motions (i.e. the lowest-frequency normal modes) were the most conserved within the large dataset of proteins.[191] These results support the notion that collective atomic fluctuations are of functional importance and must be under evolutionary selection pressure. Liu and Bahar have used Elastic Network Models to study a set of 34 enzymes representing diverse protein families and functional classes and have found that conserved residues tend to have little mobility in the global modes, while sequence

variability was found to correlate with increased mobility.[192]

However, similarly to the cases of discontinuous mapping found between sequence, structure and function space (discussed in the previous subsection), one can easily imagine such discontinuities in the sequence→dynamics and structure→dynamics interfaces as well. In other words, it may be possible that small changes in protein sequence or structure corresponds to large changes of protein dynamics. Although very little is known about the mapping from sequence/structure to dynamics, the following few examples suggest that such 'transition points' at which dynamics changes abruptly upon minor modification of sequence or structure do exist.

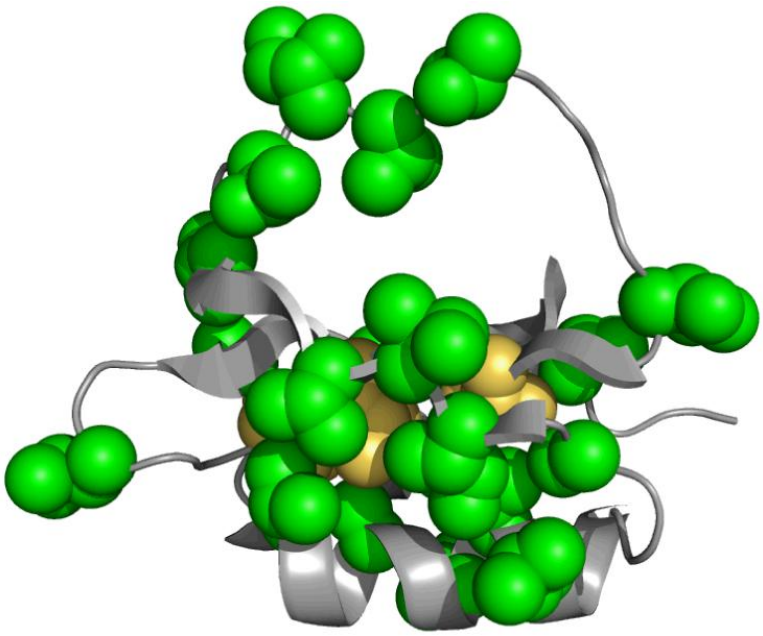

Figure 1.15: Single point mutations may have large and global effects on dynamics. In chymotrypsin inhibitor 2 (CI2), each of the five studied point mutations (at sites shown in yellow) were found to significantly change side-chain fluctuations in all regions of the protein according to ^2H NMR data.[193] Residues of consistently altered side-chain motions (i.e. in at least 4 of the 5 single mutants) are distributed throughout the structure (highlighted in green). (Image courtesy: Whitley et al. 2008[193])

For instance, Whitley et al. have used NMR relaxation experiments to study the effects of five hydrophobic core mutations on the backbone and side-chain dynamics of chymotrypsin inhibitor 2 (CI2).[193] They have found that each of the tested five point mutation

have significantly altered the picosecond-nanosecond side-chain motions as well as backbone motions, increasing the flexibility throughout the whole protein (Figure 1.15). The relative rigidity of the wild-type structure was suggested to be evolutionarily optimized and have an important role in the serine protease inhibitor function of the protein. However, minimal sequence changes can switch between entirely different global dynamic behaviours.

Similarly, the effects of single point mutations on functionally relevant molecular motions have been investigated in case of many other proteins including lysozyme[194], spermine synthase[17] and the SH3 domain of Fyn tyrosine kinase[195].

For example, Mittermaier and Kay have found based on 15N and 2H NMR relaxation data that backbone and side-chain dynamics of Fyn tyrosine kinase SH3 domain have changed considerably (i.e. the protein has become more flexible) as a result of single point mutations.[195] Verma et al. have also shown that certain point mutations in lysozyme caused small local structural perturbations accompanied by large changes of backbone flexibility.[194] Likewise, Zhang et al. have studied the effect of minor sequence differences on the conformational dynamics of spermine synthase (SMS). They have used molecular dynamics simulations to elucidate the effects of three clinically identified missense mutations known to be crucial in the development of Snyder-Robinson syndrome. Their results showed that some substitutions altered the conformational dynamics of the enzyme largely affecting its function.[17]

In terms of discontinuities in the mapping from structure to dynamics, a number of examples are known of proteins that adopt highly similar 3D-structures yet have very different dynamic properties. For instance, in case of the dimeric catabolite activator protein (CAP), binding of a cAMP ligand to one of its subunits was found to have no effect on the structure of the second subunit, yet it caused prominent changes in its intrinsic dynamics.[60] The process of negatively cooperative binding of cAMP to dimeric CAP is one of the first discovered examples in which allosteric communication is exclusively mediated by altered residue fluctuations without apparent conformational change (see more about dynamically-driven allostery in Chapter 6). On the other hand, this example also

illustrates that very similar or even identical protein structures may have entirely different internal motions. Similarly, comparison of the apo and ligand-bound forms of the phosphotyrosine-binding domain of insulin receptor substrate-1 (IRS-1) has revealed only neglectable structural difference but a significant shift in dynamics (also discussed in Section 1.2.3).[59] Thus minor local structural perturbations (due to various reasons such as interactions with ligands, chemical modifications or point mutations) may induce large differences in protein motions.

To summarize, these examples show that neither sequence similarity, nor structural similarity necessarily mean that two proteins would have similar dynamics. The existence of transition points at which protein dynamics can change abruptly due to point mutations or slight structural perturbations is likely to be remarkably important in protein evolution. First of all, the "evolutionary path" between distinct protein functions may be shorter than expected as single or only few mutational events might switch between significantly different conformational dynamics and, consequently, function. Secondly, since functionally important protein motions may depend sensitively on certain amino acid positions or structural features, these critical residues and structural properties are probably evolutionarily highly conserved. The above-mentioned examples also suggest that Elastic Network Models that represent proteins without taking into account their primary sequence have only limited accuracy.

As discussed above, in order to study the conservation of protein dynamics independent from the conservation of sequence and structure, one would need to define a quantitative measure of dynamic similarity (which does not require sequence or structure information). Figure 1.16 presents a schematic representation of six possible relations between sequence, structure and dynamics similarity of two proteins. When both the sequences and structures of the two proteins are significantly similar (Figure 1.16A and D), either pairwise sequence alignments or structural alignments can be used for comparative analysis. In case the proteins have significantly conserved 3D-structures, but highly diverged sequences (Figure 1.16B and E), structural alignments are still able to match equivalent structural regions. Transition points between the sequence and structure space may re-

sult in protein pairs that have similar sequences and dissimilar structures (Figure 1.16C), as exemplified by G_A88 and G_B88 (discussed above). Although in most cases similar sequences and structures or just similar structures are good predictor of dynamic similarity (Figure 1.16A and B), this is complicated by the existence of transition points between sequence/structure space and dynamics space (Figure 1.16D and E). Finally, Figure 1.16F represents an imaginary option when two proteins with no detectable sequence and structural similarity have similar functionally relevant motions. In this case, dynamic similarity could help to predict similar protein functions even in the absence of sequence and structure conservation. However, this option relies on the assumption that one can formulate a definition of dynamic similarity that is independent from structural correspondence.

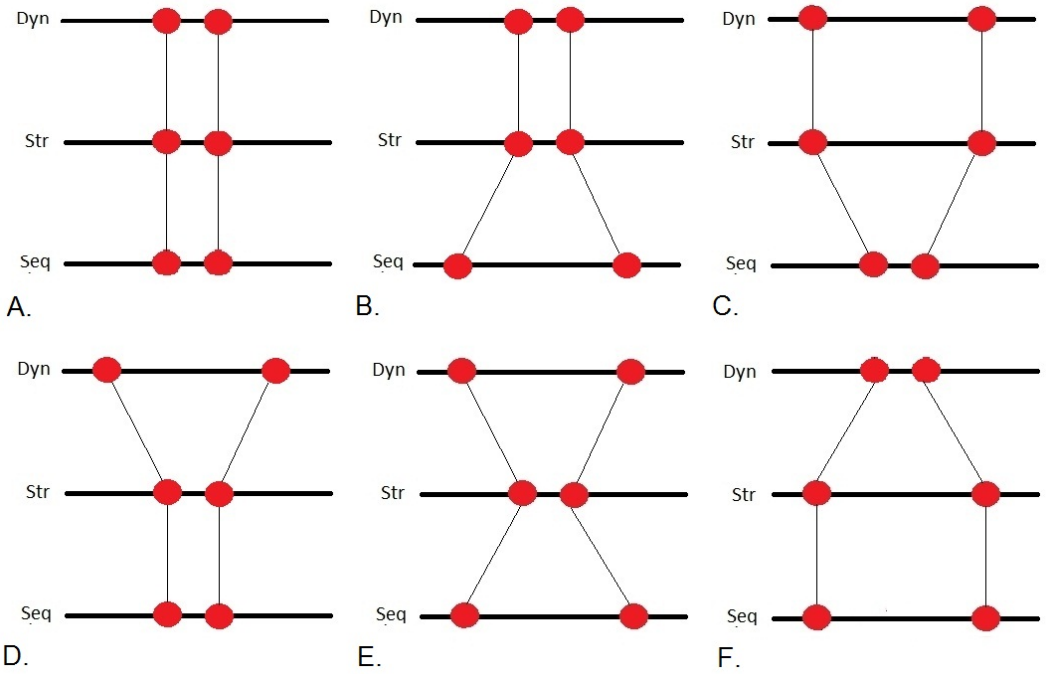

Figure 1.16: Schematic representation of six possible relations between sequence, structure and dynamics similarity of two proteins. The three levels, "Seq", "Str" and "Dyn" represent the sequence, structure and dynamics space, respectively. Similarity and dissimilarity are illustrated as small and large distances between the two points. Due to the different divergence rates in the three spaces, the existence of transition points and convergent evolution, sequence similarity does not guarantee structural similarity, structural similarity does not guarantee sequence similarity and neither sequence similarity, nor structural similarity guarantee dynamic similarity.

1.5 PDZ domains: connecting proteins

Throughout the research work covered in this thesis, the family of PDZ domains has been investigated and used as a test case to assess the novel methodology. There were several important reasons why PDZ domains have been selected as suitable systems for this research. First of all, because of their small sizes, long (>100 ns) molecular dynamics simulations were easily feasible. Secondly, substantial literature and experimentally-determined structures were available. Furthermore, PDZ domains represent proteins with highly conserved 3-dimensional structures but diverse functional properties (i.e. ligand binding specificities). As discussed in Section 1.4, such examples are interesting for studying the relationship between protein sequence, structure, dynamics and function. Finally, PDZ domains are of great clinical importance playing central role in various disease pathways.

1.5.1 What are PDZ domains?

PDZ (**P**ost-synaptic desity-95/**D**iscs large/**Z**onula occludens-1) domains are small (80-90 amino acid long) protein interaction modules commonly found in signalling proteins.[196] They play a key role in various signalling pathways by controlling the localization, targeting, clustering and anchoring of receptors, transporters and ion channels.[196] PDZ domains most often interact with the C-terminal peptides of target proteins mediating protein-protein interactions.[197] The biophysical aspects of folding and binding reactions of PDZ domains has been intensively studied[198], as well as their ligand preferences[199,200] and the background of binding specificity against a wide range of ligands[201].

Due to their biological importance, PDZ domains are found in a large number of species: bacteria, yeast, plants, invertebrates and vertebrates.[202] However, the PDZ family has undergone a evolutionary expansion and diversification: i.e. the number of PDZ-containing genes encoded in metazoan genomes is much larger than in prokaryotes, plants, and fungi. In addition, the architectures of PDZ-containing genes and the variety of associations between PDZ and other domains are much more diverse in metazoan.[203] Based on

large-scale genomic studies and *in silico* bioinformatics analysis, the number of potential PDZ domains is estimated to be ∼90 in the *Caenorhabditis elegans*, ∼130 in the *Drosophila melanogaster* and over 400 in the human genome.[204]

1.5.2 PDZ domain-containing proteins

Intracellular signal transduction from receptors at the plasma membrane to the cytoplasm and the nucleus are mediated by specific protein-protein interactions.[205] Signalling proteins that belong to the same signal transduction cascade are often found to be organized into large multiprotein complexes (i.e. signalling complexes) at the cell membrane.[144,206] There are several biophysical advantages of the formation of signalling complexes. The fact that different components of a pathway are organized in close proximity and chemical reactions can take place in these clusters instead of freely diffusible environments largely enhances signalling efficiency and can ensure its specificity. Moreover, oligomerization of signalling complexes may further increase signalling sensitivity.[144]

PDZ domains play essential role in the assembly of many signalling complexes by mediating protein-protein interactions. Their primary biochemical function is recognizing and binding to C-terminal peptides in a sequence-specific manner.[208] In addition, alternative modes of PDZ domain-mediated protein interactions have also been identified: e.g. some studies have shown that certain PDZ domains can bind to internal peptide motifs or interact with other PDZ domains.[209,210] Holding the different signalling components together via protein-protein interactions, PDZ domains act as "glue" in the formation of multiprotein complexes at the cell membrane.[196] An example in Figure 1.17 shows how PDZ domain-containing proteins participate in the organization and regulation of neurotransmitter receptors in the postsynaptic density. A similar example of the role of the PDZ-containing protein InaD in Drosophila phototransduction signalling is presented in Chapter 5.

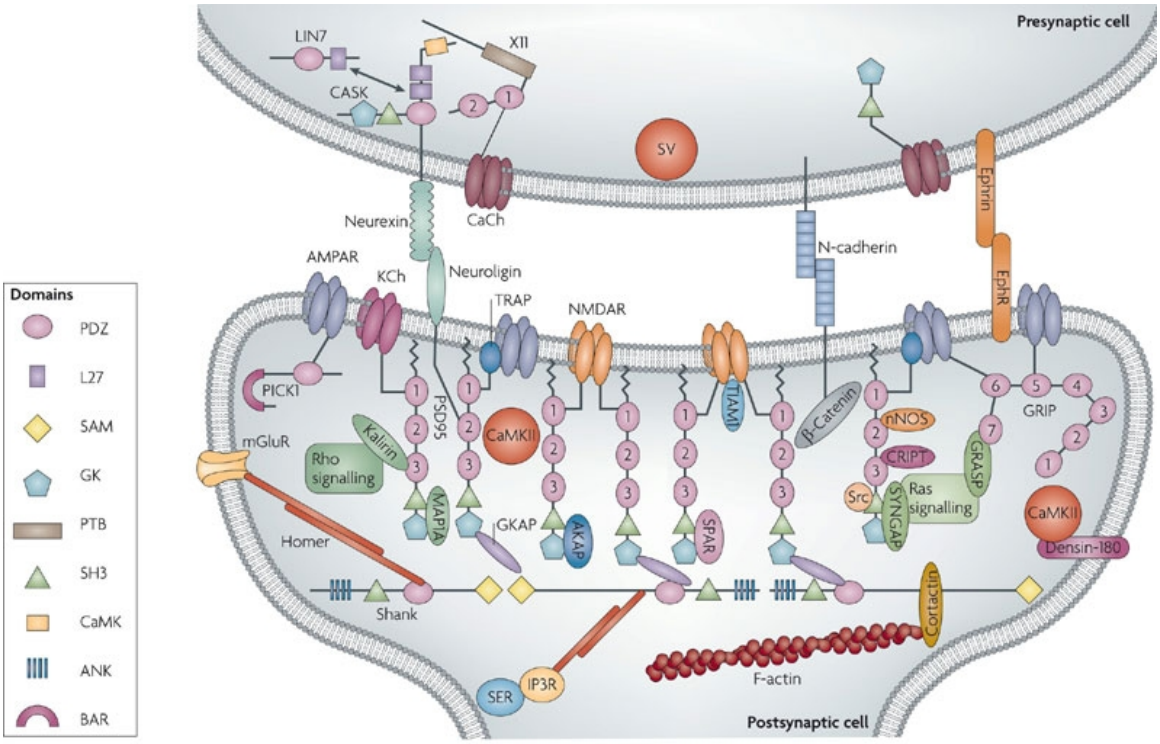

Figure 1.17: Multiprotein complexes held together by PDZ domain-containing proteins in the postsynaptic density (PSD) (Image courtesy: Feng and Zhang 2009)[207]

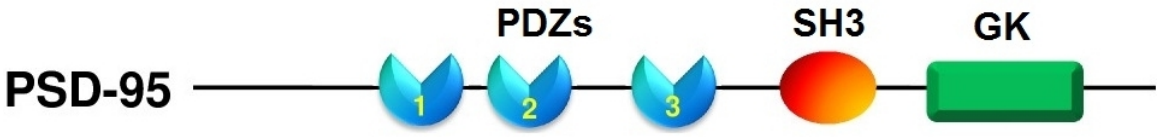

Figure 1.18: Domain architecture of the human Postsynaptic Density Protein 95 (PSD-95). The protein contains three tandem PDZ domains, an SH3 domain and a guanylate kinase (GK) domain. (Image courtesy: Lee and Zheng 2010[211])

More than 250 PDZ domain-containing proteins have been identified so far in the human proteome.[212] One of the major roles of PDZ domain-containing proteins is to serve as scaffolds for the assembly of large multiprotein complexes at the cell surface or other specific subcellular locations.[208] Many PDZ domain-containing proteins carry out only this scaffolding function as they are exclusively made up of an array of PDZ domains. Exam-

ples of such multi-PDZ proteins include InaD (Inactivation-no-after-potential D protein) and GRIP (glutamate receptor interacting protein) that are composed of five and seven tandem PDZ domains, respectively. Other PDZ-domain containing proteins, such as PSD-95 (Postsynaptic Density Protein 95) and HtrA (High-temperature requirement A protein), contain additional interaction or catalytic domains and have more complex functions than simply acting as a molecular glue between signalling proteins. As an example, Figure 1.18 shows the domain architecture of the human Postsynaptic Density Protein 95 (PSD-95) that contains three consecutive PDZ domains, an SH3 domain and a guanylate kinase (GK) domain. PSD-95 plays key role in the organization of signalling complexes in the postsynaptic density (Figure 1.17).[213] As additional examples, the domain architectures of five other PDZ domain-containing proteins are shown in Chapter 5.

With each PDZ domain having unique peptide binding specificity properties, the different combinations of PDZ domains in scaffold proteins (e.g. GRIP) determine the composition of multiprotein complexes formed around these molecular scaffolds.[208] In addition, the ability of many PDZ domain-containing proteins to multimerize (mostly via PDZ-PDZ interactions) can further increase the size and potential heterogeneity of the protein complexes.[208,214] For example, GRIP and ABP (AMPA receptor binding protein) can form both homomultimers and heteromultimers interacting via their PDZ4-6 domains.[215,216] Another example is InaD which can self-associate via its PDZ3 and PDZ4 domains to form multimers.[217]

1.5.3 The canonical PDZ structure and peptide binding

PDZ domains are classified as a separate structural family in the SCOP database.[125] Experimental structures of a large number of PDZ domains are available in the RCSB Protein Data Bank (PDB). The canonical PDZ domain structure consists of six β-strands (β1-β6) and two α-helices (α1 and α2).[198,211] While the lengths of secondary structural elements vary in different PDZ domains, the tertiary structure is highly conserved.[211] The primary sequences fold into a globular "six-stranded β-sandwich" structure.[196] The canonical peptide-binding site is formed between the β2-strand and the α2-helix and is capped by the β1/β2 loop

referred to as the carboxylate-binding loop that contains a conserved Gly-Leu-Gly-Phe (GLGF) motif.[208]

As a representative example of the canonical PDZ fold, Figure 1.19 shows the crystallographic structure of PSD-95 PDZ3 in complex with the C-terminal peptide of CRIPT (PDB: 1be9). Like in case of most PDZ-peptide interactions, the ligand binds to the extended peptide binding groove located between β2-strand and α2-helix.

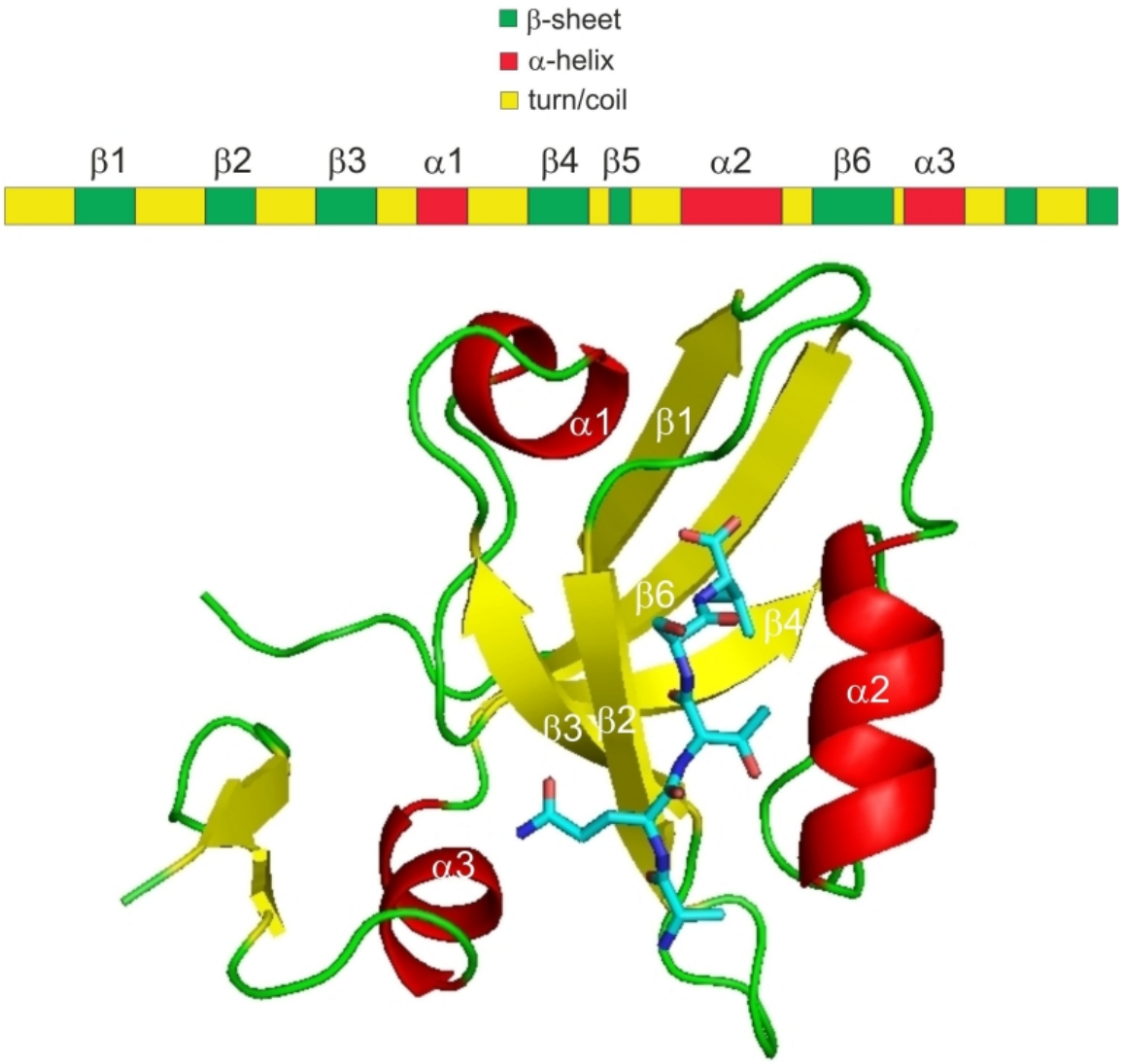

Figure 1.19: Secondary structural elements and tertiary structure of the third PDZ domain (PDZ3) of PSD-95 in complex with the C-terminal pentapeptide of CRIPT (KQTSV) (PDB: 1be9). The secondary structural annotations are predicted by the STRIDE web server.

When interacting with a C-terminal peptide, three main chain amide protons of the GLGF motif in the carboxylate-binding loop form hydrogen bonds with the free carboxylate group at the C-terminal end of the peptide ligand.[218] Furthermore, a series interactions between the main chain of β2-strand and the main chain of the peptide stabilize the ligand in the binding pocket. As a result, the peptide is inserted as an anti-parallel extension of the β-sheet on the PDZ domain surface.[197]

In addition to interacting with the main chain of β2-strand and the carboxylate-binding loop, the ligand also engages in specific side-chain interactions with the α2-helix at the opposite side of the binding site.[219] These side-chain interactions are crucial for determining the ligand specificity of the PDZ domain enabling the recognition of specific peptide motifs. (For more detailed discussion of what determines ligand binding specificity of PDZ domains, see the next subsection and Section 5.2.2.)

Although the most common interaction mode of PDZ domains is binding to C-terminal peptides, some PDZ domains have been found to also interact with internal peptide motifs.[197] Interaction with an internal peptide region usually involves a very similar binding mode as with a C-terminal peptide, since the internal peptide motif is often structured as a "pseudo-peptide" to fit into the binding pocket. For instance, the second PDZ domain of PSD-95 (PDZ2) recognize the PDZ domain of nNOS (neuronal nitric oxide synthase) in a way that is not dependent on the C-terminal sequence of nNOS PDZ, however, requires a 30-residue extension on the nNOS PDZ domain. Studying the structural basis of this binding mode has revealed that the 30-residue extension on nNOS PDZ folds into an extended β-hairpin conformation called the "β-finger" which can dock into the PSD-95 PDZ2 binding pocket mimicking a C-terminal peptide ligand.[220]

Besides the six β-strands and two α-helices found in the canonical PDZ fold, some PDZ domains such as PDZ3 of PSD-95 have additional secondary structural elements, for instance, as in the case of PSD-95 PDZ3, a third α-helix (α3). The roles of such extra elements are not always clear but they may be important for the function of the PDZ domain. For example, removal of α3-helix of PSD-95 PDZ3 reduces its ligand binding affinity by 21-fold despite the fact that α3 is located remote from the peptide binding site.[221] In some

PDZ domains (e.g. PTP-BL/BAS PDZ2), distal sites opposite the peptide binding groove have been found to be allosterically linked to the binding site via intramolecular communication pathways.[222,223] (The allosteric communication pathways of the mouse PTP-BL PDZ2 domain are studied in Chapter 6.)

1.5.4 PDZ domain specificity and promiscuity

In general, PDZ domains are involved in four different types of interactions: recognition of C-terminal peptides[211], recognition of internal peptides[211], PDZ-PDZ dimerization[224–226] and recognition of lipids[218,227]. Since PDZ domains function as "glue" in large signalling complexes, their ligand binding specificities are crucial in determining the compositions of these multiprotein complexes.

Those PDZ domains that interact with C-terminal peptides (also referred to as the classical or canonical binding mode) has been further divided by early studies into three classes (Class I-III) based on their preference against carboxyl-terminal ligand positions 0 and -2 (i.e. the very C-terminal position, p_0, and the third C-terminal position, p_{-2} of the peptide). (10). According to this original classification system, class I PDZ domains recognize peptides that have Serine or Threonine at position p_{-2} and a hydrophobic amino acid at position p_0 (a motif of **Ser/Thr-X-Φ-COOH**, where X is any amino acid and Φ is any hydrophobic amino acid). Class II PDZ domains bind to peptides that have any hydrophobic amino acids at both positions p_0 and p_{-2} (a motif of **Φ-X-Φ-COOH**). Finally, Class III PDZ domains interact with peptides that have Aspartic acid or Glutamic acid at position p_{-2} and any hydrophobic amino acids at position p_0 (a motif of **Asp/Glu-X-Φ-COOH**). As discussed in the previous subsection, the specific recognition of these peptide motifs are mediated by side-chain interactions between binding site and peptide residues.

However subsequent studies suggested that the picture is far more complicated, because PDZ domains have overlapping specificity and promiscuity toward their target peptides[198] and the PDZ binding cleft is able to interact specifically with up to seven C-terminal peptide residues[228]. A recent large scale analysis of the PDZ domain family in the human and Caenorhabditis elegans proteome found 16 distinct specificity classes sug-

gesting that the specificity map of PDZ domains is indeed surprisingly diverse and complex[200]. Using protein microarrays and quantitative fluorescence polarization to study the interactions of 157 mouse PDZ domains with 217 genome-encoded peptides, Stiffler et al. have found that PDZ domains do not really fall into discrete specificity classes, instead they are evenly distributed in selectivity space.[229]

Even applying the simplest classification system discussed above which divides PDZ domains into three distinct specificity classes, it is intriguing that many PDZ domains are able to interact with multiple peptides belonging to different classes of peptide motifs. For example, the PDZ domain of Erbin can bind both to class I peptides (e.g. the C-terminus of δ-catenin) and to class II peptides (e.g. the C-terminus of ERBB2).[230] Other PDZ domains, such as the PDZ domain of Dishevelled-2 are able to bind to both C-terminal and internal peptide ligands.[231] The promiscuity of PDZ domains, a property that is likely to be relevant for their central role in organizing signalling pathways and also essential for their evolvability, is discussed in details in Chapter 5.

Finally, it is important to note that the peptide binding preference of some PDZ domains may also be modulated by other structural elements in the protein. For example, Van Den Berk et al. have found that the binding specificity of the second PDZ domain of mouse PTP-BL (PDZ2) is regulated by the first PDZ domain of the protein (PDZ1). In isolation, PDZ2 recognizes class III peptides, however, when linked to PDZ1, it is unable to bind to class III ligands (see Section 6.2.2 for details).[232] Alternatively, PDZ-ligand interactions can be modulated via different regulatory mechanisms such as phosphorylation. For instance, Zhang et al. have recently found that phosphorylation at the atypical helical extension (α3) of PSD-95 PDZ3 that occurs *in vivo* allosterically modulates ligand binding affinity of this domain.[233]

1.5.5 Clinical importance of PDZ domains

Since PDZ domains play central roles in many signalling pathways serving as hubs in protein-protein interaction networks, their malfunctions can have serious consequences and lead to various diseases. Several heritable human diseases including Usher syndrome

[234], Dejerine-Sottas neuropathy[235] and Cystic Fibrosis[236] have been linked directly to defects of PDZ domain-containing proteins or their interaction partners. In addition, PDZ domains are crucially involved in the organization of signalling pathways implicated in many human diseases including cancer, Alzheimer's disease and schizophrenia (see Chapter 5). Because of their central role in various disease pathways and their well-defined binding sites, PDZ domains are promising targets for rational drug design.[237]

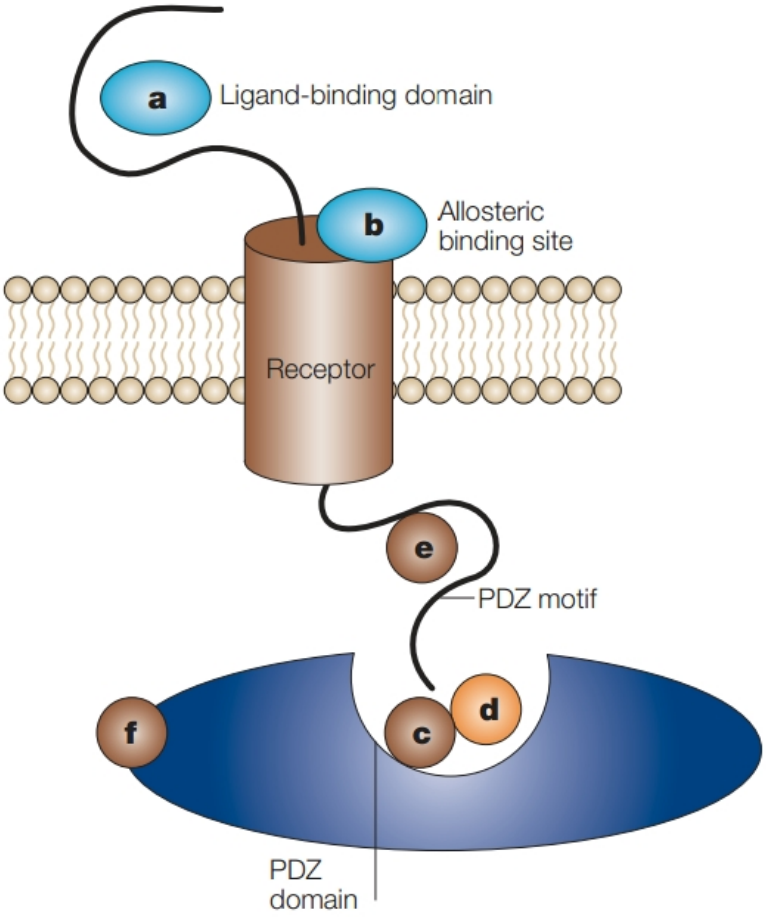

Figure 1.20: Alternative strategies of modulating signalling pathways. In the traditional approaches, receptor activity is modulated by agonist or antagonist molecules that bind to the extracellular ligand binding site (**a**) or allosteric sites of the receptor (**b**). A novel strategy could be using peptides (**c**) or small molecules (**d**) that dock into the peptide binding sites of the PDZ domains of co-activator proteins. This would result in competitive inhibition of the interaction between the co-activator and the receptor. Alternatively, small molecules could be designed to bind to the carboxy-terminal domain of the receptor, masking the peptide motif recognized by the PDZ domain (**e**). Finally, small molecules may interact with allosteric sites on the PDZ domains modulating the peptide binding affinity of the domain (**f**). (Image courtesy: Dev 2004)[237]

Interestingly, many viruses were found to encode proteins that bind to cellular PDZ proteins. These include both oncogenic and nononcogenic viruses (e.g. hepatitis B virus, rhesus papillomavirus, cottontail rabbit papillomavirus, influenza and human immunodeficiency virus).[238] Understanding how the virus modulates the cellular protein-protein interaction networks by interacting with PDZ domain-containing proteins of the host cell could also lead to novel targets for antiviral therapy.[239]

Significant effort has been made in recent years to explore the possibilities of targeting PDZ domain-containing proteins. A straightforward option for modulating protein-protein interactions is using inhibitor peptides that mimic specific motifs recognized by the PDZ domains (Figure 1.20c).[237] This results in competitive binding between the synthetic peptide and the natural ligands of the PDZ domains *in vivo*. Several successful attempts of inhibiting protein-protein interactions using small blocking peptides have been reported. A promising example was the disruption of interactions between PICK1 and glutamate receptors using synthetic peptides[240,241]. Since the the PDZ-domain containing protein PICK1 interacts with many disease-associated proteins (e.g. proteins involved in various cancers such as breast, lung and kidney cancer or psychiatric, neurological and neurodegenerative disorders such as depression, hyperactivity, schizophrenia and Parkinson's disease), modulating the interactions of PICK1 with these proteins has been suggested to open new therapeutic possibilities in various diseases.[237]

Although blocking peptides are indeed very effective in disrupting PDZ-peptide interactions, there are several difficulties of using synthetic peptides as drug compounds. For example, peptides have limited membrane permeability and their peptidic nature makes them subjects of undesired degradation and pharmacokinetics.[237]

An alternative approach is the use of small-molecule nonpeptide inhibitors designed to bind either to the PDZ peptide binding cleft, the carboxy-terminal motif on the binding partner or an allosteric site on the PDZ domain (Figure 1.20d,e,f).[237] For example, screening of ∼44,000 compounds have recently identified a small-molecule nonpeptide inhibitor that binds to the PICK1 PDZ binding site with an affinity similar to that observed for endogenous peptides. The inhibitor was found to effectively modulate the PICK1/AMPA

receptor interaction and is therefore a promising compound for the therapy of neuropathic pain, excitotoxicity and cocaine addiction.[242]

As a second example, much research have been conducted to find small-molecule inhibitors against the PDZ domain of Dvl (Dishevelled) because of its central role in the Wnt signalling pathway.[243–247] Using such inhibitors to block the association between the Dvl PDZ domain and Frizzled receptors has proven to be very effective for suppressing upregulated Wnt signalling in tumour cells.[244] More PDZ domains of clinical/pharmacological importance are discussed in Chapter 5.

While some of the drug candidates against PDZ domain targets have been identified using high-throughput biochemical compound screening (HTS) methods[242,244], virtual (*in silico*) screening has also proven to be very useful in the search for small-molecule inhibitors. Both ligand-based (i.e. pharmacophore-based)[245,247] and structure-based[243,246] virtual screening approaches have been used. Since the ligand preferences of PDZ domains are likely to be determined not only by their sequences and structures but also their conformational dynamics, structure-based (docking) studies should incorporate information about the flexibility of the PDZ targets.[248,249] Better understanding of how the intrinsic dynamics of PDZ domains are related to ligand binding could make structure-based screening approaches more reliable.

Chapter 2

Methods

2.1 Molecular Dynamics (MD) simulations

As discussed in Section 1.2.4, Molecular Dynamics (MD) simulations serve as a powerful tool for studying the conformational dynamics of proteins. The key elements of atomistic MD simulations performed in this work are briefly summarized in this section. All simulations discussed in the thesis have been prepared and run with the molecular dynamics software suite GROMACS[250] .

2.1.1 All-atom MD simulations (AT-MD)

All-atom MD (AT-MD) simulations[251] represent every atoms of the biomolecular system individually, calculating their coordinates and velocities in the course of time. The system may contain atoms of protein molecules, atoms of the solvent (in case of explicit models discussed below), solvated ions, atoms of other molecules (e.g. lipids, ligands, cofactors or drug compounds) etc. All these move in a dynamically changing potential (also referred to as the "force field") and can be accurately described with Newton's laws of motion.

Empirical force fields

The force field of the system[252] can be decomposed into multiple terms describing the potential energy contributions of different types of interactions. Generally, the total potential energy is the sum of bonded and non-bonded energies. The contributions of bonded interactions (corresponding to covalently bound atoms) include the effects of bond-stretching (V_{bonds}), angle-bending (V_{angles}), improper and proper dihedrals (V_{improp} and V_{prop}). Fur-

thermore, the non-bonded potential energy terms describe long-range electrostatic (V_{electr}) and van der Waals (V_{vdW}) interactions. The total potential energy of the system can therefore be given as the following sum:

$$V_{total} = V_{bonds} + V_{angles} + V_{improp} + V_{prop} + V_{electr} + V_{vdW} \tag{2.1}$$

For example, in the simplest case, V_{bonds}, V_{angles} and V_{improp} can be modelled as harmonic potentials. Additionally, to represent the potential energy of proper torsional deviations (V_{prop}), the Ryckaert-Bellemans function[253] is often used. Furthermore, electrostatic interactions (V_{electr}) can be described by Coulomb's law, while van der Waals interactions (V_{vdW}) are modelled with the Lennard-Jones potential[252].

Therefore a possible form of the total potential energy is given as

$$\begin{aligned} V_{total} = & \sum_{bonds} \frac{k_b}{2}(b - b_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \sum_{impropers} \frac{k_\xi}{2}(\xi - \xi_0)^2 + \\ & + \sum_{propers} \sum_{n=0}^{5} C_n(cos(\phi - 180))^n + \sum_{pairs(i,j)} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} + \\ & + \sum_{pairs(i,j)} 4\epsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) \end{aligned} \tag{2.2}$$

where the key parameters are: the equilibrium bond lengths (b_0), the equilibrium bond angles (θ_0), the equilibrium torsion values (ξ_0), the force constants of the harmonic potentials (k_b, k_θ and k_ξ), the coefficients of the Ryckaert-Bellmans function (C_n), the vacuum permittivity and relative permittivity of the solvent (ϵ_0 and ϵ_r) and the collision diameter and well depth parameters of the Lennard-Jones potential (σ_{ij} and ϵ_{ij}).

A number of different force field functions have been developed for MD simulations including the OPLS-AA[254], GROMOS[255], AMBER[256] and CHARMM[257] force fields that differ in their exact forms and their parameters optimized by fitting the potential function against experimental data. The commonly used OPLS-AA (Optimized Potentials for Liquid Simulations - all atom) force field[254] has been applied in the MD simulations per-

formed in this work.

Numerical integration of the equations of motion

In classical MD simulations, the motion of each atom can be described by Newton's second law: e.g. the second time-derivative of the position of atom i ($\mathbf{r_i}$) is directly proportional to the force acting upon this atom ($\mathbf{F_i}$):

$$\mathbf{F_i} = m_i \frac{d^2\mathbf{r_i}}{dt^2} \tag{2.3}$$

where the force vector $\mathbf{F_i}$ is the gradient of the potential field given in Eq. 2.2:

$$\mathbf{F_i} = -\nabla V(\mathbf{r_i}) \tag{2.4}$$

In such many-body systems, the analytical solution of the above described differential equations is not feasible. To overcome this problem, numerical methods such as the Verlet algorithm and leap-frog algorithm are used to integrate Newton's equations.[258,259] These numerical integration techniques aim to calculate the coordinates and velocities step by step: that is, given the current state of the system, the coordinate and velocity vectors are re-calculated for the next time step.

In the leap-frog integrator[259] used in this work, the velocities and coordinates are approximated recursively. As a first step, the velocity vector is determined at time $t + \frac{1}{2}\delta t$ (where δt is the integration time step of the MD simulation):

$$\mathbf{v}(t + \frac{1}{2}\delta t) = \mathbf{v}(t - \frac{1}{2}\delta t) + \mathbf{a}(t)\delta t \tag{2.5}$$

Here $\mathbf{a}(t)$ denotes the instantaneous acceleration vector derived from Eq. 2.3.

As a next step, the coordinate vector at time $t + \delta t$ is calculated as

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t + \frac{1}{2}\delta t)\delta t \tag{2.6}$$

Finally, the velocity vector at time t is approximated according to

$$\mathbf{v}(t) = \frac{1}{2}\left[\mathbf{v}(t + \frac{1}{2}\delta t) + \mathbf{v}(t - \frac{1}{2}\delta t)\right] \tag{2.7}$$

The above three steps are iterated in order to generate trajectories in the 3N-dimensional conformational space discussed below (where N is the number of atoms in the system).

When using bond length constraint algorithms such as RATTLE[260], SHAKE[261] or LINCS [262] to constrain bond stretching vibrations, an increased δt integration time step can be applied in the MD simulation. The LINCS (Linear Constraint Solver) method was used in this work that allowed for an integration time step of δt = 2 fs = $2 \cdot 10^{-15}$ s.

Implicit and explicit solvent models

Incorporating the effect of solvent is a crucial problem for performing realistic simulations of proteins. There are two main strategies used to represent aqueous solutions in MD simulations: applying explicit or implicit water models.

In explicit models, all water molecules added to the system are represented and simulated explicitly thereby largely increasing the number of degrees of freedom of the system. Although this strategy is clearly the most accurate way to incorporate the effect of solvent, calculating the coordinates and velocities of a large number of water molecules is time-consuming. Several different representations of water molecules have been developed such as the SPC, SPC/E, TIP3P, BF, TIP4P and ST2 models.[263–267] In this work, the SPC (Simple Point Charge) explicit water model[263] was used: that is, water was represented as a rigid 3-atom molecule assuming an ideal tetrahedral shape.

The alternative strategy is applying implicit solvation models by adding an extra energy term (i.e. solvation energy term) to the potential function of the system:

$$V(\mathbf{r}) = V_{vac}(\mathbf{r}) + \Delta G_{solv}(\mathbf{r}) \tag{2.8}$$

where $V_{vac}(\mathbf{r})$ denotes the potential energy function in vacuum and $\Delta G_{solv}(\mathbf{r})$ is the added solvation energy term. Several implicit models are available such as the GB (Generalized Born) model, ACE (Analytic Continuum Electrostatics) model and ASP (Atomic

Surface Area-based Empirical) model.[268–270] Using implicit water models results in significantly faster but less accurate simulations.

Long-range electrostatic interactions

As shown by Eq. 2.2, the contribution of electrostatic interactions (V_{electr}) to the total potential energy of the system is calculated as a sum of all pairwise Coulomb interaction energies between each pair of atoms. If the system contains a large number of atoms (especially when explicit solvent is used), the calculation of this summation is computationally very expensive.

To address this problem, one option is to define a distance cutoff value beyond which electrostatic interactions are not taken into account. However, this would introduce non-physical effects causing problems in the simulations.[271–273] A more advanced solution is the particle-mesh Ewald (PME) summation method[274] used in this work that reduces computational time but still providing an accurate approximation of the electrostatic contributions to the potential energy.

In the PME method, the electrostatic potential energy is decomposed into two parts: a first term describing short-range interactions and a second term accounting for long-range interactions. Within a distance cutoff, the short-ranged potential is calculated in real space as a summation of the pairwise Coulomb interaction energies. Beyond the cutoff, the long-ranged potential is approximated as a summation in Fourier space using fast Fourier transform (FFT) on a discrete grid after interpolating electric charges to the grid points.

Distribution of initial velocities

The initial velocity vector of each atom is generated randomly based on a Gaussian distribution called the Maxwell-Boltzmann distribution. That is, for example, the probability of the x-coordinate of the initial velocity of atom i to be assigned to value v_{ix} is given as

$$p(v_{ix}) = \sqrt{\frac{m_i}{2\pi k_B T}} \cdot exp\left[-\frac{1}{2}\frac{m_i v_{ix}^2}{k_B T}\right] \tag{2.9}$$

where m_i is the mass of atom i, k_B is the Boltzmann constant and T is the specified temperature of the simulation. Drawing initial velocities from this Gaussian distribution ensures that the total momentum of the system is approximately zero.

Temperature and pressure coupling

To maintain constant temperature and pressure of the system, these parameters are controlled with the Berendsen algorithm[275]. In particular, in the Berendsen thermostat, the system is weakly coupled to an external heat bath of temperature T_0. At every simulation step, the instantaneous temperature is corrected depending on its deviation from T_0:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{2.10}$$

where τ is the time constant that characterises the rate of the exponential decay of temperature deviation.

In practice, correcting the temperature involves rescaling the velocities of each atom at every time step of the simulation. The rescaling factor λ is given as

$$\lambda = \left[1 + \frac{\delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)\right]^{1/2} \tag{2.11}$$

The temperature coupling constant τ_T is related to above used time constant τ according to the formula:.

$$\tau_T = \frac{N_{df} k_B}{2C_V}\tau \tag{2.12}$$

where C_V denotes the total heat capacity and N_{df} denotes the total number of degrees of freedom of the simulated system.

Similarly, the Berendsen algorithm is used to maintain constant pressure by weakly coupling the system to a barostat with a reference pressure of P_0. At every time step, the instantaneous pressure P is corrected depending on its deviation from P_0:

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau} \tag{2.13}$$

At every simulation step, coordinates and box vectors are rescaled using the scaling matrix μ which can be calculated as

$$\mu_{ij} = \delta_{ij} - \frac{\delta t}{3\tau_P}\beta_{ij}\left\{P_{0ij} - P_{0ij}(t)\right\} \tag{2.14}$$

where matrix β describes the isothermal compressibility and τ_P is referred to as the pressure coupling constant.

Periodic boundary conditions

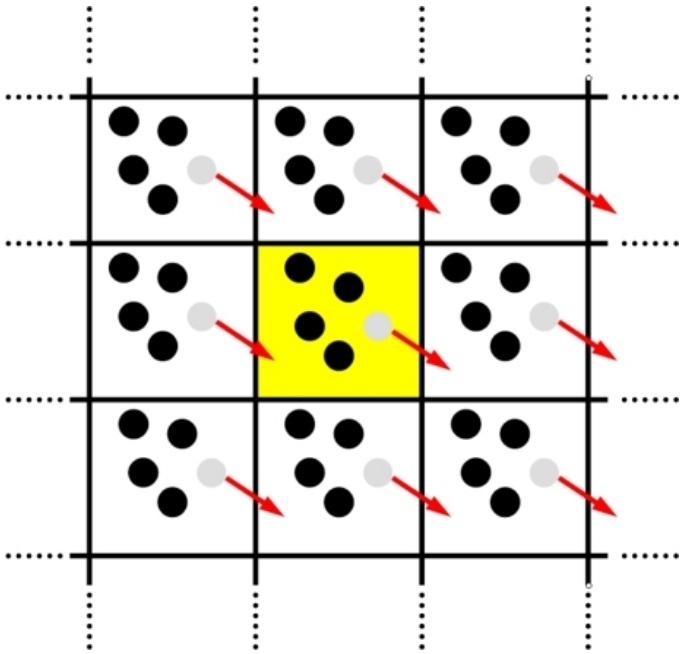

Figure 2.1: Schematic illustration of the use of periodic boundary conditions. Atoms leaving the box at one side, enter at the opposite side. (Image courtesy: Steinhauser and Hiermaier 2009[276])

Simulating a biomolecular system in a finite box can lead to severe errors (referred to as "boundary effects") arising from the fact that atoms at the edges of the box are not treated correctly. Therefore, to avoid these unwanted effects, periodic boundary conditions (PBC)

are often used in MD simulations mimicking that the protein is embedded in an "infinite" solvent environment.[276]

In practice, the application of PBC means that, for example, an atom which exits at one face of a cubic simulation box, automatically enters the box at the opposite face (Figure 2.1). This ensures that the motions of atoms are not limited by the "walls" of the box. PBC can be implemented using several different simulation box types such as cubic, rectangular, rhombic dodecahedron and truncated octahedron. These different box geometries differ in their efficiency of realizing the same periodic image distance. For convenience, in this work, MD simulations with PBC have been performed in simple rectangular boxes.

Energy minimisation, restrained and unrestrained simulations

Prior to MD simulation, the system usually undergoes an energy minimization (EM) procedure. The objective of EM is optimizing the geometry of the protein structure in order to identify an equilibrium conformation corresponding to a local minimum of the potential energy surface. The equilibrium structure is then used as the starting point of the MD simulation. There are a number of different algorithm used for energy minimization including the steepest descent, conjugate gradient and L-BFGS method.[277,278]

In the steepest descent algorithm applied in this work, the 3N-dimensional coordinate vector **r** of the protein structure is iteratively refined based on the following formula:

$$\mathbf{r}_{n+1} = \mathbf{r}_n + \frac{\mathbf{F}_n}{max|\mathbf{F}_n|} h_n \tag{2.15}$$

where $\mathbf{r}_n$ and $\mathbf{F}_n$ denote the coordinate vector and force vector at the n^{th} iteration step and $max|\mathbf{F}_n|$ is the maximum of absolute values of the $\mathbf{F_n}$ vector components. Parameter h_n denotes the maximum displacement which is initialized as $h_0 = 0.01nm$.

At each step, the force vector ($\mathbf{F_n}$) and the potential energy (V_n) are re-calculated based on the latest $\mathbf{r_n}$ coordinates. If the potential energy decreases ($V_{n+1} < V_n$), the proposed coordinate vector calculated by Eq. 2.15 is accepted and the maximal displacement is reset to $h_{n+1} = 1.2h_n$. If the coordinate refinement does not reduce the potential energy ($V_{n+1} \geq$

V_n), the proposed coordinates are rejected and $h_{n+1} = 0.2h_n$. The iteration terminates after a predefined number of steps or if $max|\mathbf{F}_n| < \epsilon$ where ϵ is a suitable cutoff value.

In order to let solvent molecules equilibrate around the protein, position restraint simulations are often applied in which atoms of the protein are restrained to some reference positions, while water atoms, ions and other molecules move in normal force field. For example, the potential energy used for protein atom i in a position restrained simulations can be given as

$$V_{pr}(\mathbf{r}_i) = \frac{1}{2}k_{pr}|\mathbf{r}_i - \mathbf{R}_i|^2 \tag{2.16}$$

where k_{pr} is the position restrained "force" constant and $\mathbf{R}_i$ denotes the reference position of atom i. Note that the potential energy of restrained protein atoms is described by a harmonic potential.

Following the energy minimization and a possible position restrained simulation, the main MD simulation (referred to as *production run*) is performed in which the whole system is governed by the normal force field defined in Eq. 2.2. In the production run, simulation frames ("snapshots") are saved for further analysis with a certain sampling frequency.

Output trajectories of the MD simulations

The output of an AT-MD simulation is the time series of K simulation frames saved for analysis that make up a conformational ensemble of the simulated protein. Let S_k denote the k^{th} conformation in this structural ensemble and the position vector of atom i in structure S_k is denoted as

$$\mathbf{r_i^{S_k}} = \left\{ r_{ix}^{S_k}, r_{iy}^{S_k}, r_{iz}^{S_k} \right\} \tag{2.17}$$

where vector $\mathbf{r_i^{S_k}}$ has three components: the x, y and z coordinates of the atom.

As each structure is characterized with 3N coordinates (where N is the number of atoms in the structure), the output $\{S_k\}_{k=1}^{K}$ ensemble corresponds to a trajectory in the 3N-dimensional configuration space. Since it describes the conformational dynamics of

the protein, this 3N-dimensional trajectory is the input dataset for further analysis.

2.1.2 Root mean square fluctuation (RMSF)

A commonly used approach to characterize residue fluctuations is calculating the root mean square fluctuation (RMSF) profile of alpha carbon atoms based on the MD trajectory. As a first step, each protein conformation in the ensemble is superposed to a common reference conformation using the least square superposition method. The RMSF profile characterizing the mobility of individual atoms can be calculated as

$$RMSF(i) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mathbf{r_i^{S_k}} - \mathbf{r_i^R})^2} \tag{2.18}$$

where $\mathbf{r_i^{S_k}}$ denotes the position of $C\alpha$-atom i in the superposed structure S_k and $\mathbf{r_i^R}$ denotes the position of $C\alpha$-atom i in a second reference structure R.

Note that two different reference structures may be involved in the calculation of RMSF profile: the reference used for superposition and the reference with regards which atomic displacements are measured (R). The two reference conformations may, however, be identical. Usually, the first structure of the ensemble is used as reference for superposition, and structure R is set to the mean of the superposed conformations.

2.2 Dimensionality reduction methods

Dimensionality reduction methods refer to a class of mathematical techniques that facilitate the analysis and comparison of high-dimensional datasets. The general objective of these methods is to reduce the dimensionality of datasets while preserving the original structure of data. To address this problem, several approaches have been proposed in the data mining literature including Principal Component Analysis (PCA), Multidimensional Scaling (MDS), Locally-Linear Embedding (LLE), Kernel Principal Component Analysis (kPCA) and Isomap algorithm.[34,279–282] Two of these methods, PCA and MDS, are summarized in this section.

2.2.1 Principal component analysis (PCA)

Principal component analysis (PCA)[34] is a widely-used data mining technique applied to analyse the hidden structure of complex multi-dimensional datasets. The goal of the method is to calculate an orthogonal linear operator that transforms the original data matrix to the coordinate system which best describes the variance of data. The axes of the new coordinate system are defined by the principal component vectors (PCs) of the data matrix. Principal components can be ordered according to the proportion of variance they explain: e.g. the first PC describes the largest variance of data, the second PC describes the second largest variance etc. PCA is a common tool in the analysis of MD trajectories as it has proven to be efficient in separating functionally important collective (correlated) atomic motions (also referred to as essential modes of motions) from uncorrelated thermal fluctuations. Given an input data matrix, calculating the principal component vectors involves the eigenvalue decomposition of the coordinate covariance matrix defined below.

Covariance and correlation matrix

Let vector $\mathbf{r_i^{S_k}}$ denote here the position of atom i in conformation S_k after each S_k conformation is superposed to a common reference structure. The NxN *atomic covariance matrix* capturing the cooperative motions of individual atoms is defined as

$$\Sigma_{ij} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{r_i^{S_k}} - \mathbf{r_i^0})(\mathbf{r_j^{S_k}} - \mathbf{r_j^0}) \tag{2.19}$$

where $\mathbf{r_i^0}$ denotes the mean position of atom i in the ensemble:

$$\mathbf{r_i^0} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{r_i^{S_k}} \tag{2.20}$$

Comparing of Eq. 2.18 and Eq. 2.19 and assuming that the reference position $\mathbf{r_i^R}$ is the mean position of atom i, we get that the root mean square fluctuation (RMSF) can be calculated as the diagonal elements of the Σ atomic covariance matrix:

$$RMSF(i) = \Sigma_{ii} \tag{2.21}$$

Furthermore the NxN *atomic correlation matrix* of displacements is given as the covariance matrix normalized by the product of standard deviations of the two atoms:

$$C_{ij} = \frac{\frac{1}{K}\sum_{k=1}^{K}(\mathbf{r_i^{S_k}} - \mathbf{r_i^0})(\mathbf{r_j^{S_k}} - \mathbf{r_j^0})}{\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mathbf{r_i^{S_k}} - \mathbf{r_i^0})^2}\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mathbf{r_j^{S_k}} - \mathbf{r_j^0})^2}} = \frac{\Sigma_{ij}}{\sigma_i \sigma_j} \tag{2.22}$$

The atomic correlation matrix (also called dynamical cross-correlation map) is commonly used for visualizing the pairwise dynamic couplings between atoms.

In addition to the above-defined atomic covariance and correlation matrices, the following 3Nx3N *coordinate covariance matrix* is used to characterize the concerted motions of individual atomic coordinates:

$$\Sigma_{ij}^{xyz} = \frac{1}{K}\sum_{k=1}^{K}(a_i^{S_k} - a_i^0)(a_j^{S_k} - a_j^0) \tag{2.23}$$

where $\mathbf{a^{S_k}}$ is the 3N-dimensional vector containing the 3N coordinates of structure S_k:

$$\mathbf{a^{S_k}} = \left\{r_{1x}^{S_k}, r_{1y}^{S_k}, r_{1z}^{S_k}, r_{2x}^{S_k}, ..., r_{Nz}^{S_k}\right\} \tag{2.24}$$

and vector $\mathbf{a^0}$ contains the mean coordinates calculated for the ensemble:

$$a_i^0 = \frac{1}{K}\sum_{k=1}^{K} a_i^{S_k} \tag{2.25}$$

Eigenvalue decomposition of the Σ^{xyz} matrix

The principal components (PCs) of the MD trajectory can be calculated by the following eigenvalue decomposition of the coordinate covariance matrix Σ^{xyz}:

$$\Sigma^{xyz} = V\Lambda V^T \tag{2.26}$$

where matrix V contains the $\mathbf{v_i}$ eigenvectors (i.e. principal components) as its columns

and the diagonal matrix Λ contains the corresponding λ_i eigenvalues. The PC vectors are usually ranked according to their ordered eigenvalues:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_{3N} \tag{2.27}$$

Note, that the eigenvalue corresponding to a PC determines the relative proportion of variance the eigenvector explains in the dataset:

$$R(k) = \frac{\lambda_k}{\sum_{i=1}^{3N} \lambda_i} \tag{2.28}$$

Essential subspace overlap

As discussed in Section 1.2.2, a number of studies have found that only a small set of PCs (referred to as the essential modes) can describe the functionally relevant large-scale motions of many proteins. When comparing the dynamics of different proteins or different simulations of the same protein, several studies have used the following root mean square inner product (RMSIP) measure[36] to quantify the overlap between the subspaces spanned by the essential PC vectors (essential subspaces) of two trajectories, A and B:

$$RMSIP(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{v_i^A} \cdot \mathbf{v_j^B})^2} \tag{2.29}$$

where $\mathbf{v_i^A}$ and $\mathbf{v_j^B}$ denote the PCs of simulation A and B, while the two sums run over the top n eigenvectors of the largest eigenvalues (a usual choice is $n = 10$). The RMSIP value of 1 means that the two subspaces are identical, while the RMSIP value of 0 indicates that the two subspaces are orthogonal.

Covariance matrix overlap

An alternative measure of similarity between the essential subspaces calculated for two simulations (A and B) is the *covariance matrix overlap*[283] defined as

$$\Omega(A,B) = 1 - \left[\frac{\sum_{i=1}^{3N}(\lambda_i^A + \lambda_i^B) - 2\sum_{i=1}^{3N}\sum_{j=1}^{3N}\sqrt{\lambda_i^A \lambda_j^B}(\mathbf{v_i^A} \cdot \mathbf{v_j^B})^2}{\sum_{i=1}^{3N}(\lambda_i^A + \lambda_i^B)} \right]^{1/2} \tag{2.30}$$

where $\mathbf{v_i^A}$ and $\mathbf{v_j^B}$ are the eigenvectors while λ_i^A and λ_j^B denote the eigenvalues of simulation A and B, respectively.

2.2.2 Multidimensional scaling (MDS) (Torgerson-Gower scaling)

Multidimensional scaling (MDS)[279] is a statistical technique that aims to transform a high-dimensional dataset to a lower-dimensional representation while preserving the pairwise dissimilarities (or distances) between individual data points. The method is used in Chapter 4 for mapping protein conformational ensembles to two-dimensional space so that they can be visualized and compared.

The input of the MDS algorithm is a KxK dissimilarity matrix D that contains pairwise structural dissimilarities of the K conformations. (Matrix D can be defined based on different structural similarity measures: see Section 2.5.2 and Section 5.3.1) The output is a KxL coordinate matrix denoted by Y, where L is the dimensionality of the lower dimensional space to which the conformations are mapped. For example, "projecting" the data points to a two-dimensional map for the purpose of visualization ($L = 2$), the output Y is a Kx2 matrix and (Y_{i1}, Y_{i2}) are the (x, y) coordinates of the point representing conformation i on the 2D map.

In classical MDS (or Torgerson-Gower scaling)[279], the goal is minimizing the following $\phi(\mathbf{Y})$ objective function:

$$\phi(\mathbf{Y}) = \sum_{ij} \left(D_{ij}^2 - \|\mathbf{y_i} - \mathbf{y_j}\|^2 \right) \tag{2.31}$$

where $\|\mathbf{y_i} - \mathbf{y_j}\|$ is the Euclidean distance between the two points representing conformations i and j in the low-dimensional space. The $\phi(\mathbf{Y})$ function measures how well the low-dimensional distances reproduce the original D_{ij} dissimilarities.

The Y matrix that minimizes the $\phi(\mathbf{Y})$ objective function can be calculated by the eigenvalue decomposition of the following G matrix (referred to as the Gram matrix):

$$G_{ij} = -\frac{1}{2}\left(D_{ij}^2 - \frac{1}{K}\sum_l D_{il}^2 - \frac{1}{K}\sum_l D_{jl}^2 + \frac{1}{K^2}\sum_{lm} D_{lm}^2\right) \tag{2.32}$$

Given the eigenvectors of matrix G, the entries of matrix Y are simply given as

$$Y_{i\alpha} = \sqrt{\lambda_\alpha} v_{\alpha i} \tag{2.33}$$

where $\{v_\alpha\}_{\alpha=1}^L$ are the top L eigenvectors of the Gram matrix G, and $\{\lambda_\alpha\}_{\alpha=1}^L$ are the corresponding eigenvalues.

2.3 Elastic Network Model (ENM)

Elastic Network Models (ENMs)[284] are simplified representations of biological macromolecules commonly used to study their large-scale dynamics. In these coarse-grained models, a protein is represented as a network of particles connected by elastic springs. In different types of elastic network models, the particles may represent individual atoms, groups of atoms (e.g. residues or secondary structural elements) or characteristic points defined for groups of atoms (e.g. center of mass of residues or side-chains). A major advantage of ENMs is their little computational cost as compared with for example MD simulations. Despite their simplicity, several studies have confirmed that protein fluctuations predicted by ENMs can reproduce MD or experimental data reasonably well. Accordingly, as discussed in Section 4.2, the ENM approach were used in comparative protein dynamics studies.

2.3.1 Gaussian Network Model (GNM)

In the Gaussian Network Model (GNM)[188] , residues are represented by their alpha carbon atoms. Linked particles of the network are connected by harmonic springs (Figure 2.2) which have a potential energy proportional to the square of their extensions. This

harmonic approximation of the potential greatly simplifies the study of protein dynamics.

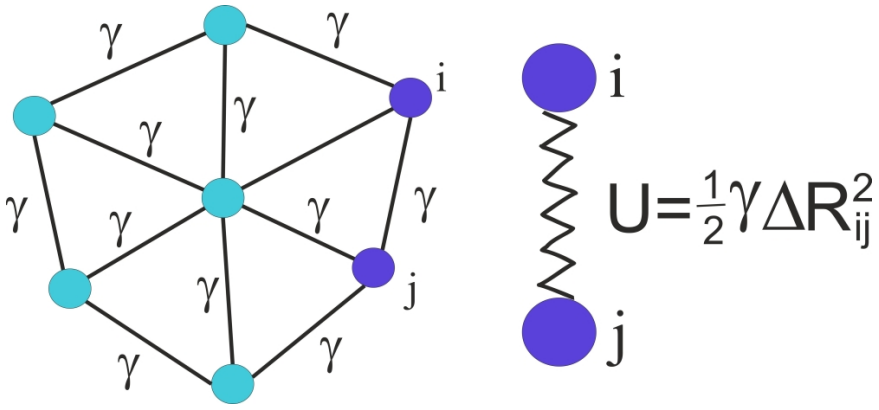

Figure 2.2: Schematic illustration of GNM. Each pair of particles (representing Cα-atoms) in the network is connected by harmonic springs that have the same γ force constant. Therefore the potential energy of an (i,j) particle pair is proportional to the square of their displacement ΔR_{ij}.

The total potential energy of the Gaussian network can be given as the sum of potential energies of all connected pairs of particles:

$$V = \frac{\gamma}{2}\left[\sum_{ij}^{N}(\mathbf{\Delta R_j} - \mathbf{\Delta R_i})^2\right] = \frac{\gamma}{2}\left[\sum_{ij}^{N}\mathbf{\Delta R_i}\Gamma_{ij}\mathbf{\Delta R_j}\right] \tag{2.34}$$

where $\mathbf{\Delta R_i}$ is the displacement of Cα-atom i, γ is the force constant and Γ denotes the Kirchhoff matrix defined as

$$\Gamma_{ij} == \begin{cases} -1 & \text{if } i \neq j \text{ and } R_{ij} \leq r_c \\ 0 & \text{if } i \neq j \text{ and } R_{ij} > r_c \\ -\sum_{j,j\neq i}^{N}\Gamma_{i,j} & \text{if } i = j \end{cases} \tag{2.35}$$

It is easy to show that the covariance matrix of atomic displacements in the GNM can be given as

$$< \mathbf{\Delta R_i} \cdot \mathbf{\Delta R_j} >= \frac{3k_BT}{\gamma}[\Gamma^{-1}]_{ij} \tag{2.36}$$

where k_B is the Boltzmann constant, T is the absolute temperature and Γ^{-1} denotes the inverse of the Kirchhoff matrix Γ.

As a consequence, the mean square fluctuations (MSF) of residues can be calculated as the main diagonal elements of the above described covariance matrix:

$$< \mathbf{\Delta R_i}^2 >= \frac{3k_BT}{\gamma}[\Gamma^{-1}]_{ii} \tag{2.37}$$

In addition, atomic fluctuations can be converted to theoretical B-factors (Debye-Waller factor) that are directly comparable with experimental B-factor profiles provided by X-ray crystallographic studies:

$$B_i = \frac{8\pi^2}{3} < \mathbf{\Delta R_i}^2 >= \frac{8\pi^2 k_BT}{\gamma}[\Gamma^{-1}]_{ii} \tag{2.38}$$

2.3.2 Normal mode analysis (GNM-NMA)

A commonly applied analysis on Gaussian Network Models is identifying the functionally most relevant modes of motions of a protein structure. The application of Normal Mode Analysis to Gaussian Network Models (GNM-NMA) has proven to be an effective approach for decomposing protein fluctuations into collective modes of atomic motions.[35] Deriving the normal mode vectors of the model involves the eigenvalue decomposition of the above-defined Kirchhoff matrix:

$$\Gamma = U\Lambda U^T \tag{2.39}$$

where the NxN unitary matrix U contains the u_i eigenvectors of the Kirchhoff matrix (normal mode vectors) and the diagonal matrix Λ contains the λ_i eigenvalues corresponding to these eigenvectors. However, because Γ is a positive semi-definite matrix, the first eigenvalue (λ_1) equals 0.

Given the eigenvalue decomposition of the Kirchhoff matrix, combining Eq. 2.36 and Eq. 2.39 the covariance matrix of atomic displacements can be written as:

$$< \mathbf{\Delta R_i} \cdot \mathbf{\Delta R_j} >= \frac{3k_BT}{\gamma}[\lambda_k^{-1} u_k u_k^T]_{ij} \tag{2.40}$$

2.4 Markov chain Monte Carlo (MCMC) methods

Markov chain Monte Carlo (MCMC) algorithms[285] refer to a set of methods mostly used in Bayesian statistical inference as a computationally feasible strategy to sample from a probability distribution of interest (called the *target distribution*).[285] The approach is based on creating a discrete time stochastic process called Markov chain (see more precise definition below) that converges to the desired target probability density. In addition, MCMC methods are also used in the field of stochastic global optimization, i.e. for finding the global minimum or maximum of an objective function, as applied in this work (in Chapter 4).

2.4.1 The Markov property ("memorylessness")

A Markov chain is defined as a sequence of random variables (indexed as X_1, X_2, X_3, ..., X_n) which satisfies the Markov property: i.e. the probability of the next state of the chain (X_{n+1}) depends only on the current state (X_n) and is therefore independent of the all previous states of the chain (Figure 2.3).[286] This property is also referred to as the memorylessness of the stochastic process:

$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n) \quad (2.41)$$

where $X_i \in \mathbf{S}$ values are the states of the Markov chain that belong to 'state space' **S**.

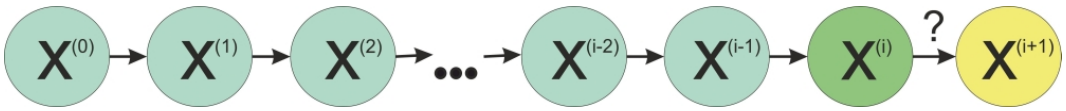

Figure 2.3: Markov property of a sequence of random variables: the chain is memoryless, i.e. the transition probability from X_n to X_{n+1} is dependent only on state X_n.

The way $P(X_{n+1} = x_{n+1} | X_n = x_n)$ probability is calculated differs in different MCMC methods. The Metropolis-Hastings algorithm[287,288] detailed below was applied in this work to perform global optimization.

2.4.2 Metropolis-Hastings algorithm

In the Metropolis-Hastings algorithm[287,288], a potential next state of the Markov chain (X') is generated first by sampling from a proposal probability density function $q(y|x)$:

$$X' \sim q(y|x^{(i)}) \tag{2.42}$$

Next, the proposed state X' is accepted or rejected as the next state of the chain according to an acceptance probability function that depends on the current state of the chain:

$$X^{(i+1)} = \begin{cases} X' & \text{with probability} \quad p(x^{(i)}, X') \\ X^{(i)} & \text{with probability} \quad 1 - p(x^{(i)}, X') \end{cases} \tag{2.43}$$

where p(x,y) is called the Metropolis-Hastings acceptance probability calculated as

$$p(x,y) = min\left\{1, \frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}\right\} \tag{2.44}$$

where $f(x)$ is the target probability density function to which the Markov chain is set to converge. It can be shown that using the acceptance probability presented in Eq. 2.44, the equilibrium distribution of the states of the Markov chain is equal to the $f(x)$ distribution.

In case of using a symmetric proposal density defined as

$$q(x|y) = q(y|x) \tag{2.45}$$

the acceptance probability is also simplified to the form of

$$p(x,y) = min\left\{1, \frac{f(y)}{f(x)}\right\} \tag{2.46}$$

Suppose, for example, that the target probability density $f(x)$ has the form of the Boltzmann distribution defined as

$$f(x) = \frac{1}{Z}e^{-\frac{E(x)}{T}} \tag{2.47}$$

where $E(x)$ is introduced as the energy of state x and T denotes the "temperature" parameter. The normalization factor Z is referred to as the canonical partition function and is calculated as a summation over the **S** state space:

$$Z = \sum_{x \in \mathbf{S}} e^{-\frac{E(x)}{T}} \tag{2.48}$$

Given the Boltzmann target density, combining Eq. 2.46 and Eq. 2.47, the acceptance probability can be calculated as

$$p(x, y) = min \left\{ 1, e^{\frac{1}{T}(E(x) - E(y))} \right\} \tag{2.49}$$

Note that in Eq. 2.49, the acceptance probability does not depend on the normalization factor, Z, which means that one can use the MCMC algorithm to sample from the target Boltzmann density even if $f(x)$ is known only up to a constant of proportionality. As shown by Eq. 2.49, the acceptance probability depends on the $\Delta E = E(x) - E(y)$ energy difference between the current state and the proposed state. If ΔE is non-negative (i.e. the proposed state has lower or equal energy than the current state), the acceptance probability is 1, therefore the proposed state is always accepted. On the other hand, if ΔE is positive (i.e. the proposed state has higher energy than the current state), the proposed state is rejected with certain probability also depending on the temperature parameter. In case the proposed state is rejected, the new state of the Markov chain is set to the previous state: $X_{i+1} := X_i$ (see Eq. 2.43).

According to Eq. 2.47, the target probability distribution $f(x)$ has its global maximum where the energy function $E(x)$ has its global minimum. Thus the above-described MCMC algorithm using the Boltzmann target density can also be applied for global minimization of an arbitrary $E(x)$ function. Since the distribution of states of the Markov chain converges to the $f(x)$ distribution, the state of interest which corresponds to the global maximum of $f(x)$ (and therefore the global minimum of $E(x)$) is sampled with relatively high probability. The lowest-energy state sampled in the MCMC algorithm can therefore be used as an approximation of the global minimum of the energy function.

2.5 Sequence, structure and dynamics comparison

As discussed in Section 1.4.1, in order to study the relationship between the topology of protein sequence, structure, dynamics and function space, one needs precise definitions of sequence, structural, dynamic and function similarity. In *global* pairwise protein alignment methods, in which two proteins are matched from end to end, the alignment score serves a measure of global similarity. By contrast, in *local* pairwise alignments, in which only segments of proteins are matched, the alignment score quantifies local sequence or structural correspondence. Some basic protein comparison methods are discussed here.

2.5.1 Sequence alignments and similarity

The Needleman-Wunsch algorithm[102] is a global pairwise sequence alignment method in which the exact solution (global maximum of the alignment score and corresponding alignment) is calculated in polynomial time by the dynamic programming approach. The input is the two amino acid sequences, A and B of lengths N and M, respectively. The output is an NxM binary matrix X referred to as the alignment matrix, where $X_{ij} = 1$, if residue A_i and residue B_j are aligned and $X_{ij} = 0$ otherwise. The total sequence alignment score is given as

$$Score = \sum_{ij} X_{ij} \cdot S(A_i, B_j) - N_g \cdot g \tag{2.50}$$

where $S(A_i, B_j)$ denotes the 20x20 substitution matrix that defines the match and mismatch scores between each pair of the 20 standard amino acids, N_g is the number of gaps in the alignment and g is the gap penalty score (assuming linear gap penalty scheme). Alternative S substitution matrices include the BLOSUM[104] and PAM matrices[105].

To find the optimal pairwise alignment that maximizes the score function defined in Eq. 2.50, a dynamic programming matrix F is built recursively using the following simple step:

$$F_{ij} = max \begin{cases} F_{i-1,j-1} + S(A_i, B_j) \\ F_{i,j-1} - g \\ F_{i-1,j} - g \end{cases} \tag{2.51}$$

Note, that matrix entry F_{ij} stores the maximal score of all possible subalignments between the first i and j residues of sequences A and B, respectively. Finally, the optimal global alignment is determined in a backtracing process that starts from the highest scoring matrix cell and identifies the optimal-scoring path in matrix F.

2.5.2 Structural alignments and similarity

When different conformations of the same protein are compared, the most common structural similarity measure is the root mean square deviation (RMSD) which is defined between conformations A and B as

$$RMSD(A, B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x_i^A} - \mathbf{x_i^B})^2} \tag{2.52}$$

where N is the number of atoms in the protein. Importantly, to calculate RMSD, the two conformations first need to be superposed using the least square superposition method. Alternatively, structural similarity measures that do not depend on structural superposition such as the distance root mean square deviation (dRMSD) (defined in Section 5.3.1) can be used.

On the other hand, if the structures of different proteins are compared, structural alignment tools are applied to find the optimal mapping between the two structures. As in the case of sequence alignments, the output of structural alignment is a NxM binary alignment matrix X defined in Section 2.5.1.

One of the most commonly used structural alignment algorithms, DALI (Distance Matrix Alignment)[117], is briefly discussed here. The input of the DALI algorithm is the pair of *distance matrices* of the two protein structures to be aligned. The entries of D distance matrix are defined as the pairwise distances between the alpha-carbon atoms in the structure:

$$D_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\| \tag{2.53}$$

DALI aims to identify the optimal match between two submatrices of the two input distance matrices (D^A and D^B) maximizing the following structural alignment score:

$$Score = \sum_{i=1}^{L} \sum_{j=1}^{L} \phi(i,j) \tag{2.54}$$

where the sums run over the two sets of the L aligned residues and $\phi(i,j)$ denotes the contribution of a pair of aligned residues to the score given as

$$\phi(i,j) = \begin{cases} \left(\theta - \frac{|D_{ij}^A - D_{ij}^B|}{D_{ij}^*}\right) w(D_{ij}^*) & if \quad i \neq j \\ \theta & if \quad i = j \end{cases} \tag{2.55}$$

where θ is the similarity threshold parameter, D_{ij}^* is the mean distance of the two alpha-carbon atoms and w is an envelope function applied as weighting.

As a first step, DALI creates a list of similar 6x6 submatrices identified in the D^A and D^B matrices. The method then uses a Monte Carlo algorithm to assemble these 6x6 patterns into larger, consistent, high-scoring pairs of submatrices. Finally, the alignment score is compared to a random background score distribution and is converted into Z-score in order to assess its statistical significance. Besides predicting the optimal structural alignment (i.e. matrix X), the Z-score of the alignment can be used as a pairwise similarity measure between protein structures.

In addition to DALI, several other structural alignment algorithms and tools are available including SSAP (Sequential Structure Alignment Program), CE (Combinatorial Extension method), VAST (Vector Alignment Search Tool) and MatAlign (Matrix Alignment algorithm). All these approaches offer different quantitative measures of global structural similarity. Note that once two proteins are aligned, the RMSD and dRMSD measures can also be calculated between the two sets of aligned residues.

2.5.3 Local alignments of sequences and structures

Some other methods focus on identifying *locally similar* sequence or structural regions in proteins. These tools are especially useful when one is looking for evolutionarily conserved functional sites in the absence of detectable global similarity.

The Smith-Waterman algorithm[103] aims to find the best *local sequence alignment* between two proteins by trying to match sequence segments of all possible lengths optimizing the alignment score defined in Eq. 2.50. This means that non-aligning regions can be excluded from the alignment without gap penalties and mismatch scores. The exact solution of the problem (i.e. identification of the maximum-scoring local alignment) can be calculated in polynomial time by the dynamic programming approach.

In the Smith-Waterman method, similarly to the Needleman-Wunsch algorithm, a dynamic programming matrix H is built recursively using the following simple step:

$$H_{ij} = max \begin{cases} H_{i-1,j-1} + S(A_i, B_j) \\ H_{i,j-1} - g \\ H_{i-1,j} - g \\ 0 \end{cases} \tag{2.56}$$

The important difference between the H matrix and the dynamic programming matrix F used in the Needleman-Wunsch algorithm (Eq. 2.51) is that while F_{ij} entries can be negative, the H_{ij} entries are always non-negative. Similarly to the Needleman-Wunsch algorithm, the optimal local alignment is determined in a backtracing step starting from the highest scoring matrix entry and identifying the optimal-scoring path in matrix H.

Furthermore, more advanced heuristic algorithms such as BLAST (Basic Local Alignment Search Tool)[106] and FASTA (FAST-All)[108] are available that enable significantly faster local sequence alignments more suited for large database searches. The BLAST algorithm first identifies short matches (termed as "words") between the input sequences and then assembles the high-scoring words into longer alignments. Although BLAST is much faster than the original Smith-Waterman method, it does not guarantee to find the optimal local alignment.

Additional tools that enable *local structural alignments* of proteins are available including PINTS[121], ProBiS[122], eF-seek[289] and MolLoc[290]. For example, PINTS (Patterns In Nonhomologous Tertiary Structures) uses a depth-first search strategy to rapidly identify geometrically similar local patterns in two protein structures. Other methods such as ProBiS, use graph theoretical approaches (e.g. maximum clique algorithms) to find local similarities between otherwise globally dissimilar structures.

2.5.4 Dynamics alignments and similarity

In contrast with the problem of aligning 1-dimensional protein sequences or 3-dimensional structures, the comparison of high-dimensional MD trajectories is an even more difficult challenge. However, one may largely simplify the problem either by applying dimensionality reduction methods to the input trajectories (Section 2.2) or using simple measures (e.g. RMSF profiles) for characterizing and comparing residue fluctuations (Section 2.1.2).

The 'residue matching problem'

In addition, analysing the pairwise similarity of protein dynamics is further complicated by the fact that the motions of two non-identical sets of residues need to be compared. Therefore most approaches used to study the dynamic similarity of proteins require them to be aligned. For example, in order to compare the essential dynamics subspaces by calculating the RMSIP similarity measure (Section 2.2.1), one would need an alignment that defines the corresponding residues in the two proteins for which the covariance matrices are calculated. Similarly, an alignment is required when overlaying RMSF profiles calculated for different proteins. In other words, the comparison of protein dynamics is inherently dependent on the alignment matrix X introduced in Section 2.5.1.

Defining matrix X for the purpose of comparative MD analysis will be hereafter referred to as the "residue matching problem" (Figure 2.4).

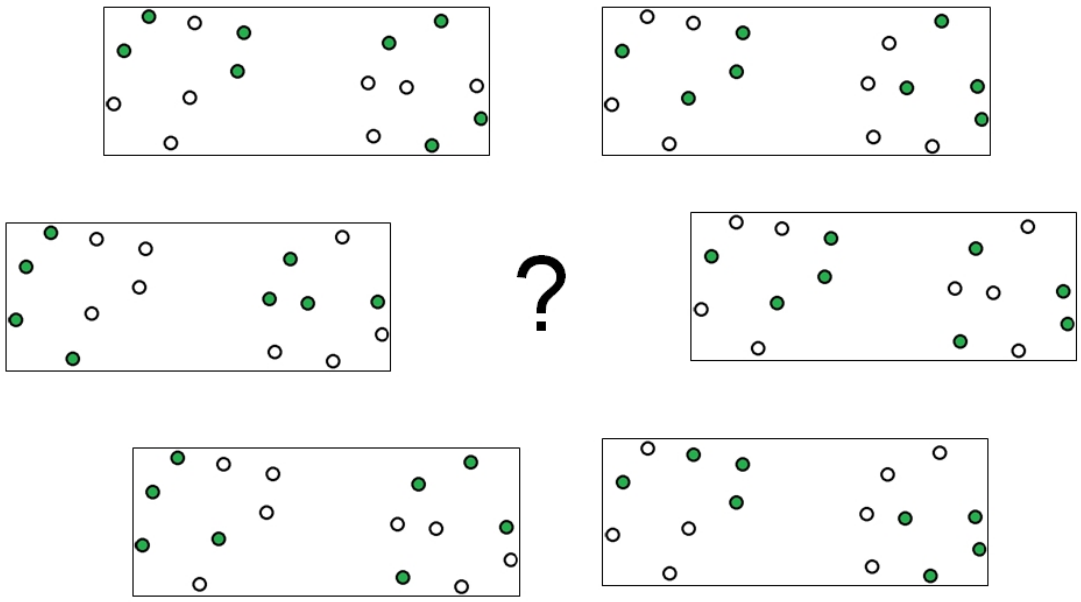

Figure 2.4: In order to study the dynamic similarity of different proteins, a mapping between the non-identical sets of residues is required ("residue matching problem"). As illustrated by this schematic figure, the size of *alignment space* (i.e. the number of possible alignments between two proteins) is exponentially large. Searching in the alignment space for the optimal alignment is therefore a problem of high computational complexity.

Note that in case of comparing different MD trajectories of the *same* protein, the mapping between residues is trivial: matrix X equals the I_N identity matrix.

Two strategies of comparative MD analysis

Two different strategies are discussed here that can be undertaken to address the above-described *residue matching problem*. The first option is to apply prior sequence or structural alignments (e.g. Needleman-Wunsch or DALI alignments) to calculate the X matrix between two proteins and use it for comparing their conformational dynamics. As shown in Figure 2.5A, in this case, information of protein motions (MD data) and information of the prior alignment are both inputs of the comparative analysis, while the output is a similarity score quantifying the dynamic correspondence of the two proteins (given a suitable definition of dynamic similarity).

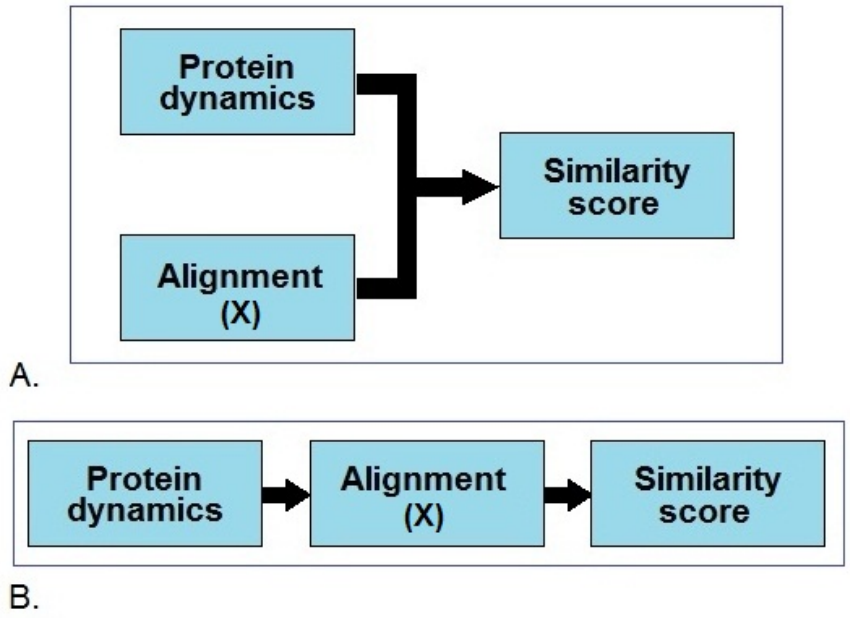

Figure 2.5: Difference between the basic data analysis pipelines of the two described strategies of comparative MD analysis. In the prior alignment-based comparison (**A.**), the data of protein dynamics and prior sequence or structural alignment are both inputs of the comparative analysis which outputs a similarity score measuring dynamic correspondence of the proteins. By contrast, in the dynamics-based alignment (**B.**), only MD data are taken as input and the alignment of proteins is created "on the fly", and is the output of the analysis. The dynamic similarity score also given as output measures the extent the two proteins can be aligned based on their dynamics.

Alternatively, the second option is to create a pairwise alignment "on the fly" based exclusively on the dynamics of proteins. Such alignment could be derived by searching in the "alignment space" to identify the optimal alignment that maximizes the dynamic similarity of the mapped residues. As shown in Figure 2.5B, in this strategy, information of protein motions (MD data) are the only input of the comparative analysis. Importantly, the alignment matrix X is not the input but the output of the pipeline. Furthermore, the similarity score corresponding to the optimal dynamics-based alignment can be used as the *dynamic similarity score* (or *dynamics-based alignment score*) between two proteins.

Note that examples from the literature which apply the above described two strategies of comparative analysis are discussed in details in Section 4.2.

Chapter 3

JGromacs and ABCD

3.1 Summary

One of the objectives of the thesis was to develop novel ways for analysing protein dynamics and to implement these ideas into reusable program code and/or user-friendly applications. In this chapter two software tools that were created are described. The first is JGromacs, a Java library that facilitates the development of cross-platform data analysis applications for Molecular Dynamics simulations. The JGromacs API builds on the strengths of object-oriented programming in Java by providing a multilevel object-oriented representation of simulation data to integrate and interconvert sequence, structure and dynamics information. The easy-to-learn, easy-to-use and easy-to-extend framework is intended to simplify and accelerate the implementation and development of complex data analysis algorithms. Furthermore, a basic analysis toolkit is included in the package. The programmer is also provided with simple tools (e.g. XML-based configuration) to create applications with a user interface resembling the command-line UI of GROMACS applications. The second tool introduced in this chapter is ABCD, a visualisation interface for comparing protein fluctuation patterns. The program implements the idea of prior alignment based comparative analysis described in Chapter 5. The features of both software tools are summarized and examples are presented below.

3.2 Introduction

As discussed in Section 1.2.4, Molecular Dynamics simulations provide a powerful method to study the native dynamics of biological macromolecules with atomistic resolution.[251,291]

Due to recent advances in hardware and software, as well as the development of enhanced sampling techniques, computer simulations now can sample biologically relevant timescales (microsecond and beyond).[292] On the other hand, while simulations can better explore the conformational space of interest, the large number of conformations sampled requires increasingly sophisticated methods for analysis.[293] Data mining and visualization techniques are therefore becoming essential in the extraction of functional information from MD data.

GROMACS[250], mentioned in the previous chapter, is one of the four most commonly used molecular dynamics simulation suites (together with CHARMM[294], AMBER[295] and NAMD[296]). However, GROMACS is the only package of the four that is open-source. The GROMACS suite also includes a series of tools to process and analyse trajectories generated by simulations. Although these in-built tools cover a wide spectrum of standard analysis methods (from principal component analysis to density calculations to clustering), one may need to develop their own analytical tools that process GROMACS trajectories. Even though it is possible to modify or extend the open source GROMACS code written in the programming language C, it would often be more convenient to build applications from scratch that operate on GROMACS data files.

In this chapter I discuss JGromacs, a Java API (Application Programming Interface) which provides full freedom in developing data analysis tools that can directly process GROMACS data. The library contains parsers for GROMACS file formats allowing simulation data to be accessed through the Java code. Data read from input files are stored in an object-oriented architecture representing different levels of structural information (from sequences to structures and trajectories). Processed data can be saved back to GROMACS formats enabling integration of GROMACS and Java-based tools into a data analysis pipeline.

The goal of developing JGromacs was to simplify the analysis of protein motions within the framework of Java, one of the most popular programming languages in academic software development and in particular bioinformatics. The most important reason for the popularity of Java is that it makes cross-platform GUI application development very easy,

and GUIs are often essential to visualise bioinformatics results. At the same time, Java is a powerful and robust object-oriented language[297]. Many existing bioinformatics tools and packages were written in Java (including programming libraries such as BioJava[298], analysis and visualization tools such as StatAlign[299], JMol[300] or Jalview[301] and complete bioinformatics analysis platforms such as Geneious[302]).

BioJava is a mature open-source project providing a framework for the analysis of biological data in general. It provides Java classes representing biological objects and a large collection of analytical and statistical routines covering a wide range of fields of bioinformatics. By contrast, JGromacs is designed to focus on the particular problem of processing and analyzing molecular dynamics (MD) trajectories; therefore, it is a much smaller API with more focused functionality. Packages developed for similar purposes in different programming languages include MDAnalysis[303] and MMTK (Molecular Modeling Toolkit)[304] designed for Python, LOOS (Lightweight Object-Oriented Structure library)[305] designed for C++ and OpenStructure[306] designed for Python/C++. While all frameworks mentioned offer object-oriented design, they have different support for reading and writing trajectory and coordinate file formats. From this point of view, MDAnalysis and LOOS are the most versatile, as they can import and export formats used by multiple MD suites such as GROMACS, CHARMM, AMBER, and NAMD. Unlike the other three packages, MMTK also enables setting up and running MD simulations. MDAnalysis, LOOS, and OpenStructure all offer an atom selection feature; i.e., atom groups can be selected using descriptors and boolean operators. Since JGromacs has been designed to process GROMACS trajectories, it defines atom groups via index sets used by GROMACS tools. By contrast to other packages, it also supports input/output of sequences and multiple alignments and enables the joint analysis of sequence and structural/dynamics data.

The JGromacs API (and detailed documentation) is released under a GPL (GNU General Public) licence and is freely available as an open-source package from the project's website (http://sbcb.bioch.ox.ac.uk/jgromacs). For most of the work covered in this thesis the JGromacs library was used to implement the new analysis ideas.

The structure and main features of the JGromacs API are discussed here first, followed

by an example presenting a simple JGromacs code and its application on a sample MD trajectory. The detailed documentation of the API (including a quick start guide, examples and description of all subpackages, classes and methods) is available on the JGromacs website mentioned above.

In addition, a novel visualisation tool, ABCD (Alignment-based Comparison of Dynamics) is introduced in this chapter. The goal of developing ABCD was to facilitate the comparison of fluctuation patterns (see Section 5.3.2) based on prior pairwise sequence alignments. The GUI was designed to help recognizing equivalent regions in two structures that show different extent of mobility. The interface provides a visual mapping between fluctuation matrix entries and the corresponding positions of the sequence alignment. (For more details about prior alignment-based comparative analysis see Chapter 5). ABCD has also been developed in Java using the JGromacs API and Swing[307] , the primary Java GUI Widget toolkit. The interface has proven to be useful in the comparative analysis of PDZ domains (discussed in Chapter 5). The main features of ABCD are explained in this chapter followed by an example analysis.

3.3 JGromacs API

The most important features of the JGromacs library are summarized in this section and two examples are presented.

3.3.1 Structure and features of the package

Object-oriented description

The JGromacs library comprises 5 subpackages, each of which is a collection of Java classes sharing a distinct function (see the UML (Unified Modelling Language)[308] diagram of the five subpackages in Figure 3.1). The core subpackage, jgromacs.data contains 13 classes representing different levels of structural data from single atoms and amino acid residues to protein structures to complete MD trajectories. The subpackage also contains classes to handle amino acid sequences, multiple sequence alignments, atomic index sets, simulation

frame index sets and mathematical objects such as 3-dimensional points, point sets, angles, matrices and vectors. The objects defined in jgromacs.data are the basic building blocks of JGromacs applications and can be interconverted between each other in many ways.

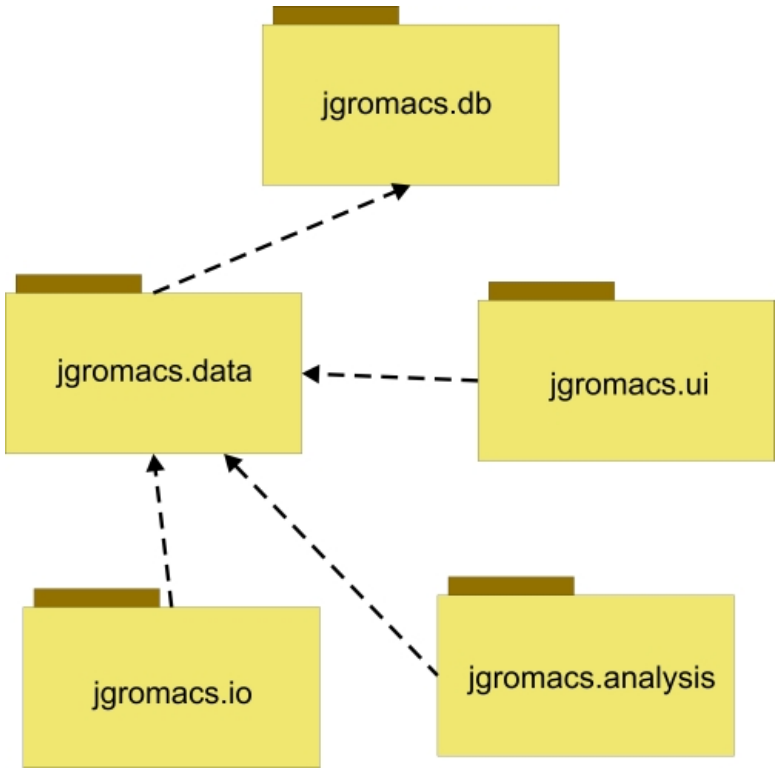

Figure 3.1: UML package diagram representing the inner structure of the JGromacs library. Dependencies between the five subpackages are shown by arrows (an arrow is pointing from package A to package B if package A uses the classes of package B). These five subpackages of JGromacs contain a total number of 25 Java classes (which are also listed and explained in details on the JGromacs website).

Figure 3.2 shows how these hierarchically related classes represent multiple levels of sequence, structure and trajectory information. The class Structure, for example, can be used to store single structural models read from coordinate files and separate polypeptide chains. A Structure object is also a collection of Residue objects that represent amino acid residues, water and other molecules in the structure. On the other hand, a Residue object is a collection of Atom objects representing the atoms in the residue. Atomic coordinates are stored by objects of the Point3D class. JGromacs defines groups of atoms with the help

of index sets, analogously to the index (.NDX) files in GROMACS.

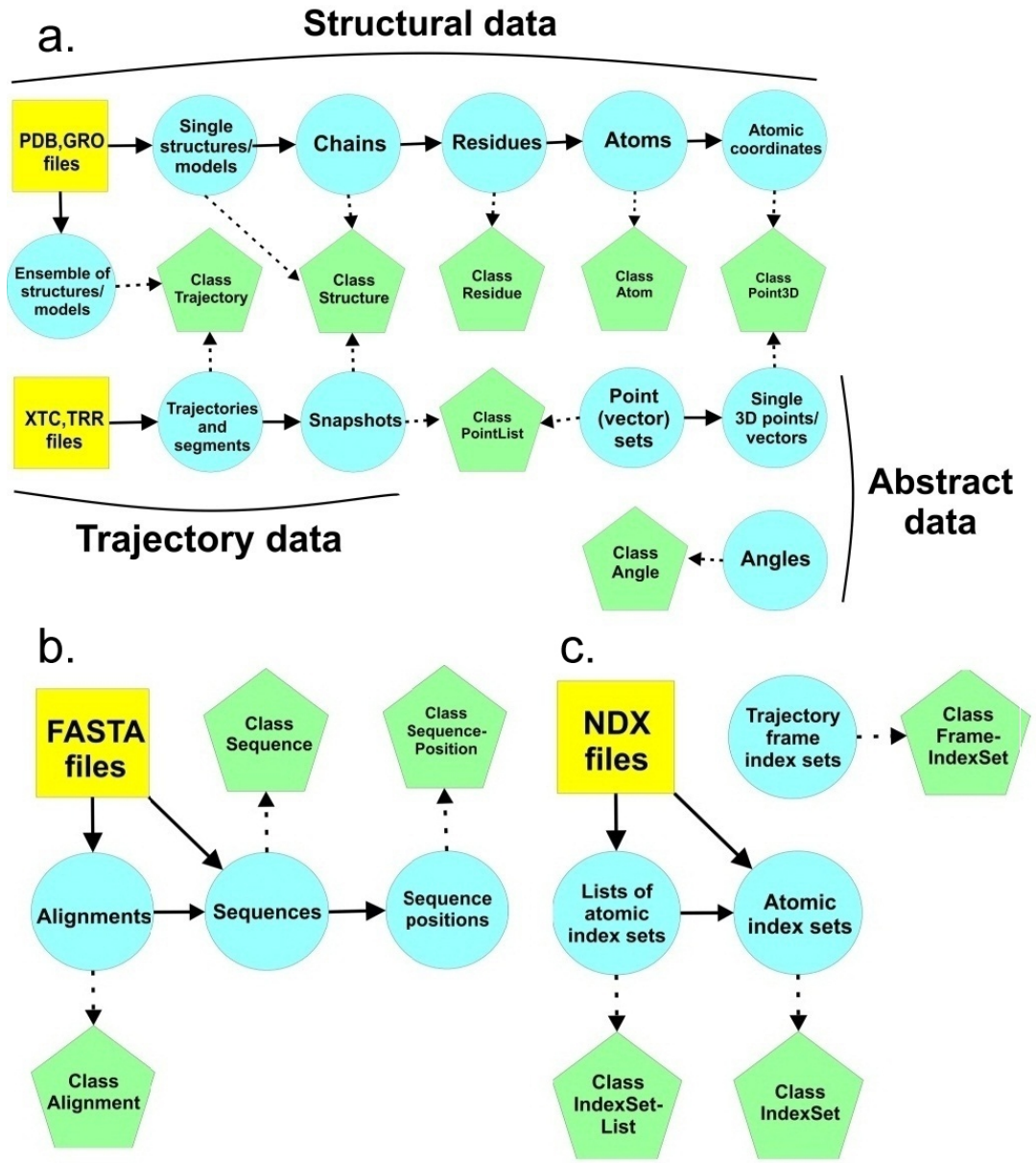

Figure 3.2: JGromacs classes and multiple levels of data represented: **A.** structures and trajectories; **B.** sequences and alignments; **C.** atomic index sets and MD simulation frame index sets. Blue circles depict different levels of data, green pentagons depict Java classes. Arrows between two blue circles mean hierarchical relationships between JGromacs classes. Arrows between blue circles and green pentagons mean mappings between JGromacs classes and the data levels they represents.

MD trajectories and structural (e.g. NMR) ensembles are stored in objects of the Trajectory class. Frames of a trajectory can be retrieved either as Structure or PointList objects which are used to extract atomic coordinates. The Sequence and Alignment classes are designed to represent amino acid sequences and multiple sequence alignments. Atom and residue types are defined in subpackage jgromacs.db.

The classes in jgromacs.data provide methods for retrieving and modifying the properties of data objects such as rotating and translating atoms, calculating interatomic and interresidue distances, extracting trajectory segments, retrieving amino acid sequence from a protein etc. For further information on the functionalities of subpackage jgromacs.data see the API's documentation (available on the JGromacs website).

Parsing GROMACS files

The jgromacs.io subpackage provides parsers for GROMACS data files (including support for PDB, GRO, XTC, TRR and NDX formats) making it simple to import structures, trajectories and index groups to JGromacs objects (which can be saved back to GROMACS files with the output routines of jgromacs.io). The jgromacs.io subpackage also offers parsers and output functions for FASTA format to import and export sequences and alignments. Importing and exporting data between GROMACS files and JGromacs objects enables to connect Java tools and GROMACS tools as an integrated data analysis pipeline. Furthermore, the subpackage jgromacs.io provides an option to execute any GROMACS commands from within the Java code and automatically import the output files as JGromacs objects.

In-built analysis toolkit

The subpackage jgromacs.analysis offers a collection of analytical routines covering various areas from calculating dihedral angles to extracting contact matrices to weighted superposition of structures. Making use of the toolkit one can for example retrieve the mean distance matrix or covariance matrix of a trajectory, calculate the RMSIP (root mean square inner product) between conformational subspaces, look at the cumulative variance profiles

in PCA, extract time series of interatomic distances or dihedral angles, find the simulation snapshot where two atoms are in closest proximity, use Gaussian network models and many more. These analysis functions operate on the objects defined in subpackage jgromacs.data. The toolkit can easily be extended with additional routines that fit into this framework.

User Interface support

Finally, subpackage jgromacs.ui provides a simple way to add a user-friendly interface to JGromacs applications. The UI can easily be set up with an XML configuration file. It supports help messages, log files and command line argument parsing and in many aspects resembles the UI of GROMACS tools.

3.3.2 First example: dynamical networks

As demonstrated in the examples below, complex concepts that would normally take hours to code up from scratch can be implemented in a matter of minutes with the help of the JGromacs library. The first example illustrates how JGromacs simplifies the implementation of the idea of *dynamical networks*.[309]

Dynamical networks

The definition of dynamical networks was introduced by Sethi et al.[309] to study allosteric signalling in tRNA:protein complexes. Their idea was to represent a tRNA:protein complex as a weighted graph in which each amino acid residue and nucleotide of the complex is represented by a single node. Two nodes are connected in the network if the monomers are in contact: i.e. their closest heavy atoms are within 4.5 Å of each other for at least 75% of the MD simulation frames. An edge between nodes i and j is weighted by the absolute value of the C_{ij} correlation between the two monomers calculated over the course of the MD simulation. The weight of a link estimates the probability of information transfer between the two residues. The length of a link was defined as $-\log|C_{ij}|$. Adding information about dynamics, these networks give a more realistic picture about the system than

the unweighted protein structure networks (PSN) constructed based on the contact pattern of a single structure. Sethi et al. used network analysis concepts (i.e. shortest path, betweenness centrality, suboptimal path and community analysis) to identify nodes and paths in the network crucial for intramolecular signal transduction, highlighting possible allosteric communication pathways within the complex. (See more details of this analysis in Chapter 6 where it was applied to study the allosteric communication in the second PDZ domain of the mouse PTP-BL protein.)

Implementation in JGromacs

The following short JGromacs code calculates the weight matrix of the dynamical network of a protein from a GROMACS MD trajectory:

```
Structure s = IOData.readStructureFromGRO("example.gro");
Trajectory sim = IOData.readTrajectory(s,"example.xtc");
int d = sim.getNumberOfResidues();

Matrix contact = Distances.getFrequencyContactMatrix(sim,
                    Distances.CLOSESTHEAVY, 0.45, 0.75);

IndexSet alphaCarbons = s.getAlphaCarbonIndexSet();
sim = sim.getSubTrajectory(alphaCarbons);

Matrix correl = Dynamics.getAtomicCorrelationMatrix(sim);

Matrix W = new Matrix(d,d,0);
for (int i=0; i<d; i++) {
  for (int j=i+1; j<d; j++) {
    if (contact.get(i,j)==1) W.set(i,j, Math.abs(correl.get(i,j)));
    else W.set(i,j, Double.NaN);
    W.set(j,i, W.get(i,j));
  }
}
```

As a first step, the example code imports structure and trajectory data from GRO and XTC files. It then determines the frequency-based contact matrix using 4.5 Å distance cutoff and 0.75 contact probability cutoff. After extracting the trajectory of alpha carbon

atoms, it calculates their correlation matrix. Finally, the contact and correlation matrices are combined into the output matrix W that defines the connectivity and weights of the dynamical network. The weight matrix W is the input of further analysis.

Application to example data

Given here are the results of executing the short code presented above on an example dataset; a 20 ns MD simulation of the N-terminal PDZ domain of InaD (Inactivation no afterpotential D) protein from Drosophila. Figure 3.3 shows the 94x94 weight matrix outputted by the JGromacs code. Figure 3.4 presents the resulting dynamical network as depicted by the network analysis and visualisation software Pajek[310].

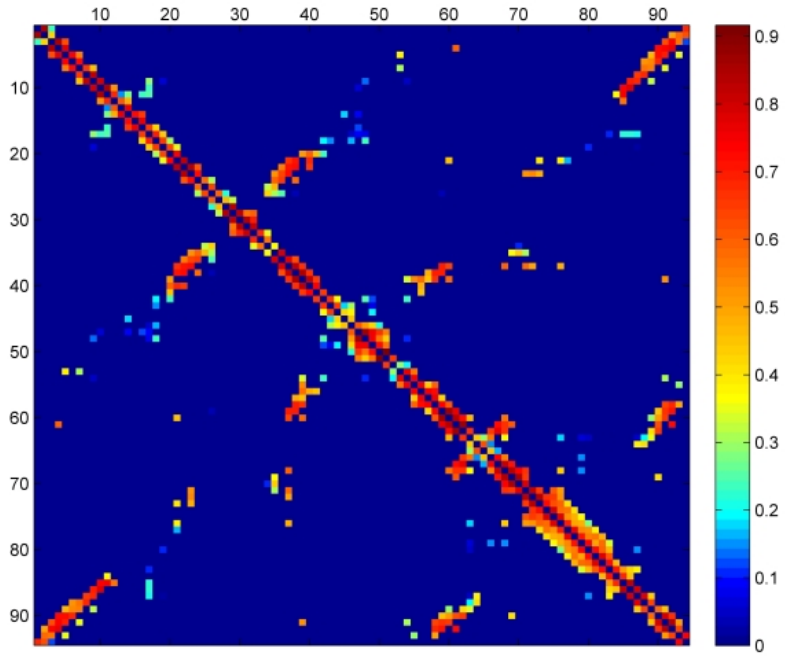

Figure 3.3: Weight matrix describing the dynamical network of the PDZ1 domain of InaD protein from Drosophila based on a 20 ns MD simulation.

We can see that even in the case of very compact structures such as PDZ domains the network may contain both sparsely and densely connected regions which have different

signal propagation capabilities. Starting from scratch, generating this network from an MD trajectory file would be time-consuming, but JGromacs significantly reduces programming time.

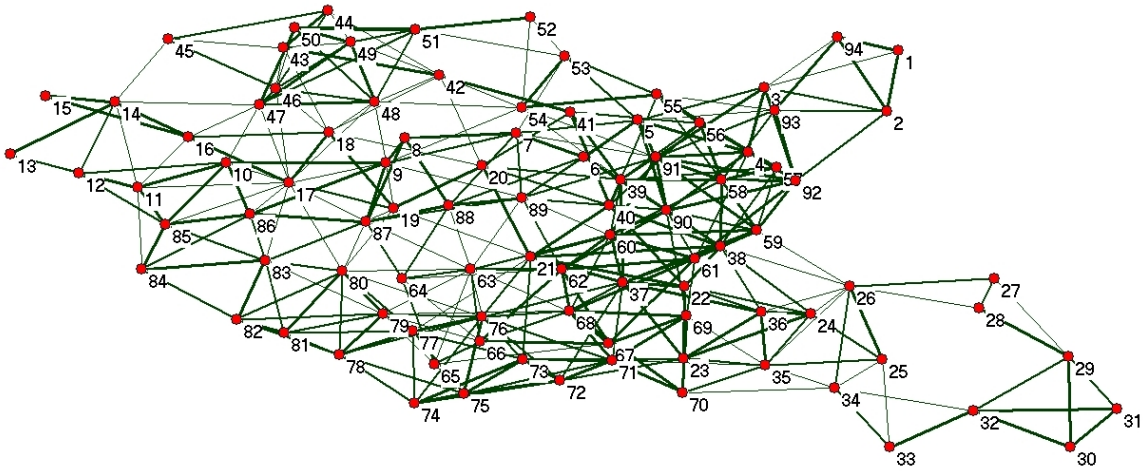

Figure 3.4: Dynamical network of the PDZ1 domain of InaD protein from Drosophila based on a 20 ns MD simulation. The layout is created with the network program Pajek[310] using the Kamada-Kawai algorithm[311]. Nodes represent residues, edge widths are proportional to link weights.

3.3.3 Second example: weighted superposition

Dynamically weighted superposition

In studying a conformational ensemble of a protein, the first step usually is the superposition of structures on each other or on a common references structure. The standard method is finding the optimal rigid transformations (rotation and translation) that minimise the RMSD difference between two structures. The Kabsch algorithm[312] provides a simple way to calculate the optimal rotation matrix between two sets of points in a three dimensional space. As a first step, both sets of coordinates are translated so that their centroids coincide with the origin of the coordinate system. Then the covariance matrix between the two coordinate matrices is computed. Finally, the optimal rotation matrix is calculated from the singular value decomposition (SVD) of the covariance matrix.

In the standard procedure all points are treated equally assuming that they are equally

important in determining the optimal fit between the two structures. As a generalization each point i is given a weight w_i to reflect their relative importance in the superposition. A number of weighted schemes have been proposed including weighting by the atomic masses, giving different weights to backbone and side chain atoms or assigning larger weights to atoms belonging to secondary structural elements. The generalization of the Kabsch algorithm for the weighted superposition problem involves centering the two structures by their weighted center of mass, computing the weighted covariance matrix and calculating the optimal rotation matrix from the SVD of the covariance matrix.[313]

Wu and Wu[314] proposed that the vector of root mean square fluctuation (RMSF) values predicted by a Gaussian Network Model (GNM) could be used as a sensible weighting scheme in superposition (see Section 2.3.1). The weight assigned to atom i was set to $d_i = (B_i)^{-m}$, where B_i is the RMSF of atom i calculated from the GNM and m is an integer. The larger fluctuation an atom has in the GNM, the less weight it is given in the superposition. As this method automatically underweights flexible (e.g. hinge) regions, Wu and Wu could use it successfully in a number of cases in which unweighted superposition failed including protein domain identification and multiple structural alignments.

Implementation in JGromacs

The following short JGromacs code calculates the dynamically weighted superposition of two structures where weights depend on the root mean square fluctuations derived from a Gaussian Network Model.

As a first step, the example code imports two structure from PDB files. It then creates a Gaussian Network Model based on the first structure using only the alpha-carbon atoms and 7 Å contact distance cutoff. Next the mean square fluctuation profile is computed and converted to the weight vector (using $m = 6$). Finally, the code performs weighted superposition of the two structures with the weight vector calculated in the previous step.

```
Structure s1 = IOData.readStructureFromPDB("example1.pdb");
Structure s2 = IOData.readStructureFromPDB("example2.pdb");

GNM gnm = new GNM(s1, 0.7, Distances.ALPHACARBON);

ArrayList<Double> msf = gnm.getMSFProfile();

ArrayList<Double> W = new ArrayList();

for (int i=0; i<msf.size(); i++)
W.add(Math.pow(Math.sqrt(msf.get(i)),-6));

s1 = Superposition.weightedSuperposeTo(s1, s2, W);
```

Application to example data

Given here are the results of executing the short code presented above on two conformations of the RAN (Ras-related Nuclear) protein taken from two X-ray structures (PDB IDs: 1BYU and 1RRP). RAN is a trans-membrane protein responsible for importing proteins into the nucleus and exporting RNA molecules.[315]

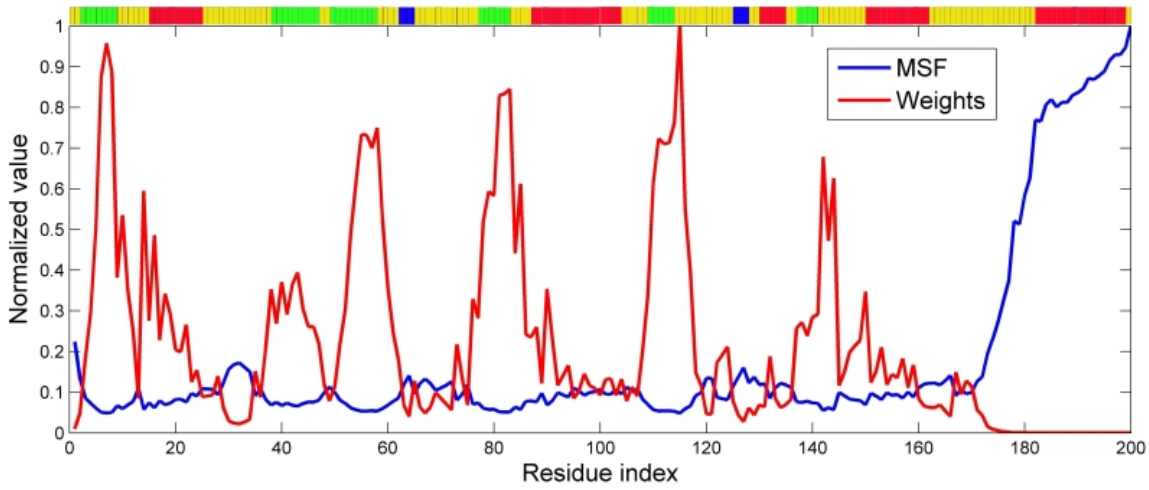

Figure 3.5: The normalized mean square fluctuation profile (blue line) predicted by a GNM model (based on the 1RRP conformation) that is transformed into the normalized weight profile (red line) used in the superposition process. Secondary structural annotations provided by Stride are also shown: green, red, blue and yellow regions represent beta-sheets, alpha-helices, 3_{10}-helices and random coils, respectively.

Figure 3.5 shows the normalized mean square fluctuation profile predicted by GNM

based on the 1RRP conformation. The normalized weight profile calculated from the RMSF values is also shown. In addition, Figure 3.5 depicts the secondary structural annotations as predicted by Stride[316] . We can see, for example, that there is a 38 residue-long extremely flexible region (random coil + alpha helix) at the C-terminal end of the protein that is given a weight close to zero in the superposition process.

Figure 3.6 shows the result of weighted superposition of the two conformations compared to the standard unweighted RMSD superposition. As it is clear from the figure, the dynamically weighted superposition provides a better fit of the two structures than the unweighted superposition which fails due to the presence of the highly mobile C-terminal region.

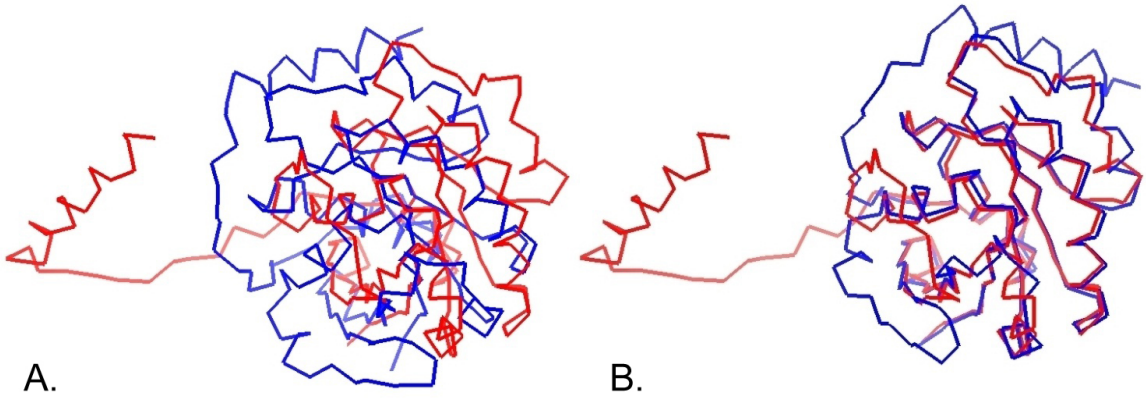

Figure 3.6: Unweighted and weighted RMSD superpositions of two conformations of the RAN protein; 1BYU (blue) and 1RRP (red). Only the alpha-carbon traces are shown by tube representations. **A**. result of unweighted superposition; **B.** result of weighted superposition.

3.4 ABCD

The most important features of the ABCD (Alignment-based Comparison of Dynamics) program are summarized in this section and an example illustrated with screenshots is presented.

3.4.1 Features of the program

ABCD provides a simple interface for comparing two protein backbone fluctuation matrices calculated from two MD trajectories as defined in section 5.3.2. As a first step the required seven input files are imported: two GROMACS trajectory (XTC or TRR) files, the corresponding coordinate (GRO) files with secondary structure (Stride) annotations and a pairwise sequence alignment (FASTA) file. For both simulations the program calculates the fluctuation matrices of the conserved residues; i.e. residues not aligned to gap in the sequence alignment. (See the detailed explanation of the methodology in Section 5.3).

The two fluctuation patterns can be compared with the help of a simple GUI (graphical user interface) that enables the user to locate differences between the two matrices. With the upper tabs one can switch between three different views: showing the fluctuation matrix of the first or second protein (F_1 and F_2) or the difference of the two fluctuation matrices ($\delta F = F_1 - F_2$). Positive and negative entries in the difference pattern indicate that two residues are more mobile with regards to each other in the first or second trajectory, respectively. On the other hand, matrix entries close to zero indicate pairs of residues that have similar relative fluctuation in the two simulations.

The color coded matrices make it simple to recognize regions of different mobility between the two simulations. The user can change the color scale and also apply a threshold to the matrix entries to be visualized. While positioning the mouse cursor to a certain matrix entry, the program automatically highlights the corresponding residues in an alignment panel where secondary structural motifs are also annotated. The integrated visualization of 2D fluctuation patterns and 1D sequence information serves as a useful tool for finding key residues that have different dynamics in the compared conformational ensembles.

The user can have a closer look at a certain subpattern by selecting and zooming on an arbitrary region of the fluctuation matrix. The zooming feature is particularly useful if one is studying large structures (consequently dealing with large fluctuation matrices) or is interested in the dynamics of specific sites (e.g. binding site) of the proteins. The sequence regions for which the fluctuation patterns are enlarged are highlighted in alignment panel.

When a submatrix is selected, ABCD automatically calculates the mean fluctuation value of the pattern (defined as the mean of all matrix entries in the submatrix; see in Section 5.3.2) that characterizes the relative motion between two subsets of residues. The mean fluctuation values summarizing the selected pattern are calculated for both trajectories and can therefore be directly compared.

An additional feature offered by ABCD is the automatic detection of all maximal subpatterns in the fluctuation matrix that fit certain requirements selected by the user. One can search for subpatterns in which matrix entries are lower or higher than a specified threshold value or lie in a given interval. The maximality of the outputted patterns means that no larger submatrices containing these patterns match the search requirements. The user can also set the minimal size of submatrices to be listed. The program highlights all patterns matching the search criteria and lets the user navigate through the submatrices one by one. This tool is useful when one would like to find equivalent structural elements that show different extent of fluctuation in the two simulations. For example, searching the difference fluctuation matrix (δF) for patterns of entries larger than a positive threshold will result a list of submatrices that describe regions more mobile in the first than in the second trajectory.

Finally, it is important to note that in addition to comparing the dynamics of two different (homologous) structures, ABCD can be used to compare the fluctuation patterns of the same protein simulated under different conditions (e.g. temperatures, force fields, ligands) or to compare different segments of the same simulation. In that case the fluctuation matrices are calculated for the whole set of residues and no sequence alignment is needed.

3.4.2 Example analysis

The example in this subsection illustrates how the graphical interface of ABCD can be used to perform comparative analysis of two MD simulations. Two 20 ns GROMACS trajectories are imported to the program: apo simulations of the PDZ domain of Alpha-1 Syntrophin from mouse (PDB: 1QAV) and the human Dishevelled-2 PDZ domain (PDB: 3CBX). The input alignment of the two proteins was created using the Needleman-Wunsch

pairwise sequence alignment algorithm[102]. The number of conserved residues in the alignment is 78, therefore the fluctuation matrices compared in the analysis are of the size 78x78. After selecting the 7 input files detailed in the previous section, the program calculates the matrices and the main window shows up.

Looking at pairwise residue fluctuations

Figure 3.7 gives a screenshot of the main window of ABCD when visualizing the difference fluctuation matrix (δF) of the two trajectories. In the the color-coded matrix in the left side panel, red and blue areas indicate positive and negative difference fluctuation values; i.e. regions that have higher mobility in the first or in the second trajectory, respectively. The alignment panel at the lower right corner gives a schematic view of the input sequence alignment (the matched residues only, columns with gaps are removed). The alignment panel is annotated with secondary structural information read from the input Stride files (green, red, blue and yellow regions of the sequences represent beta-sheets, alpha-helices, 3_{10}-helices and random coils, respectively.)

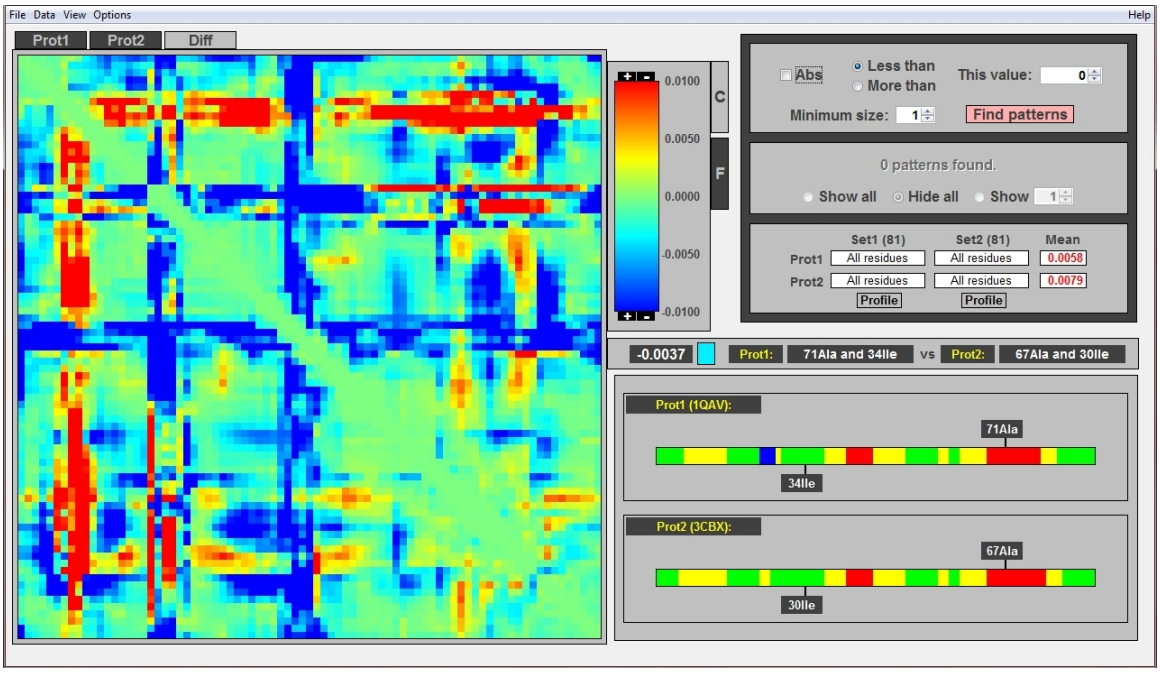

Figure 3.7: Screenshot of ABCD showing the difference fluctuation matrix.

In figure 3.7, two residues are selected from the matrix panel: 34Ile and 71Ala of the Alpha-1 Syntrophin PDZ domain aligned to 30Ile and 67Ala of the Dishevelled-2 PDZ domain. The exact locations of these residues are automatically shown in the alignment panel: in both sequences they are found in the β3 strand and α2 helix of the PDZ domains. However, these two equivalent pairs of residues have different relative motion in the two trajectories. The difference fluctuation value corresponding to the selected matrix entry is -0.0037 as outputted by the program. The negative difference means that 34Ile and 71Ala fluctuate less with regards to each other in the Alpha-1 Syntrophin PDZ simulation than 30Ile and 67Ala do in the Dishevelled-2 PDZ simulation.

Selecting and zooming on subpatterns

The screenshot in Figure 3.8 illustrates how the zooming feature can be used to focus the analysis on a specific region. A 10x6 submatrix including the binding site residues (i.e. residues of the α2 helix and β2 strand) is selected in the matrix panel and enlarged using the zoom tool. The selected subpattern describes the fluctuation of the 20Ile-25Gly region with regards to the 68His-77Lys region in Alpha-1 Syntrophin PDZ and the 14ILe-21Asn region with regards to the 64Asn-73Asp region in Dishevelled-2 PDZ. While the color-coded matrix panel is visualizing the enlarged submatrix, the selected sequence regions are also highlighted in the alignment panel. The enlarged submatrix shows interesting pattern: while some pairs of binding site residues (e.g. 25Gly and 77Lys of Alpha-1 Syntrophin PDZ) have larger pairwise fluctuation in the first simulation than their equivalents have in the second, other pairs of residues (e.g. 15Ser and 67Ala of Dishevelled-2 PDZ) are more mobile in the second protein than their equivalents in the first one. The overall mobility of the binding pocket is slightly larger in the first trajectory than in the second as compared by the mean fluctuation values of the two patterns (0.0131 vs. 0.0110) outputted by the upper right panel of the GUI.

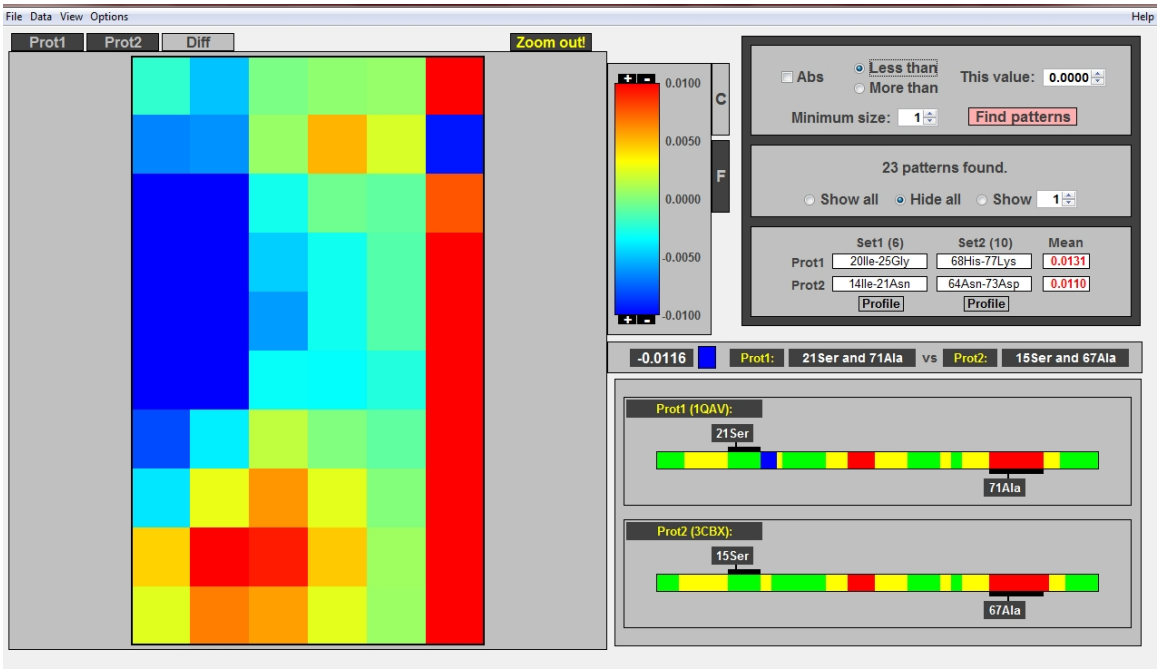

Figure 3.8: Screenshot of ABCD: zooming on a 10x6 pattern of binding site residues.

Automatic pattern detection

Figure 3.9 illustrates the use of the automatic pattern detection feature of ABCD. The screenshot shows the result of searching the difference fluctuation matrix for submatrices in which the absolute value of all entries are larger than 0.0031. (The minimal size of submatrices was set to 4x4.) The search criteria are selected using the search panel in the upper right corner (see in Figure 3.8). The program finds 23 patterns satisfying the search criteria that are automatically highlighted in black rectangles in the matrix panel. Each pattern can be selected individually and enlarged with the zoom tool. In addition, the mean fluctuation values calculated for each pattern are outputted by the program and can be compared between the two trajectories. Since in this case the search criteria was set to detect submatrices in which difference fluctuation values cannot be close to zero, the 23 patterns found correspond to regions where pairwise residue fluctuations are consistently dissimilar in the two proteins. On the other hand, the automatic pattern detection feature can be used to find regions where fluctuation values are consistently similar in the two

simulation.

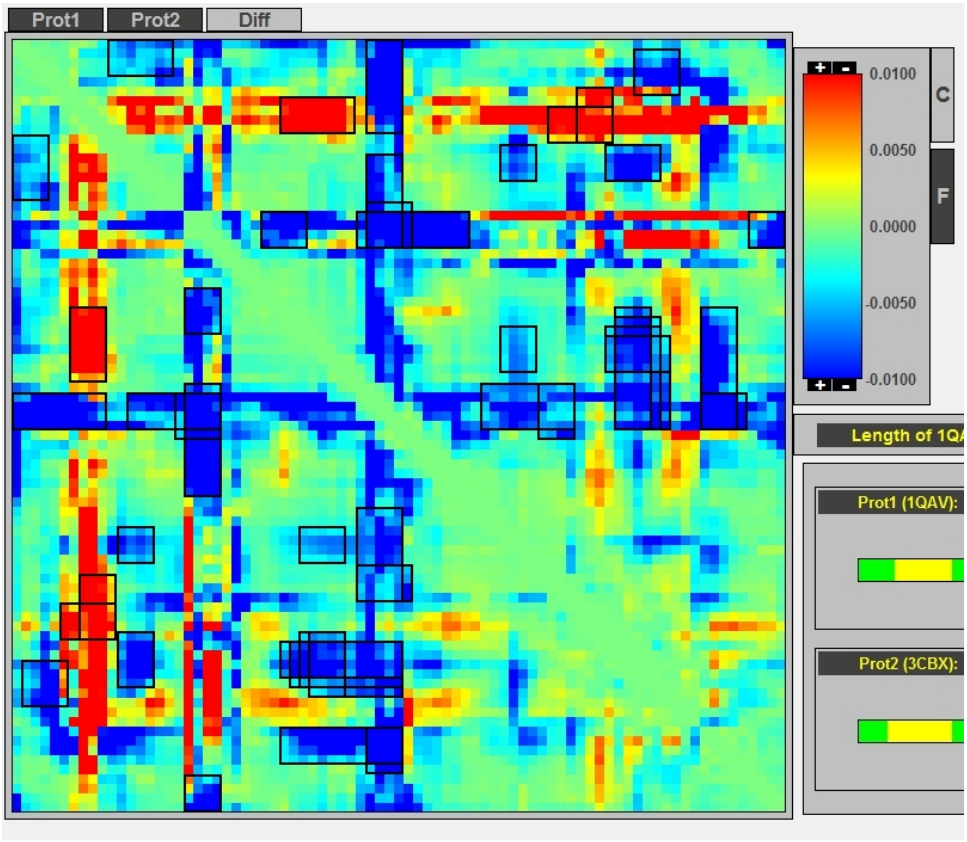

Figure 3.9: Screenshot of ABCD: 23 patterns highlighted (in black rectangles) in the matrix panel visualizing the search results.

3.5 Conclusions

As computer simulations are becoming more and more effective in sampling the conformational dynamics of biological macromolecules, the storage, management and analysis of the generated massive simulation data are increasing challenge. To address these problems, the BioSimGrid platform[317], for example, aimed to serve as an online repository for biomolecular simulations and to offer a rich analysis suite for molecular dynamics trajectories. The objective of BioSimGrid was to provide an extensive toolkit of standard analysis routines facilitating cross-comparison of the deposited trajectories. On the other hand, molecular dynamics software packages such as GROMACS and CHARMM have

their own in-built analysis tools providing the significant advantage of performing simulations and analysis within the same framework. However, in addition to making use of the standard analysis routines implemented in these platforms, one may also need a flexible framework for developing their own novel tools for analysing MD data.

The product of this research work, JGromacs is a lightweight Java library supporting simple and fast development of analytical tools for datasets produced with the commonly-used MD software GROMACS. The objective of the project was to create a framework for implementing increasingly complex analytical routines that can be used through simple user interfaces. Since in research the goal is not always to develop ready-made applications but to experiment with new ideas as quickly as possible, simplicity of the package was of utmost importance.

While JGromacs also contains a standard analysis toolkit, its main advantage is that it provides an object-oriented framework for novel tool development. The programmers can easily build up their own algorithms and applications based on the basic JGromacs classes and analytical routines already implemented in the package. Furthermore, the library provides options for integrating Java and GROMACS analysis tools.

Two brief examples were presented in this chapter to illustrate how JGromacs package simplifies the development of analysis tools and reduces programming time. More example codes and a step-by-step quick start guide are included in the documentation available on the JGromacs website where a more detailed description of all subpackages and Java classes is also presented. The self developed JGromacs API has been used effectively in most parts of the research project that are discussed in the following chapters of this thesis.

An important advantage of a Java analysis framework is that it can easily be integrated with GUI development tools to combine data analysis and visualization. As an example to this, a graphical interface, ABCD, has been developed that is built on JGromacs and implements the idea of prior alignment based comparative analysis of protein dynamics (discussed in Chapter 5). The key feature of the ABCD interface is the integrated visualization of dynamical and sequence data facilitating the discovery of equivalent (conserved

or identical) regions of interest that show different extent of fluctuation in two MD simulation. An example was presented in this chapter to demonstrate the usefulness of the program. ABCD has been used effectively in the alignment-based comparative analysis of PDZ domains (discussed in Chapter 5).

Both software products, JGromacs and ABCD, have been developed with the hope that they would serve as useful tools in MD data analysis projects for other researchers too.

Related publication:

Münz, M. and Biggin, P.C. (2012). JGromacs: a Java package for analyzing protein simulations. *J Chem Inf Model*, 52(1):255-9

Chapter 4

Dynamics-based alignment of proteins

4.1 Summary

As discussed in Section 1.2, the dynamics of many proteins are central to their function. It therefore follows that the dynamic requirements of a protein are evolutionary constrained. In order to quantify this, one needs to compare the dynamics of different proteins. Comparing the dynamics of distinct proteins may also provide insight into how protein motions are modified by variations in sequence and, consequently, by structure (see Section 1.4). The optimal way of comparing complex molecular motions is, however, far from trivial. The majority of comparative molecular dynamics studies performed to date relied upon prior sequence or structural alignment to define which residues were equivalent in 3-dimensional space. A novel approach of comparing the dynamics of proteins is introduced here following the strategy discussed in Section 2.5.4. The aim of the method is to derive pairwise protein alignments exclusively from protein motions based on MD data. The methodology therefore does not require prior alignment information. As demonstrated, it is possible to align proteins based solely on their dynamics and that these dynamics-based alignments can be used to quantify the dynamic similarity of proteins. The method was tested on 10 representative members of the PDZ domain family. As a result of creating pairwise dynamics-based alignments of PDZ domains, evolutionarily conserved patterns have been detected in their backbone dynamics. The dynamic similarity of PDZ domains is highly correlated with their structural similarity as calculated with DALI. However, significant differences in their dynamics can be detected indicating that sequence has a more refined role to play in protein dynamics than just dictating the overall fold.

4.2 Introduction

It is well established that conformational flexibility plays a key role in the biochemical functions of proteins.[15,318] As discussed in Section 1.2.3, several studies have managed to the relate internal protein motions to biochemical functions[58,319], and in particular the characterization and prediction of large-scale conformational changes via the use of normal modes[320] and elastic-network models[321] has been very successful. However, as discussed in details in Section 1.4, it is not clear if slight variations in structure can lead to large variations in dynamics, or very similar protein structures always have similar motions.[187] A number of examples are known (e.g allosteric interaction in PDZ domains and catabolite activator protein[58,60,221,222,322]) where large changes of the dynamics can occur without notable changes of the structure. These examples suggest that structural similarity is not necessarily a good predictor for dynamic similarity of proteins.

As described in Section 1.2.4, molecular dynamics (MD) simulations can be used effectively to explore the conformational energy landscape accessible to proteins[73,251] giving an insight into how protein dynamics relates back to structure and sequence. Comparative MD studies address the above described questions by performing MD simulations of multiple proteins and comparing their dynamic trajectories. Previous comparative MD studies of proteins fall into two main categories. Studies in the first class compared the dynamics of the same protein simulated under different conditions[323–326]. In this case, the question is how the motion of the protein is altered by the new condition, for example the presence of a ligand. By contrast, the second class of studies compared the dynamics of different proteins simulated under the same condition[89,185,186,327] in order to pinpoint similarities and differences in functionally important movements.

If the dynamics of non-identical proteins are compared, a mapping between the different structures is often required (see the "residue matching problem" in Section 2.5.4). Therefore, as discussed in Chapter 1, a common point of previous comparative MD studies of homologous proteins is that they use prior sequence or structure alignments to find residue equivalencies between the proteins. For example, to compare the fluctuation of dif-

ferent cold-active enzymes, Spiwok et al.[89] used structural alignment to define equivalent residue pairs between the proteins. Papaleo et al.[327] performed MD simulations of different elastases to compare their molecular flexibility. Here the correspondences between residues of different proteins were derived from pairwise sequence alignments. Another example for alignment-guided comparison of protein dynamics is the work of Pang et al. [185] who simulated a series of proteins within the same fold family. To compare the fluctuations as well as the principal components of dynamics of only the structurally conserved residues across the set of proteins, they used structural alignment to define conserved positions. Alternatively to MD, simplified models called Gaussian Network Models were used to explore the common dynamics of the globin family[328] and the protease superfamily[23]. In these studies, however, comparative analysis of dynamics also relied on prior alignments of the proteins.

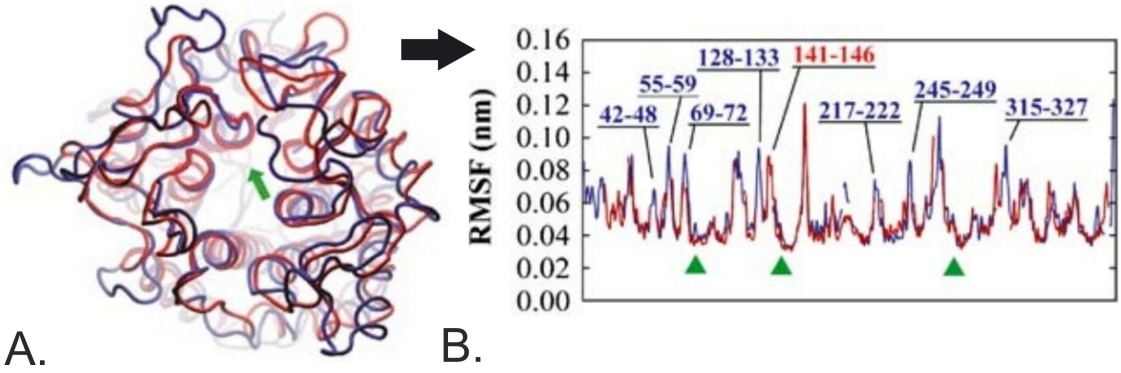

Figure 4.1: Example of the most common approach of comparing protein motions, the prior alignment-based analysis. Spiwok et al. used the Combinatorial Extension (CE) structural alignment method[118] to match the flexibility profiles of cold active enzymes and their mesophilic or thermophilic counterparts. **A:** Structural alignment of cold-active xylanase from *Pseudoalteromonas haloplanktis* (PDB: 1h12) and thermophilic xylanase from *Clostridium thermocellum* (PDB: 1cem), shown in blue and red, respectively. The location of active sites is shown by green arrow. **B:** RMSF profiles of the two proteins overlaid based on the CE alignment. Active site residues are shown by green triangles. (Image courtesy: Spiwok et al. 2007)[89]

By contrast the approaches used in these studies in which alignments were a priori inputs of the comparison of dynamics, an alternative method is introduced in this chapter to measure the similarity of backbone dynamics of proteins without the use of any prior alignment information (a strategy introduced in Section 2.5.4). The method creates pair-

wise alignments of proteins based exclusively on their backbone motions without taking into account their sequence and structure. The input of the algorithm are the MD simulation trajectories of the two proteins to be aligned.

The reason this dynamics-based alignment methodology is proposed here is twofold. First, dynamically equivalent regions of proteins may not match to sequentially and structurally similar regions. Therefore sequence or structural alignments may mislead the comparison of protein motions. An alignment, however, that is built based on solely the dynamics of the proteins, could match parts of two structures that have similar motions even if there is low global structural similarity between the two proteins or the sequence/structural alignment does not match these dynamically similar regions. As the aligned motion patterns may be of functional importance, the dynamics-based alignment approach might be helpful to understand the relationship between functionally similar but structurally different proteins.

Secondly, the pairwise alignment method also provides a simple way to quantify the similarity of two proteins. Alignment algorithms usually perform optimization of some scoring (or objective) function. The optimal alignment is defined as the alignment that gives the maximal score and the highest score found is referred to as the alignment score between two proteins. In other words, the alignment score shows the alignability of two proteins, therefore serving as a natural measure of their similarity. For example, the alignment score calculated with the Needleman-Wunsch global sequence alignment algorithm is a single-number measure of the similarity of the two aligned sequences[102] (Section 2.5.1). Likewise, the optimal score given by the DALI structural alignment algorithm serves as a reliable measure of structural similarity[117] (Section 2.5.2). Similarly, the scores of dynamics-based alignments are used here to quantify the dynamic similarity of proteins. While clustering proteins based on their pairwise sequence or structural similarity scores may not give functionally homogeneous clusters, in some cases the dynamic similarity score may correlate better with functional similarity. For example, two proteins that have high structural similarity score may have low dynamic similarity score indicating their functional dissimilarity despite their structural agreement.

The dynamics-based alignment method introduced in this chapter is therefore an attempt to include an additional layer of description (i.e. the "dynamics space") of the protein universe (Section 1.4.4). As discussed in details in Section 1.4, the distribution of proteins in the sequence, structure and function spaces have been well studied[99,113] and the nature of mapping between these spaces has also been investigated[126,164,169]. However, describing how proteins are distributed in the dynamics space (a layer between the structure and the function spaces) and how sequence/structure map to dynamics could lead to a better understanding of protein evolution and function. Importantly, as suggested in Section 1.4, in order to analyse the interrelationships of the protein sequence, structure, dynamics and function spaces, one needs quantitative similarity measures (or metrics) to describe the topology of these spaces. The dynamics-based alignment score proposed in this chapter might be applied to study the topology of the protein dynamics space.

Recently, Zen et al.[183,184] developed a similar method that takes a combined measure of spatial and dynamic consistency to derive an alignment on the fly that can be used to compare the proteins. They have used coarse-grained (β-Gaussian) elastic network models to calculate the lowest-energy modes that dominate the equilibrium fluctuation dynamics of proteins. A stochastic search algorithm has been used to explore the space of putative alignments to optimize a scoring function that measures both the agreement of the spatial position and the accord of the concerted movements of residues in the two proteins. This method has also been implemented in a web server called ALADYN[329] that takes the structure files of the two proteins as input and presents the calculated dynamics-based alignment in an interactive graphical interface. The approach has successfully been applied for comparing some interesting protein pairs such as HIV-1 PR and human β-secretase or Exonuclease III and human adenovirus proteinase to identify dynamically similar but structurally dissimilar regions.

The method proposed in this chapter differs from the approach used by Zen et al. and the ALADYN web server in several points. The most important difference is that the method introduced here takes two all-atom MD trajectories as input, while Zen et al. use a coarse grained elastic network model to approximate the dynamics of the systems.

In other words, the movements aligned by ALADYN are based on the structures of the proteins only as the elastic network model excludes sequence information. For many proteins, the ENM approach has been shown to give reasonably good approximation of the equilibrium motions[87] and these calculations are very fast compared to time-consuming MD simulations. However, if variations of the sequences of proteins significantly modulate their dynamics, the ENM method may fail as it lacks sequence information relying solely on the spatial orientations of residues. For example, the results of this chapter and the next one show that PDZ domains of highly conserved structures but largely dissimilar sequences may have very different dynamic properties, suggesting that their primary sequences have considerable modulating effect on their intrinsic motions. Since the objective of this study was to develop a tool that can be used to detect differences between the details of dynamics of proteins, even if the compared proteins are structurally similar, the ENM approach was not suitable for this purpose and for that reason all-atom MD simulations was performed.

The second main difference of the approach introduced here and the method published by Zen et al. is that their scoring function rewards both the similarity of essential motions and the agreement of the 3D structures, while the objective function used here depends only on the similarity of dynamics; the pairwise alignments created here are therefore purely dynamics-based alignments.

The method described below has been tested on members of the PDZ domain family. PDZ domains have been used as a first test case of the approach because their dynamics-based alignments could be compared with reliable sequence/structural alignments, provided that these domains are structurally highly conserved and standard sequence or structural alignment methods are able to match their 3D structures very well. On the other hand, as discussed in Section 1.5, PDZ domains have diverse functional properties (i.e. largely varying ligand binding specificity profiles). It is therefore important to study whether they have similar dynamics as a consequence their structural similarity or their equilibrium fluctuations are different to some extent reflecting the dissimilarity of their functional properties. Here it is found that while some pairs of PDZ domains have very

similar fluctuations (as shown by their high pairwise dynamics similarity scores and their dynamics-based alignments), other pairs of PDZ domains, despite their structural similarity, have considerably dissimilar dynamics.

4.3 Methods

Besides summarizing key parameters of the MD simulations performed for this study, this subsection gives a detailed introduction to the novel alignment algorithm and provides a definition of the dynamics-based alignment score which is used as a quantitative dynamic similarity measure. In addition, a framework is described to evaluate the statistical significance of the dynamics-based alignment score between proteins of variable sizes.

4.3.1 Molecular Dynamics Simulations

All MD simulations carried out in this study (i.e. simulations of PDZ domains and members of the Reference Set described in Section 4.3.6) have been performed under the following protocol. The protein has been solvated in a box of explicit solvent (using SPC water model), while water molecules of the original X-ray structure located within 4 Å from the protein were also included. Na^+ and Cl^- ions were added at random locations to achieve 150 mM/L salt concentration. The resulting system has been energy minimized. After that, a short (200 ps) position restrained MD simulation of the protein has been performed to let the water molecules and ions equilibrate around it. The position restrained simulation was followed by a 20 ns unrestrained production run.

MD simulations have been performed with the GROMACS software package[250] using the OPLS force field[254]. The integration time step was 2 fs. Periodic boundary conditions have been applied to the box. The simulations have been performed at constant temperature of 300 K (using Berendsen thermostat[275] with coupling constant of τ_T = 0.1ps), and at constant pressure of 1 bar (using Berendsen barostat[275] with an anisotropic coupling constant of τ_P = 1.0 ps and a compressibility = $4.5x10^{-5}bar^{-1}$). Furthermore, the protein and the group of water molecules and ions have been coupled to different thermostats. The

LINCS algorithm[262] has been used to constrain bond lengths. Long-range electrostatics were calculated using the PME method[274] (with a real-space cutoff of 1 nm and a cutoff of 1 nm for the van der Waals interactions). Snapshots from the trajectories have been saved at every 5 ps for analysis.

4.3.2 Dynamic Fingerprint Matrix

In order to obtain a simple representation of the equilibrium fluctuations of proteins, a novel way is introduced here to characterize their backbone dynamics. The underlying idea is that in a moving protein each residue is fluctuating with regards to all other residues, therefore a detailed description of motion should capture all inter-residue fluctuations. While static structures are often characterized by the matrix of inter-residue distances (i.e. the distance matrix), this representation is not applicable for a moving protein in which inter-residue distances are constantly changing. It is possible, however, to characterize the relative motion of any two residues by the distribution of their distance over time. The extent to which residue i and residue j are fluctuating relative to each other is measured by

$$DFM_{ij} := \sigma\{d_{ij}(t)\} \tag{4.1}$$

where DFM_{ij} is calculated as the standard deviation of the distribution of d_{ij} (distance of the two residues) in the whole conformational ensemble generated by MD simulation. In this initial investigation only distances between $C\alpha$-atoms are considered, but the technique can easily be extended to a more detailed description of each amino acid residue. The standard deviation of the distance distribution reflects how much the two residues fluctuate relative to each other. DFM_{ij} values are calculated for each residue pair and are collected into a matrix denoted by DFM, which will be referred to as the Dynamic Fingerprint Matrix. Similarly to a distance matrix that characterizes a single conformation, a dynamic fingerprint matrix characterizes an ensemble of conformations.

Comparison of the DFM and the mean distance matrix calculated for the same MD trajectory (see Figure 4.9 as an example) shows that these two matrices contain essentially

different information and therefore provide complementary descriptions of the conformational ensemble. As illustrated by the scatter plot in Figure 4.2 in the example analysis of the PSD-95 PDZ3 domain, there is only medium correlation (0.49) between the entries of the DFM and the mean distance matrix suggesting that the mean pairwise distance of two atoms is a poor predictor of their relative mobility. While there appears to be a strong linear correlation considering only those data points lying between the 0.5-3.5 nm mean distance interval and the 0.0-0.1 nm DFM interval, still a very large number of data points indicate uncorrelated relationship between the two measures. In other words, one can see large variation in the DFM values for a given mean pairwise distance.

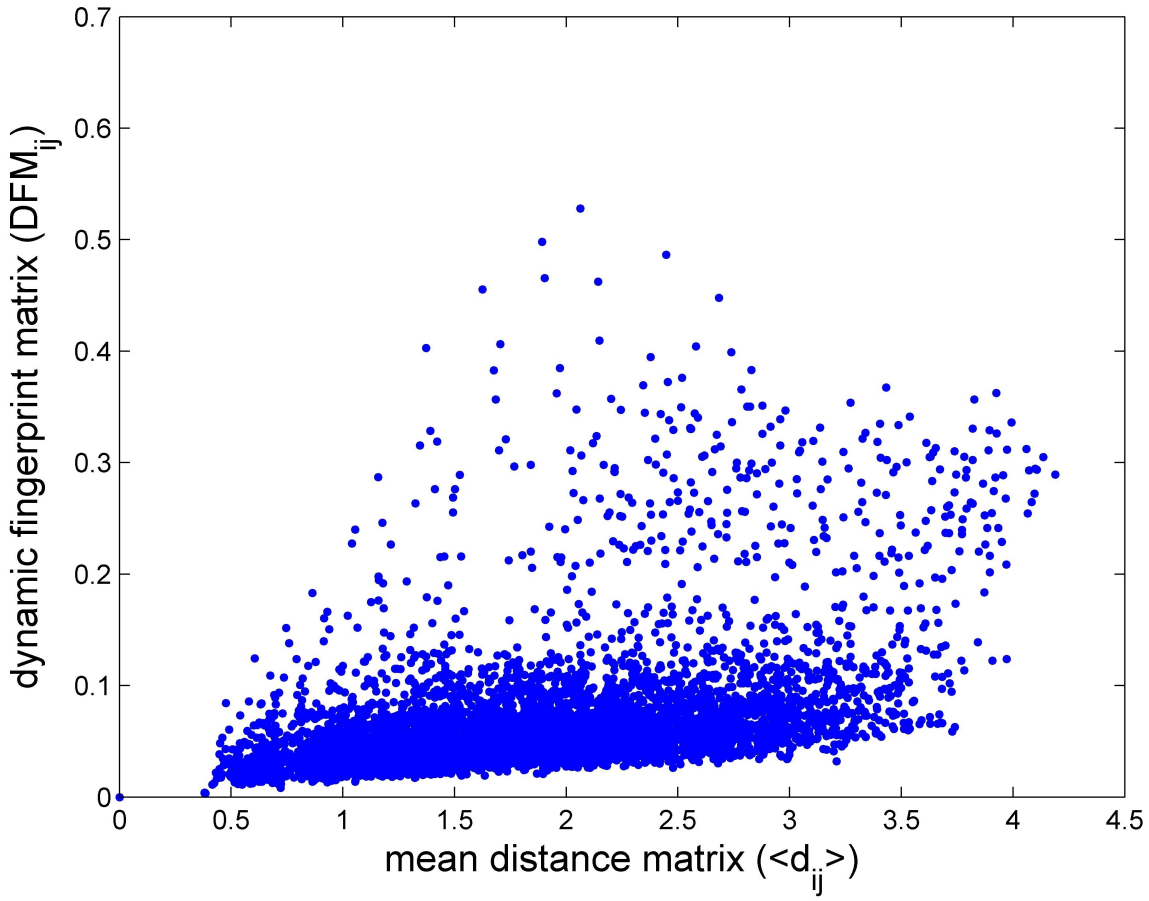

Figure 4.2: Scatter plot showing the relationship between the mean distance and DFM values of the same pairs of atoms calculated based on a 20 ns MD simulation of the PSD-95 PDZ3 domain.

4.3.3 Comparing DFMs using prior alignment

If a prior alignment of the proteins to be compared is known, the comparison of DFMs is straightforward. Given a pairwise sequence alignment of protein A and B, let α and β be the index vectors of the aligned residues of sequence A and B, respectively. That is, the k^{th} match column, (i.e. columns not containing a gap in the alignment) aligns residue α(k) of protein A with residue β(k) of protein B. Thus each pairwise alignment can be characterized by an (α,β) pair. Let DFM^A and DFM^B be the DFMs of the two proteins. The (α,β) alignment define a submatrix of size of $|\alpha|x|\alpha|$ of both DFMs. The (i, j) entries of the two submatrices are given by

$$DFM^A_{\alpha(i)\alpha(j)} \;\; and \;\; DFM^B_{\beta(i)\beta(j)} \tag{4.2}$$

Note that the two submatrices describe the pairwise fluctuations of the aligned residues only. The (i, j) entries of the two submatrices are considered equivalent as they describe the fluctuations of equivalent pairs of residues. (A simple explanation of the relationship between a pairwise sequence alignment and matrix alignment is shown in Figure 4.3.)

The dynamic similarity score of protein A and B based on a prior (α,β) alignment is defined as:

$$S^{AB}(\alpha,\beta) = \sum_{i=1}^{|\alpha|} \sum_{j=i+1}^{|\alpha|} s(i,j) \tag{4.3}$$

where each pair of equivalent matrix entries are compared one-by-one and their contribution to the overall score is given by

$$s(i,j) = \frac{s_+ - s_-}{1 + e^{\lambda(\Delta(i,j)-\omega)}} + s_- \tag{4.4}$$

where

$$\Delta(i,j) = \frac{|DFM^A_{\alpha(i)\alpha(j)} - DFM^B_{\beta(i)\beta(j)}|}{(DFM^A_{\alpha(i)\alpha(j)} + DFM^B_{\beta(i)\beta(j)})/2} \tag{4.5}$$

is the relative difference of the two equivalent matrix entries.

In Eq.4.4, s(i, j) is an S-shaped logistic function that assigns positive score (s_+) to highly

similar matrix entries and negative score ($s\_$) to highly dissimilar entries. An user-adjustable cut-off parameter, T, defines the critical $\Delta(i, j)$ over which s(i, j) turns negative. The relationship between T and parameter ω of s(i, j) is discussed in subsection 4.3.7 as well as the choice of parameter values. The key difference from using a discrete threshold is that the parameter λ can be tuned to set the steepness of the S-shaped function to make s(i, j) less dependent on the cut-off parameter T. Since s(i, j) is associated with a match column pair in the alignment, it will be referred to as the Pairwise Match Score (PMS) of columns i and j and is discussed in more details in subsection 4.3.7.

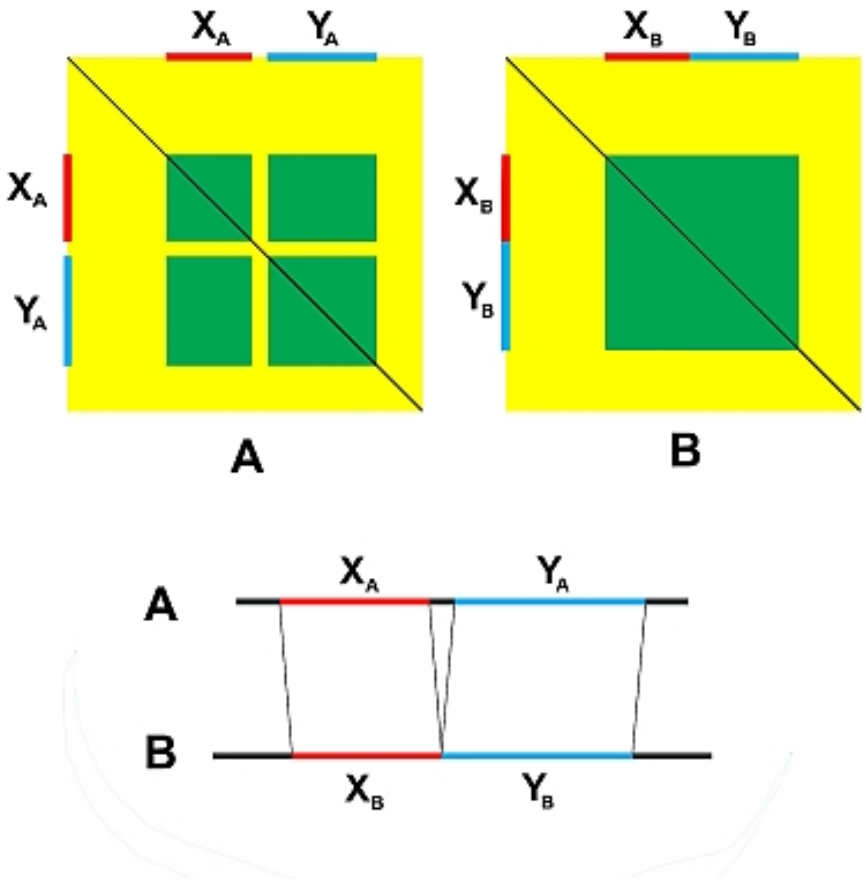

Figure 4.3: Correspondence between sequence alignment and matrix alignment of two proteins, A and B. Sequence region X_A in protein A is aligned to sequence region X_B in protein B and region Y_A in protein A is aligned to region Y_B in protein B. The two aligned segments correspond to two DFM submatrices of equal size. Because the sequence alignment contains a gap in protein B between X_B and Y_B, the submatrix of protein A is composed of separate regions. The DFM algorithm presented below generates local alignments (discussed in Section 2.5.3) of the two proteins.

4.3.4 Comparing DFMs without prior alignment

The way of comparing the DFMs of two proteins using a prior sequence/structural alignment has been introduced above. The main goal of this chapter, however, is to find the optimal alignment of two proteins based on exclusively their DFMs. Note that it is the opposite strategy of most previous comparative MD studies which relied upon prior alignments. The aim is to find the (α,β) pair corresponding to the maximal similarity score. Let (α*, β*) be the pair of index vectors for which $S^{AB}(\alpha, \beta)$ is maximal. $S^{AB}(\alpha^*, \beta^*)$ is then called the dynamic similarity score of protein A and B and is simply denoted by S^{AB}. The sequence alignment problem is hereby transformed into a matrix alignment problem. Structural alignment algorithms DALI[330] and MatAlign[331] aim to solve the same question when aligning distance matrices. The search space of (α,β) pairs is exponentially large and the global optimization problem is in fact NP-hard. In this case to find the maximum score S^{AB} a Simulated Annealing protocol was employed (see details of the algorithm below).

4.3.5 Matrix Alignment Algorithm

To find a good approximation for the global maximum of $S^{AB}(\alpha, \beta)$ in a reasonable time, a heuristics approach have been developed based on the multiple restart Simulated Annealing (SA) algorithm using the MCMC (Markov chain Monte Carlo) method described in details in Section 2.4. The algorithm performs an MCMC optimization search in the space of (α, β) pairs. The Markov chain starts from a random initial alignment, and in each step the alignment is modified by inserting or removing one residue pair. The Metropolis acceptance criterion[287,288] was used to decide the next state of the chain.

The parameter called "temperature" which controls the acceptance probability is gradually reduced according to an exponential decay annealing schedule, leading to the convergence to a high-scoring and potentially optimal alignment. The Markov chain was let to explore the search space at a given constant temperature: the chain has to go through a minimal number of accepted steps before the temperature is further reduced.

The initial temperature is calibrated using the method proposed by Johnson et al.[332].

The whole SA procedure terminates when the acceptance ratio goes below a critical value. To overcome the stochastic nature of SA and the possible existence of local optima, the process is restarted for a number of times from random initial states and the best result of the multiple runs is selected as the final output of the algorithm.

Main steps of the search algorithm

As discussed in Section 4.3.4, the goal of the algorithm is to find the (α*, β*) index vector pair for which S^{AB}(α*, β*) is the global maximum of the S^{AB}(α, β) similarity function. For convenience, the variable of x=(α, β) is introduced to represent a single element of the search space. Note that although a global maximization problem is addressed, the algorithm presented here aims the global minimization of the E(x)= -$S^{AB}(x)$ function, which is equivalent to the original problem. E(x) will be referred to as the energy of state x. The inputs of the algorithm are the two DFMs, while the output is the x*=(α*, β*) best alignment encountered. The main steps of the algorithm are presented in the flowchart in Figure 4.4.

The algorithm uses a Markov chain Monte Carlo (MCMC) method to perform stochastic optimization. In **Step 1**, the initial state of the Markov chain is set to a randomly generated $x_0 = (\alpha_0, \beta_0)$ alignment.

In the Simulated Annealing (SA) heuristics, the temperature (T) of the Markov chain is gradually reduced during the search. The SA algorithm is governed by two nested loops. In the inner loop (from **Step 8** to **Step 19**), a Markov chain of constant temperature is being created. A proposal function (see in the next section) is used in **Step 9** to randomly modify the current state of the Markov chain. The probability of accepting the modified state (x') as the next state of the chain is calculated in **Step 10** according to the Metropolis acceptance criterion[287]:

$$p := Metrop(x_i, x') = min\{1, e^{\frac{1}{T_j}(E(x_i)-E(x'))}\} \tag{4.6}$$

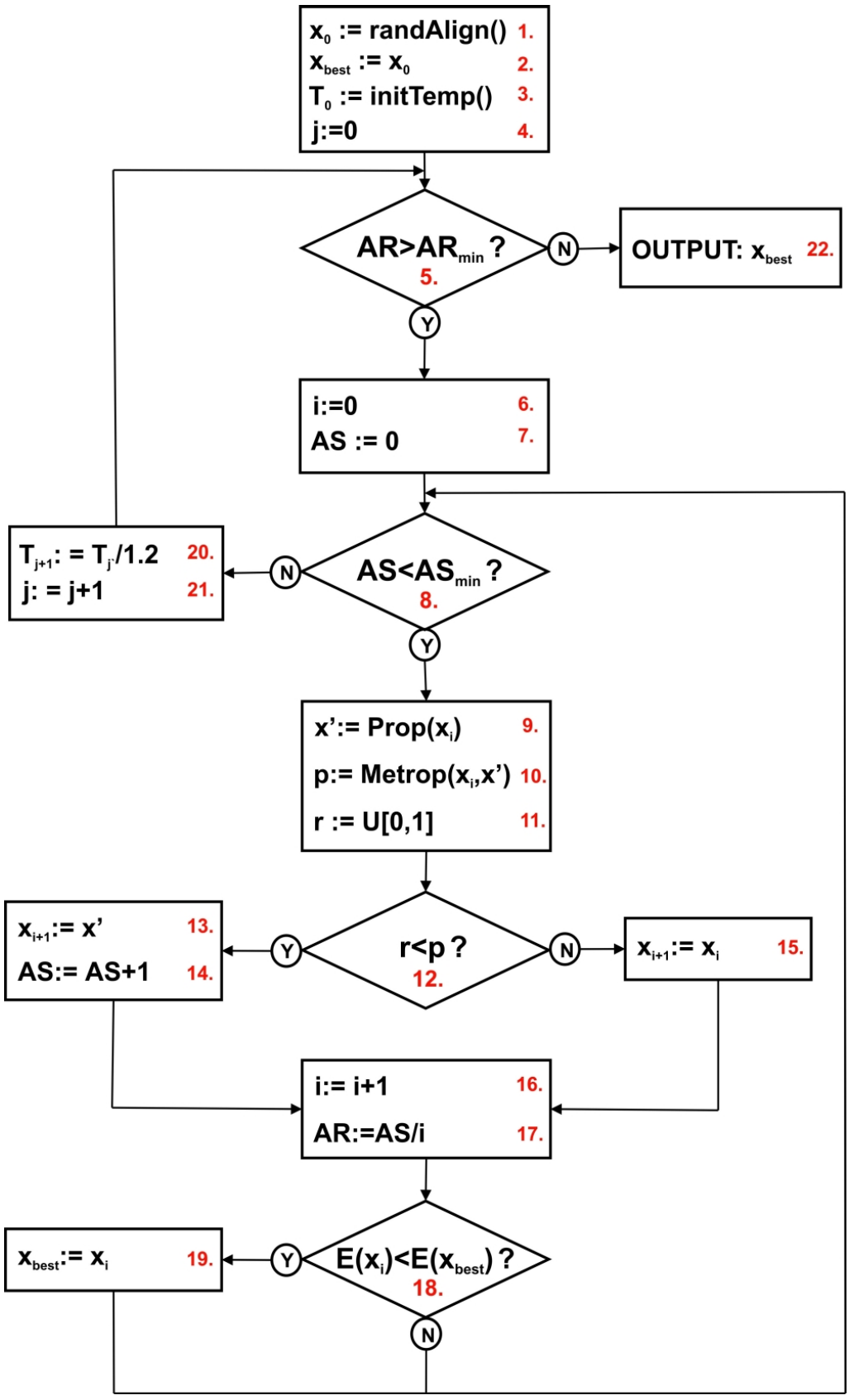

Figure 4.4: Flowchart of the matrix alignment algorithm based on the Simulated Annealing (SA) method. (See detailed explanation of each step in the main text.)

If a random number (r) drawn from uniform distribution (**Step 11**) is lower than the calculated probability (p), the next state of the Markov chain will be the modified state, otherwise it will be the previous state (**Step 12** to **Step 15**). The best state (with the lowest E) found so far is saved to x_{best}, which is initialized in **Step 2**. As the Markov chain grows, it is always checked in **Step 18** if the current state of the chain is better than the saved x_{best} state. If so, x_{best} is updated (**Step 19**).

The number of accepted steps (in which x_{i+1}:=x') is denoted as AS and is counted by **Step 14**. The inner loop does not terminate until the number of accepted steps exceeds a predefined threshold (AS_{min}), ensuring that the chain appropriately explores the accessible states (**Step 8**).

Note, that in the Metropolis criterion, the probability of accepting less-optimal states depends on the temperature of the Markov chain. When the inner loop terminates, the temperature parameter is reduced according to an exponential decay annealing schedule (**Step 20**). The whole process then restarts: a Markov chain of the new constant temperature is being generated in the inner loop.

In other words, the outer loop (**Step 5** to **Step 21**) controls the annealing process in which the temperatures of the consecutive Markov chains are getting lower and lower. The outer loop terminates (**Step 5**) if the acceptance ratio (denoted as AR) calculated in **Step 18** goes below pre-defined a cutoff (AR_{min}). The acceptance ratio is calculated as the number of accepted steps (AS) divided by the total number of steps (i). Note that both counters are set to zero when the inner loop is restarted (**Step 6** and **Step 7**). Low acceptance ratio indicates that the Markov chain is trapped into one or few states due to the low temperature.

The initial value of the temperature parameter (T_0) is calculated in **Step 3** using the method proposed by Johnson et al.[332]. In this strategy, a short trial MCMC run is performed and the optimal initial temperature is estimated from the average energy- (similarity score-) differences. If the outer loop terminates, the algorithm ends (**Step 22**) and gives the output of the best state found during the search (x_{best}).

The proposal function

As shown in the previous subsection, the role of the proposal function, Prop(x) is to introduce a random modification into the current state of the Markov chain. Let the current state of the chain be x=(α, β) and let x'=(α', β') be the modified state, where x'=Prop(x). The proposal function simply adds or removes an index pair to or from the α, β index vectors. In case a new index pair (i_A and i_B) is added, the resulting index vectors are given by

$$\alpha' = \alpha \cup \{i_A\} \quad and \quad \beta' = \beta \cup \{i_B\} \tag{4.7}$$

In case an existing index pair (i_A and i_B) is removed, the modified index vectors are

$$\alpha' = \alpha \backslash \{i_A\} \quad and \quad \beta' = \beta \backslash \{i_B\} \tag{4.8}$$

In each step the probability of adding a new index pair is 0.6, while the probability of removing an old index pair is 0.4. The proposal function first decides whether it adds or removes an index pair and then it randomly selects the index pair to be added or removed.

While it is straightforward to select a removable index pair, selecting an insertable index pair is not trivial. To understand this, let (i_A^1,i_B^1), (i_A^2,i_B^2) and (i_A^3,i_B^3) be three aligned index pairs for which the order is $i_A^1 < i_A^2 < i_A^3$ and $i_B^1 < i_B^2 < i_B^3$. Let a randomly generated index pair be denoted by (i_A^*,i_B^*). In case $i_A^1 < i_A^* < i_A^2$ and $i_B^1 < i_B^* < i_B^2$, the (i_A^*,i_B^*) pair can be inserted into the alignment. Similarly, if $i_A^2 < i_A^* < i_A^3$ and $i_B^2 < i_B^* < i_B^3$, the (i_A^*,i_B^*) pair is also insertable. However, inserting the (i_A^*,i_B^*) pair would cause a conflict in the order of aligned residues in case $i_A^1 < i_A^* < i_A^2$ and $i_B^2 < i_B^* < i_B^3$. Consequently, not all randomly generated index pairs are insertable into the current state x. The proposal function therefore selects randomly only from the subset of insertable index pairs.

Random restarts

The method of Simulated Annealing does not guarantee to find the exact (global) maximum of the scoring function. In some cases, at low temperature the Markov chain is

getting trapped by a local maximum. The more local maxima the scoring function has, the larger the chance for this quasi-ergodic behaviour which is also called the freezing problem. Because of the stochastic nature of the algorithm, its output is not always the same. Even if in some cases the algorithm succeeds to find the global maximum of the scoring function, in other cases it may fail. Therefore, in order to increase the reliability of the algorithm, the whole Simulated Annealing process can be restarted again and again from different random initial states. The number of restarts is an important parameter which be considered as a trade-off between the reliability and the runtime of the optimization.

Parameters of the algorithm

Besides the number of restarts, there are another two parameters that are able to improve the reliability of the algorithm at the cost of the runtime: AS_{min} and AR_{min}. By increasing the value of AS_{min}, one can help to achieve better convergence in each constant temperature sections of the Markov chain. Since the value of AR_{min} determines the point when the algorithm terminates, it should be low enough to let the whole Simulated Annealing process achieve convergence. Reducing AR_{min} increases the probability that SA will not be terminated too early. Another strategy, however, may be to set the parameters to get relatively short runtime and run multiple restarts. The values of AS_{min}=30000 and AR_{min}=0.05 were used in this study.

Speeding up the algorithm

It is time consuming to re-calculate E(x') each time when the proposal function generates a new x' state. A much better strategy is to calculate only the ΔE(x) = E(x')-E(x) difference resulting from the modification of the previous state x. If a new index pair is added to state x by the proposal function, the energy difference is given by

$$\Delta E = -\sum_{\substack{k=1 \\ k \neq k^*}}^{|\alpha'|} s(k, k^*) \tag{4.9}$$

where α'(k*)=i_A and β'(k*)=i_B are the inserted indices. On the other hand, if an index

pair is removed from state x by the proposal function, the energy difference is given by

$$\Delta E = \sum_{\substack{k=1 \\ k \neq k^*}}^{|\alpha|} s(k, k^*) \tag{4.10}$$

where α(k*)=i_A and β(k*)=i_B are the removed indices.

The energy of the modified state is then calculated as E(x')=E(x)+ΔE(x). This strategy results in a significant speedup of the algorithm.

4.3.6 Significance Analysis

To assess the statistical significance of dynamic similarity scores between PDZ domains, 20 ns MD simulations of 12 evolutionarily and functionally unrelated proteins of different sizes have been performed referred to as the Reference Set (see Table 4.1). Reference proteins have been aligned using the dynamic fingerprint alignment algorithm to measure the background distribution of similarity scores. Using the background distribution, the significance of dynamic similarity of any two proteins can be expressed by the p-value of their similarity score.

The background score distribution, however, may and does depend on the lengths of the two protein sequences aligned (i.e. the sizes of the two DFMs). Therefore, when analysing the significance of dynamic similarity score between two proteins, one should always take into account the lengths of the sequences. The strategy used to measure the length-dependency of the background distribution is described here.

45 different length combinations tested

Let $\Theta(L_A, L_B)$ be the background score distribution resulting from alignments of unrelated proteins of the length of L_A and L_B. The distribution is approximated for a set of different (L_A,L_B) pairs. Nine equally distributed points were selected in the [70,110] protein length interval: 70, 75, 80, 85, 90, 95, 100, 105 and 110. Using these points all different (L_A,L_B) pairs were generated. These are: (70,70), (70,75), (70,80), (70,85), (70,90), (70,95), (70,100), (70,105), (70,110), (75,75), (75,80), (75,85), (75,90), (75,95), (75,100), (75,105), (75,110), (80,80),

Protein	PDB	Length	Source organism	Resol. (Å)
γ subunit of the dissimilatory sulfite reductase (DsrC)	1sau	114	*Archaeoglobus fulgidus*	1.12
S-adenosylmethionine decarboxylase (ch. A)	1tlu	117	*Thermotoga maritima*	1.55
Origin binding domain of large T antigen (ch. A)	2ipr	127	*Simian virus 40*	1.5
YueI protein (ch. A)	2ohw	128	*Bacillus subtilis*	1.4
Lysozyme	1lz1	130	*Homo sapiens*	1.5
Carbohydrate binding module (ch. A)	1uxz	131	*Cellvibrio mixtus*	1.4
Soluble Secreted Antigen MPT53	1lu4	134	*Mycobacterium tuberculosis*	1.12
Hypothetical protein Atu0741	1zhv	134	*Agrobacterium tumefaciens str. c58*	1.5
Endonuclease V	2end	137	*Enterobacteria phage t4*	1.45
BclA protein	2r6q	138	*Bacillus anthracis*	1.43
Cutinase	1agy	197	*Nectria haematococca*	1.15
Antiviral protein DAP-30	1rl0	255	*Dianthus caryophyllu*	1.4

Table 4.1: Reference Set: 12 evolutionarily and functionally unrelated proteins of different sizes used for inferring the background similarity score distribution.

(80,85), (80,90), (80,95), (80,100), (80,105), (80,110), (85,85), (85,90), (85,95), (85,100), (85,105), (85,110), (90,90), (90,95), (90,100), (90,105), (90,110), (95,95), (95,100), (95,105), (95,110), (100,100), (100,105), (100,110), (105,105), (105,110) and (110,110). (Note that only one of (x,y) and (y,x) is included in the list.) The $\Theta(L_A, L_B)$ distributions corresponding to the above listed 45 different length combinations were approximated

Alignment of 66 reference protein pairs

Not only the full DFMs of the 12 reference simulations, but their submatrices of any sizes can be used as inputs of the alignments. For example, if one wants to align a protein of

length L_A and a protein of length L_B, it is possible to take a pair of reference proteins and use $L_A x L_A$ and $L_B x L_B$ submatrices of their DFMs.

The method of selecting a submatrix from an original DFM was the following. An integer (r) was randomly selected from the [0,N-S] interval, where N is the size of the DFM and S is the size of the future submatrix. The (i,j) entry of the submatrix was then defined as the (i+r,j+r) entry of the original DFM. In other words, a random subset of consecutive residues from the protein was selected, and the resulting submatrix was the DFM describing the selected residues only.

As all the 12 reference DFMs are larger than 110x110, each DFMs could be used to extract input matrices for the 45 above-listed $\Theta(L_A, L_B)$ distributions. Since there are 66 different pairs of the 12 reference proteins, 66 alignments were performed for each (L_A,L_B) length combinations, in which the aligned $L_A x L_A$ and $L_B x L_B$ submatrices were taken from the 66 possible pairs of reference DFMs. As a result, each of the 45 $\Theta(L_A, L_B)$ distributions consists of 66 values: the dynamic similarity scores of 66 different alignments.

A total of 2970 alignments

Since the background distribution is studied for 45 different length combinations, and for each combination 66 alignments were performed, the total number of alignments carried out was 45x66=2970. All alignments were performed using the same parameter values (as described in subsection 4.3.5).

Extreme Value Distribution

Since the optimal alignment score of two proteins is the maximum of the scores of their possible alignments, it follows a type I Extreme Value Distribution. To give an example, Figure 4.5 shows the measured Θ(80,80) distribution (histogram) that was generated by the method described above. However, to further-increase the accuracy of the distribution in this single example, not only one but ten independent alignments were performed for each reference protein pairs by repeating the random submatrix selection ten times. It therefore resulted in a distribution of 10x66 = 660 alignment scores. The extreme value distribution

fitted to the 660 data points is also shown in Figure 4.5.

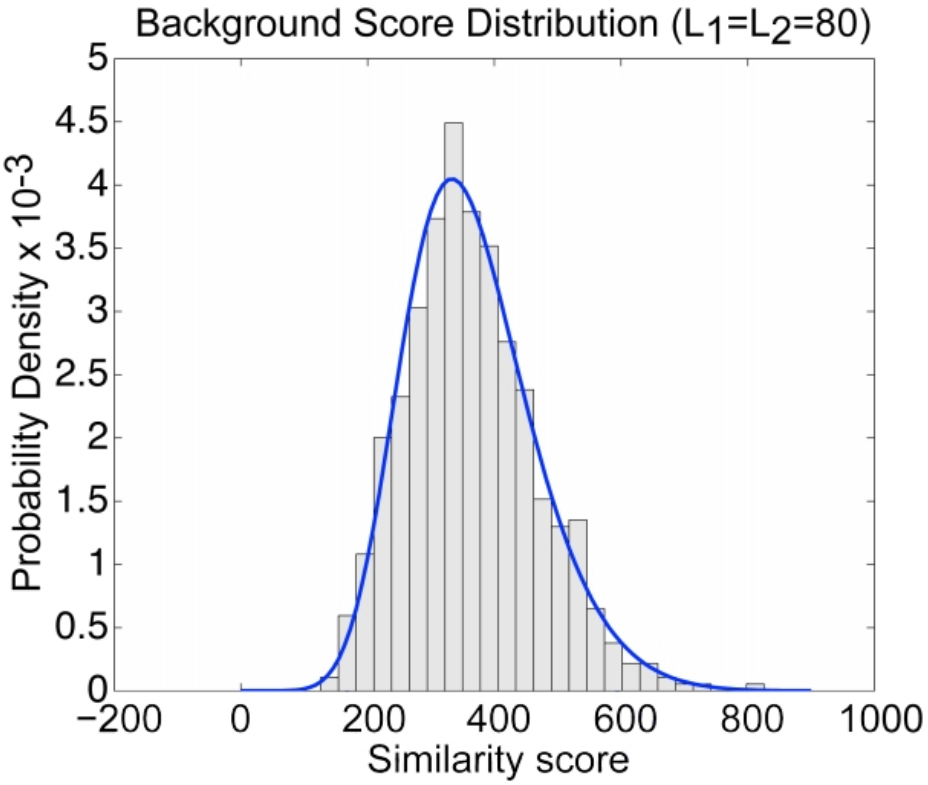

Figure 4.5: Background score histogram (and the type I Extreme Value Distribution fitted on the data) resulting from 660 independent alignments of unrelated proteins of the lengths of 80-residues.

Length-dependency of the parameters

Since type I. Extreme Value Distribution has two parameters: the location parameter (μ) and the scale parameter (σ), only the $\mu(L_A, L_B)$ and $\sigma(L_A, L_B)$ functions were approximated. To simplify the problem, a new variable $L = L_A L_B$ was used calculated as the product of the two sequence lengths. Hence the goal is to approximate the $\mu(L)$ and the $\sigma(L)$ functions. For each measured $\Theta(L_A, L_B)$ distributions, a type I. Extreme Value Distribution was fitted to the 66 data points. The Maximum Likelihood estimates of the location and scale parameters were then plotted against L. As presented in Figure 4.6, linear functions fit well to the 45 data points in $\mu(L)$ and $\sigma(L)$.

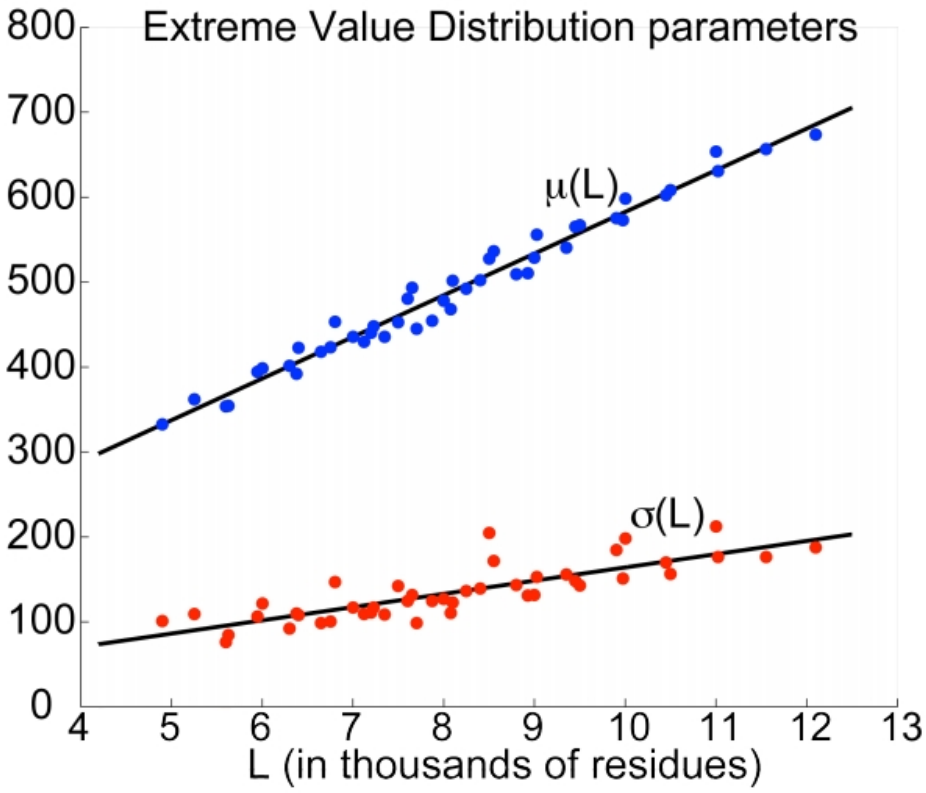

Figure 4.6: Linear dependence of the location and scale parameters (μ and σ) on the product of the lengths of the two proteins (L). Lines were fitted with the Maximum Likelihood method on the points representing the parameter values of the 45 different extreme value distributions.

As it is clear from Figure 4.6, the background score distribution indeed depends on the lengths of the aligned proteins. Both the location (μ) and the scale (σ) parameters of the distribution increase with the size of the aligned DFMs. Consequently, if the length-dependency was not taken into account when calculating the p-values, the statistical significance of similarity could be over-estimated for large protein pairs.

The equations for the best fit lines are given by

$$\mu(L) = 0.049 \cdot L + 92.23 \tag{4.11}$$

$$\sigma(L) = 0.016 \cdot L + 8.18 \tag{4.12}$$

These expressions for $\mu(L)$ and $\sigma(L)$ were used in the Cumulative Distribution Function

(CDF) of type I Extreme Value Distribution to calculate the (one-tailed) p-value of a given S^{AB} alignment score:

$$p(A,B) = exp\left[-exp\left(\frac{S^{AB} - \mu(L)}{\sigma(L)}\right)\right] \tag{4.13}$$

4.3.7 Pairwise Match Score (PMS) and parameter values

As introduced earlier, s(i,j) is referred as the Pairwise Match Score (PMS) of match columns i and j and it is an S-shaped logistic function which has four free parameters, s_+, s_-, λ and ω. The first two parameters define the maximum and minimum of s(i,j). For highly similar matrix entries s(i,j) is close to s_+, while for highly dissimilar entries s(i,j) is close to s_-. Note that changing the ratio of s_+ and s_- has a similar effect to changing the ratio of match and mismatch scores in a standard sequence alignment algorithm. Using high s_-/s_+ ratio, the matrix alignment algorithm can be forced to exclude match columns that align non-similar matrix entries with other match columns.

The parameter ω of the logistic function determines the zero level of the Pairwise Match Score. Let T be defined as a user-adjustable cut-off value determining the point where s(i,j) turns from positive to negative. In other words, if the relative difference (Δ) of two matrix entries is smaller than the cut-off parameter T, their PMS increase the total alignment score, otherwise it has a negative contribution:

$$\Delta(i,j) \leq T \rightarrow s(i,j) \geq 0 \quad and \quad \Delta(i,j) > T \rightarrow s(i,j) < 0 \tag{4.14}$$

From Eq.4.4, we get that the transformation between the user-adjustable parameter T and parameter ω of s(i,j) is given by

$$\omega = T - \frac{1}{\lambda} ln\left(-\frac{s_+}{s_-}\right) \tag{4.15}$$

The fourth parameter, λ, can be used to set the steepness of the logistic function (the transition between s_+ and s_-) to make s(i,j) less dependent on the choice of the cut-off parameter T.

Figure 4.7 shows the S-shaped s(i,j) function with the parameter values used in this

study (s_+=1, s_-=-1, λ=100 and T=0.2). These parameters have been selected empirically by trying to increase the discriminative power of the scoring function as much as possible. That is, the goal was to find parameter values which result in a large difference between similarity scores of PDZ domains and similarity scores of random proteins.

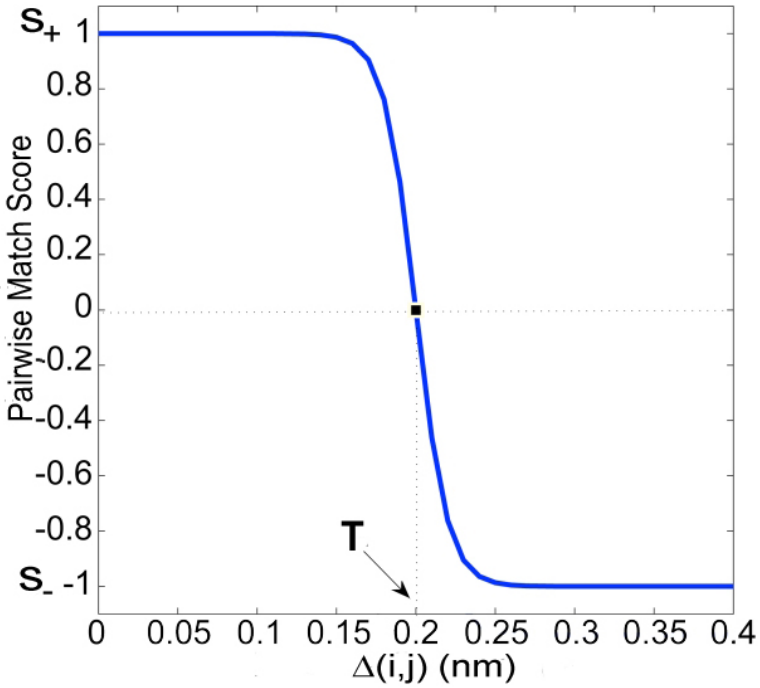

Figure 4.7: Example of the S-shaped logistic function describing the dependence of the pairwise match score (PMS) on the relative difference of the two compared matrix entries. (The parameter values of the curve are given in the main text.)

For example, if the value of the cut-off parameter T is too small, it may give so strict requirement on the similarity of DFM entries that is rarely satisfied regardless we are aligning PDZ domains or random proteins. On the other hand, too large values of parameter T result in equally good scoring alignments of random proteins and PDZ domains. The choice of T=0.2, which proved to be reasonably good, means that two equivalent DFM matrix entries can differ by a maximum of 20 per cent (relative to their mean) in order to give a positive contribution to the alignment score.

The rationality behind using s_+=1, s_-=-1 was to let the highly similar and highly dissimilar matrix entry pairs have the same weight in the total alignment score. It is analo-

gous to the case when matches and mismatches are equally weighted in standard sequence alignment algorithms. When the s_-/s_+ ratio is increased, the resulting dynamics-based alignments typically contain less match columns, but the number of columns corresponding to negative PMS values is also reduced.

4.3.8 Single Match Score (SMS)

Although the PMS scores corresponding to a given match column depend on the other match columns in the alignment, it is useful to compare the total contributions of each individual column to the alignment score. For match column i, the sum of PMS scores with respect to all other match columns will be referred to as the Single Match Score (SMS):

$$s(i) = \sum_{\substack{j=1 \\ j \neq i}}^{|\alpha|} s(i,j) \tag{4.16}$$

In other words, the SMS of a match column is the score by which the total alignment score decreases in case of removing that match from the alignment. Matches of negative SMS values are beneficial to remove in optimizing the alignment. Hence the optimal dynamics-based alignment contains only positions of non-negative SMS values. Either studying a prior (sequence/structural) or a dynamics-based alignment, the SMS-profile represents our confidence in each aligned pairs of residues.

4.4 Results

4.4.1 Analysis of the motion of PSD-95 PDZ3

Before the comparative analysis of PDZ domains is discussed, it is first demonstrated that the DFM protocol is appropriate by characterizing the dynamics of the third PDZ domain (PDZ3) of PSD-95 (Postsynaptic Density Protein 95) from Rattus norvegicus. PSD-95 plays an important role in controlling synaptic strength and plasticity in the central nervous system.[333] The 110-residue-long PDZ3 is one of the most well studied PDZ domains[198] with a canonical PDZ-domain structure which (as it is described in Section 1.5.3) consists

of six β-strands (β1-β6) and two α-helices (α1 and α2). The peptide-binding groove is located between the β2-strand and β2-helix (see Figure 4.8). As described in Methods, a 20 ns MD trajectory was used to calculate the Dynamic Fingerprint Matrix (DFM) of PDZ3 (see Figure 4.9B).

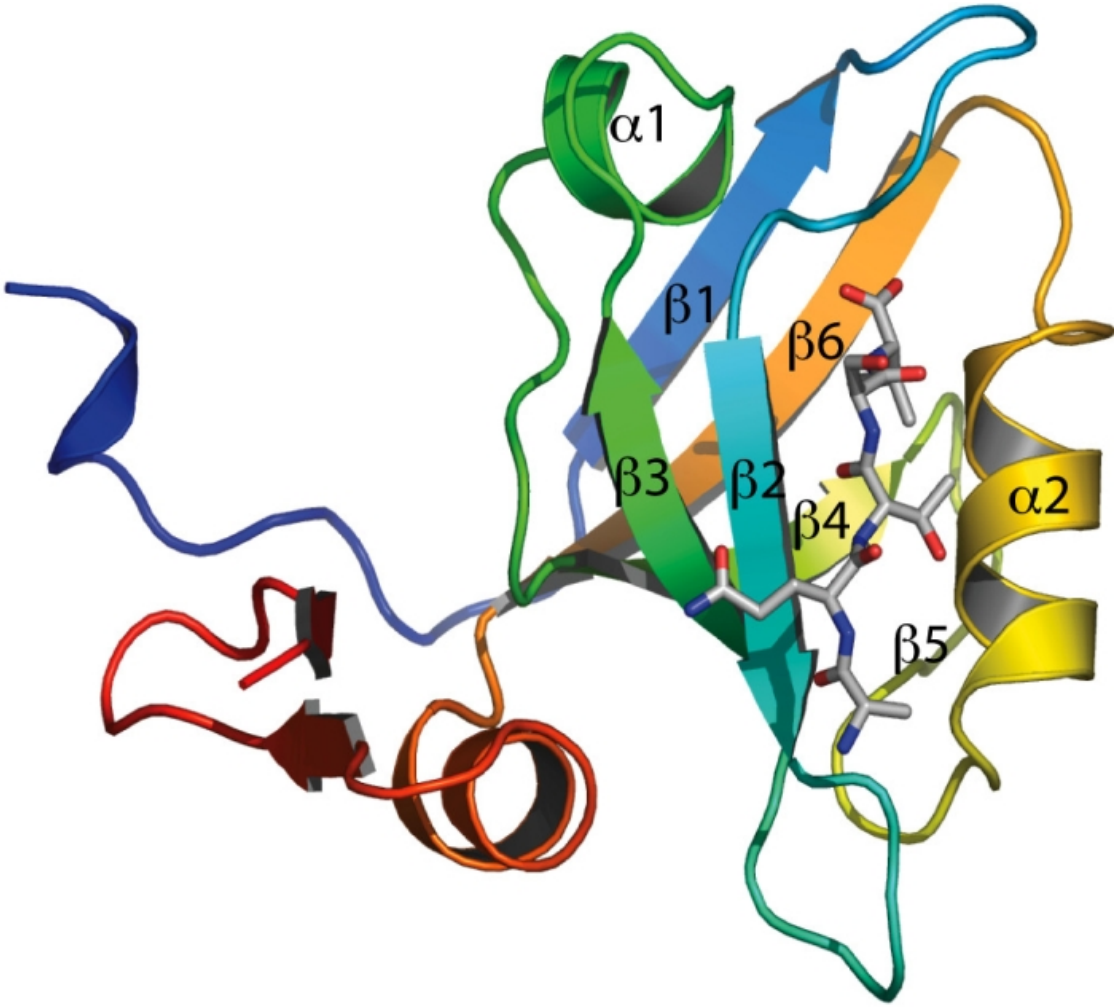

Figure 4.8: The third PDZ domain (PDZ3) of Postsynaptic Density Protein-95 (PSD-95) in complex with the C-terminal peptide of CRIPT. (PDB code: 1BE9)

Simple analysis of the DFM revealed that the most dynamic part of the domain is the α2-helix (His372-Ala382); this region has the highest fluctuation with regards to the rest of the protein. This observation is in accordance with De Los Rios et al.[334] who performed Normal Mode Analysis of a Gaussian Network Model of PDZ3. As the binding pocket

is formed between the α2-helix and the β2-strand, it seems likely that the considerable relative motion of these two structural components may be related to the ligand binding capacity of the PDZ domain. Such relationship between the relative fluctuation of β2-strand vs. α2-helix and the ability to bind multiple peptides is discussed in details in Chapter 5 in which a direct link is found between the binding site flexibility and ligand binding promiscuity of PDZ domains.

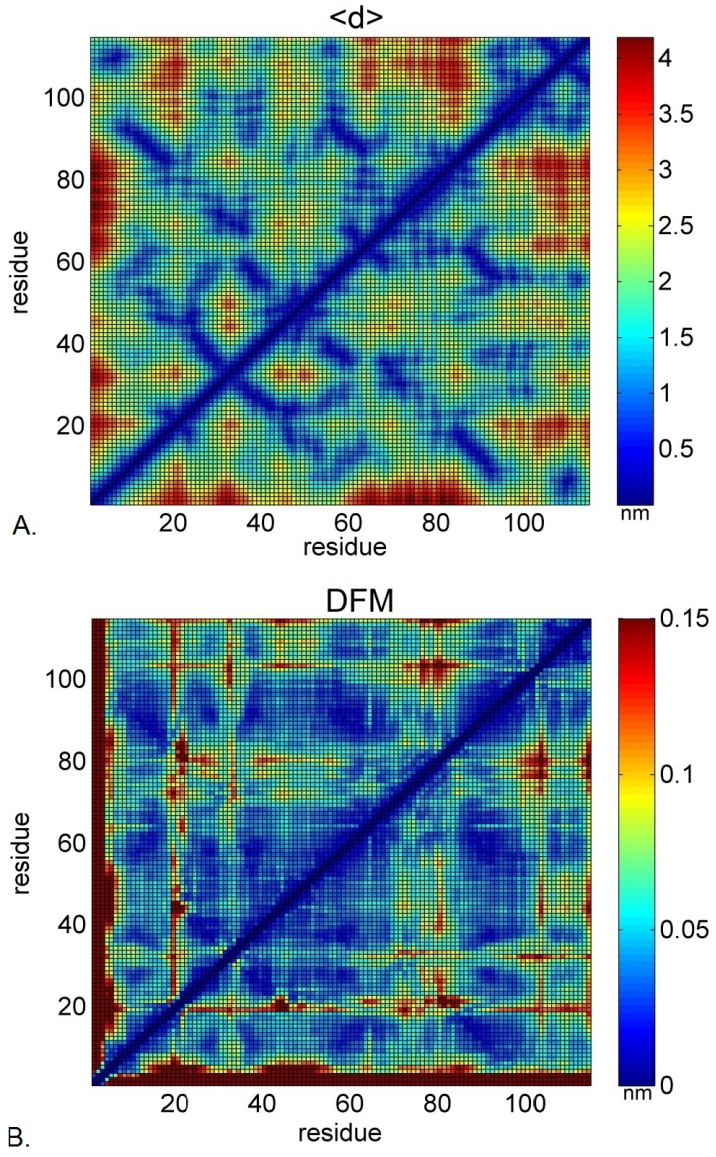

Figure 4.9: Mean distance matrix (**A.**) and dynamic fingerprint matrix (DFM) (**B.**) of PSD-95 PDZ3. In the DFM, regions where the distance fluctuation gives a high standard deviation, σ, are indicated as red while low σ values are indicated as blue.

To reveal how a particular residue fluctuates relative to all other residues, individual rows of the DFM were examined. The k^{th} row of the DFM will be referred to as the "dynamic profile" of residue k. The dynamic profile of Phe325, for example, shows us that it is the $\beta 1/\beta 2$ loop and the $\alpha 2$-helix which fluctuates the most relative to this residue (see Figure 4.10). Note, that Phe325 is located at the N-terminal end of the $\beta 2$-strand, right next to the $\beta 1/\beta 2$ (L1) loop that interacts with the carboxylate terminal of the bound peptide. Therefore, the relative motion of Phe325 and the $\alpha 2$-helix at the other side of the binding pocket along with the L1 loop may be responsible for structural deformations that are necessary for peptide binding. A series of examples are shown in Chapter 5 in which the intrinsic fluctuations of the PDZ binding pocket are closely related to its peptide binding properties.

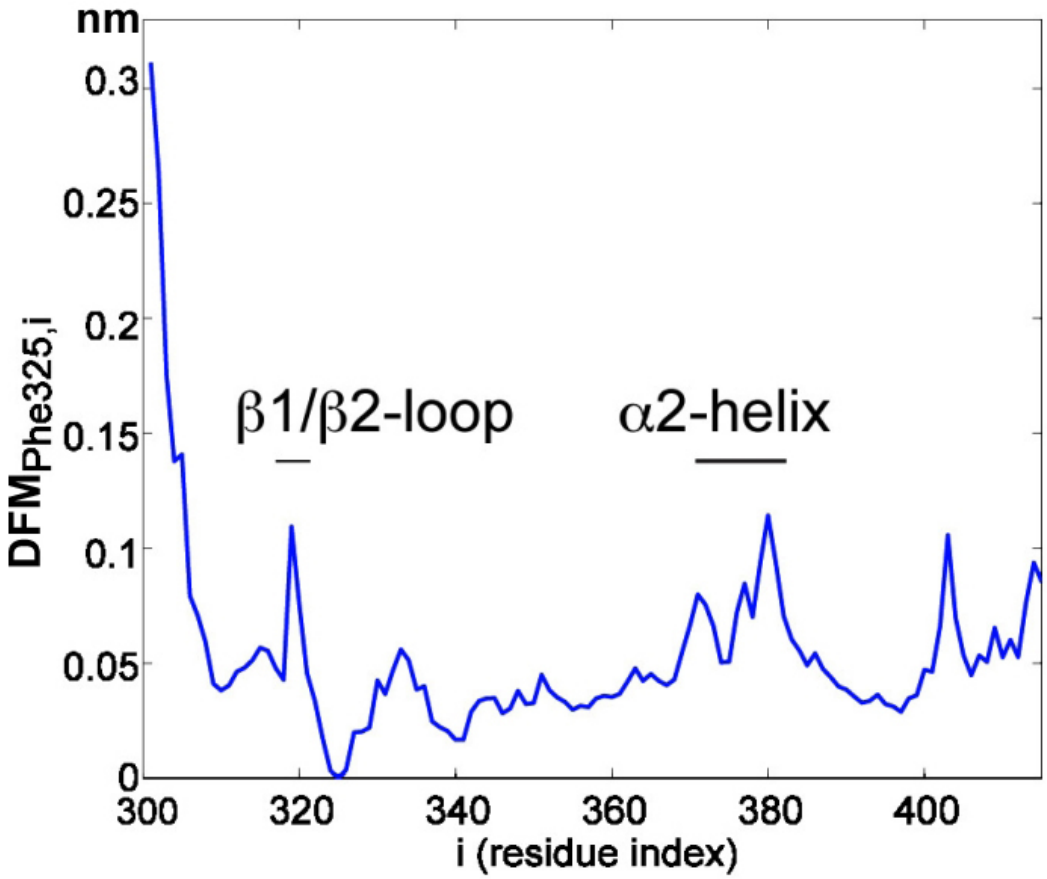

Figure 4.10: Example dynamic profile calculated for residue 25 (Phe325 of PSD-95 PDZ3). The sequence regions of the $\beta 1/\beta 2$-loop and $\alpha 2$-helix are highlighted.

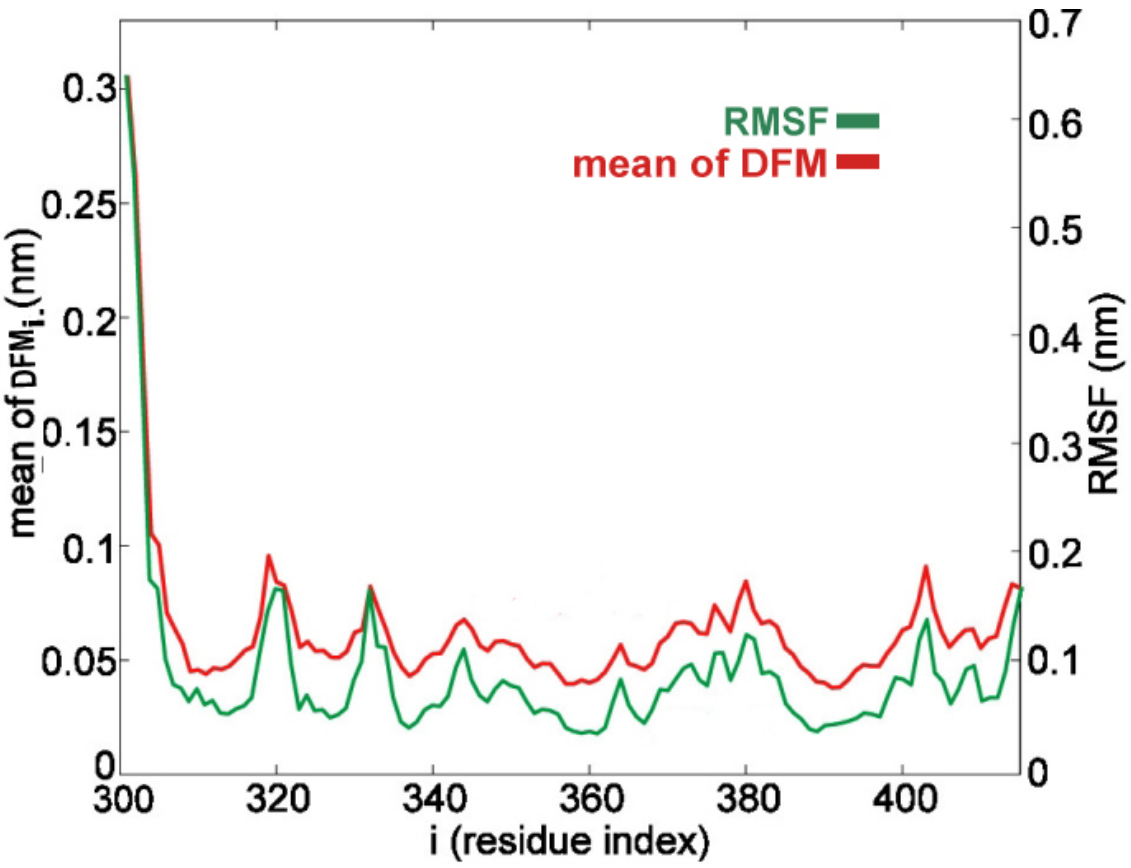

Figure 4.11: Average fluctuation profile (AFP) (mean value of each row of the DFM matrix) compared to the RMSF (root mean square fluctuation) profile of PSD-95 PDZ3.

The mean value of each row of the DFM have also been calculated and will be referred to as the *average fluctuation profile* (AFP) (see Figure 4.11). Atomic fluctuations are often characterized by the RMSF (root mean square fluctuation). As shown by Figure 4.11, the RMSF profile is very similar to the AFP. There was 0.94 correlation between the average fluctuation profile and the RMSF profile. It can be concluded that a DFM contains the same information as a standard RMSF plot, however, by describing pairwise relative inter-residue fluctuations, it gives us a more detailed representation of protein flexibility. More importantly, it does away with the dependency on a single "native" reference structure for defining fluctuations and is simple to compute. As shown by several studies, the standard least square superposition method that involves rigid transformations to overlay alternative conformations (see Section 3.3.3) has several weak points, particularly when it is used

for analysing flexible proteins (e.g. those that contain highly mobile loops or hinged domains).[335] Relying on the rigid superposition process, the RMSF measure is also biased by the arbitrary choice of atoms used for the least square RMSD calculations. Therefore it is proposed here, that the average fluctuation profile calculated from the DFM would generally serve as a more reliable measure for residue fluctuations than RMSF since the former does not depend on superposition step. The benefits of the AFP over RMSF is probably most prominent in the case of highly flexible proteins.

4.4.2 Dynamics-based alignments of PDZ domains

Thus far it has been demonstrated that the DFM methodology can be a useful way to analyse protein motions, but the power of the approach is that it enables us to compare the dynamics of two or more different proteins. Furthermore this information can be used to derive an alignment. To illustrate this, 10 PDZ domains have been selected from a range of organisms (see Table 4.2) and 20 ns explicit MD simulations of these proteins have been performed. The DFM for each protein were then calculated and the dynamics-based alignments of each pairs of proteins have been built using the matrix alignment algorithm described in Methods.

Figure 4.12 presents an example: the alignment of the third PDZ domain (PDZ3) of PSD-95 and the PDZ domain of neuronal nitric oxide synthase (nNOS). The alignment process does not require any prior sequence or structural information; the two DFMs are the only inputs of the algorithm (Figure 4.12A). The most similar submatrix pair found by the algorithm have 77x77 entries which are highlighted (in white) in Figure 4.12B. The optimal submatrix pair corresponds to a pairwise alignment consisting of 77 aligned residues. Removing the rows and columns of the DFMs that correspond to gaps in the alignment, the remaining matrices will be referred to as the "collapsed DFMs" (which are identical to the submatrices identified). Although one cannot see notable similarity between the original DFMs, the collapsed DFMs appear to be visually similar patterns (Figure 4.12C).

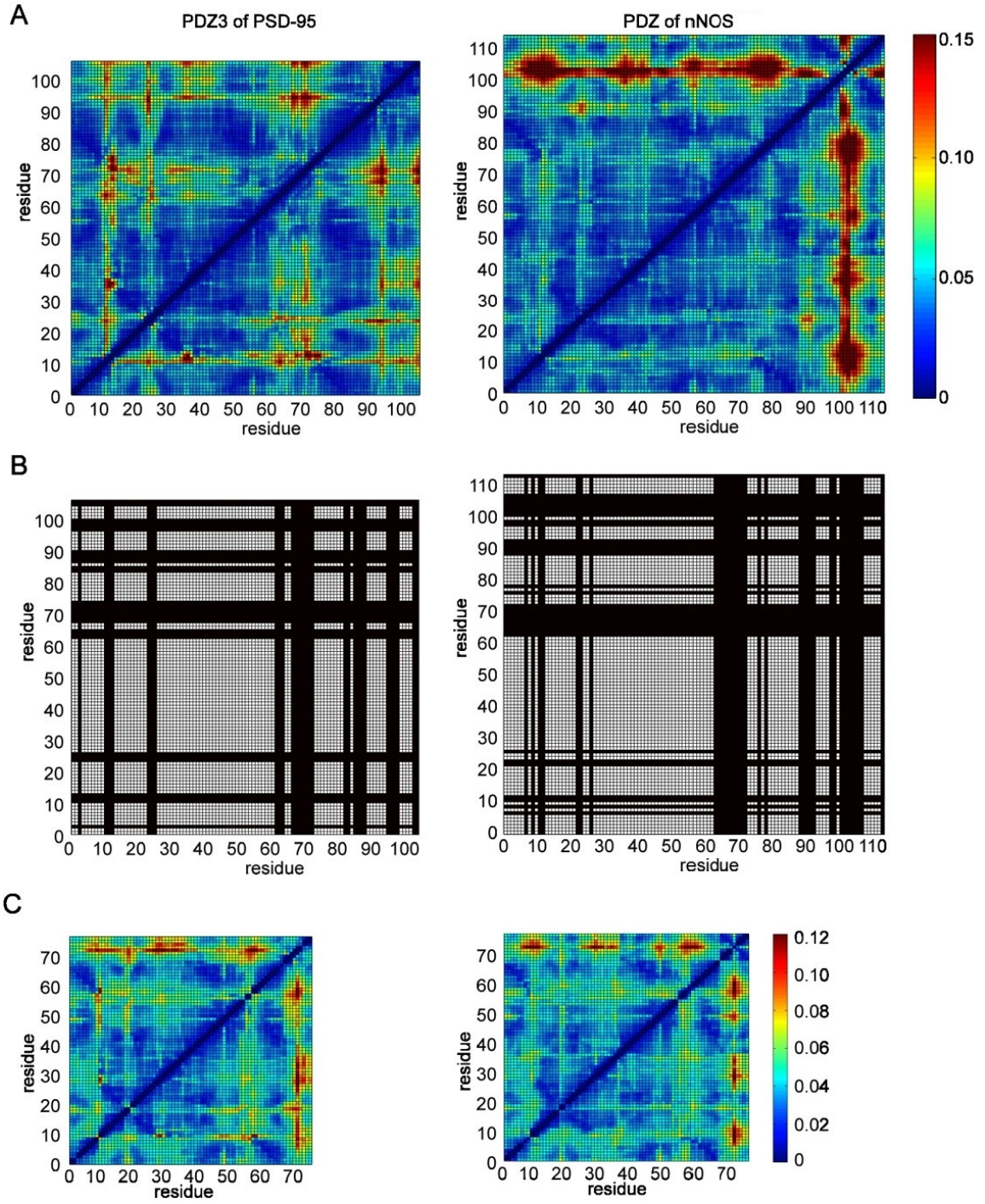

Figure 4.12: **A:** Comparison of the DFMs calculated for the third PDZ domain of PSD-95 and the PDZ domain of nNOS. **B:** Identification of similar submatrices, containing 77 residues in this case, from the DFMs. Aligned matrix entries are highlighted in white. **C:** Collapsed DFMs (77x77 submatrices) representing the set of aligned residues. Unlike the two full DFMs, the two collapsed DFMs are similar patterns.

The derived dynamics-based alignment was compared to a structural alignment created by pairwise DaliLite[330] and a pairwise sequence alignment created by the Needleman-Wunsch algorithm[102] using EMBOSS-Align[336]. Figure 4.13 presents the three alignments of the same pair of proteins annotated by the secondary structure elements of the canonical PDZ-domain fold (i.e. six β-strands, β1 to β6, and two α-helices, α1 and α2). For the DFM-alignment and the sequence alignment, the SMS score of each column is also presented, reflecting our confidence in individual aligned positions.

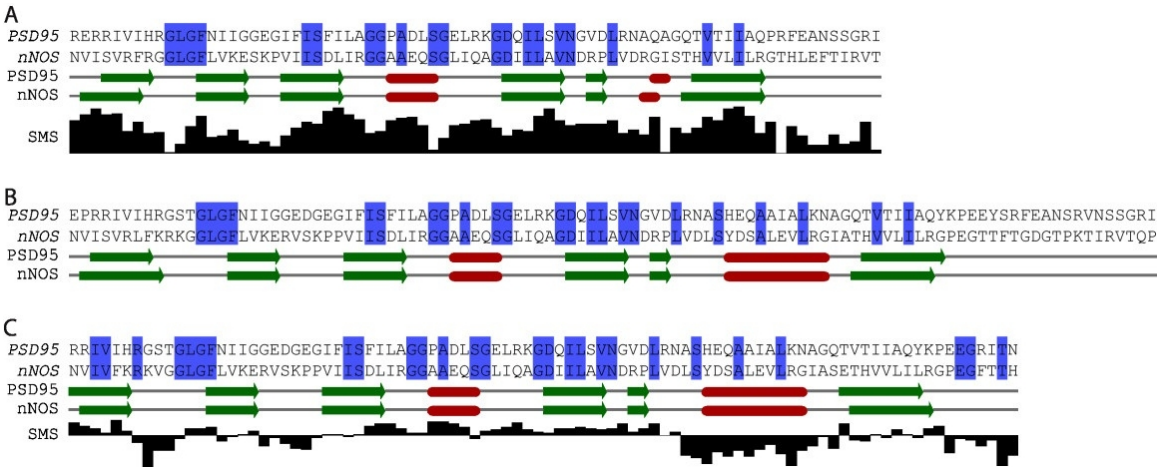

Figure 4.13: The resulting dynamics-based alignment derived from the collapsed DFMs (**A**) is compared to the alignments derived with the DALI (**B**) and Needleman-Wunsch (NW) algorithms (**C**). Identical residue pairs are indicated by blue boxes. The single match score (SMS) is depicted underneath the DFM and NW alignments. It can be seen that the region corresponding to the second α-helix in the NW alignment gives negative SMS values indicating that the dynamic similarity is not preserved in this region.

As shown by Figure 4.13A, equivalent secondary structure elements of the two proteins align very well in the dynamics-based alignment, suggesting that dynamics, just like sequence and structure, may contain enough information to match proteins at the secondary structure level. Moreover, the DFM-based alignment includes 20 pairs of identical residues, out of which 18 and 20 are also present in the Needleman-Wunsch and DALI alignments, respectively. Despite all these similarities, however, a striking difference can be seen between the DFM-based alignment and the sequence/structural alignments. The second α-helix (α2), included both in the Needleman-Wunsch and DALI alignments, is almost completely missing from the DFM-alignment, indicating that, although conserved

at the sequence and structure level, this helix has different dynamics in the two proteins. Characterized above, the α2-helix has high mobility in the PDZ3 of PSD-95 unlike in the PDZ of nNOS, that makes the two regions dynamically non-alignable. This is a clear example, when the dynamics-based alignment gives similar information as sequence and structural alignments, but at the same time, it provides new insights into the properties of proteins, that cannot be detected through standard alignment methods.

PDZ domain	**PDB**	**Source organism**	**Resolution (Å)**
nNOS (neuronal nitric oxide synthase) PDZ	1qau	*Rattus norvegicus*	1.25
InaD (inactivation no afterpotential D) PDZ1	1ihj	*Drosophila melanogaster*	1.8
PSD-95 (postsynaptic density protein 95) PDZ3	1bfe/1be9	*Rattus norvegicus*	2.3/1.82
tricorn protease PDZ	1k32	*Thermoplasma acidophilum*	2.0
GRIP2 (glutamate receptor interacting protein 2) PDZ4	1x5r	*Homo sapiens*	NMR structure
hypothetical protein Rv0983 PDZ	1y8t	*Mycobacterium tuberculosis h37rv*	2.0
photosystem II D1 protease PDZ	1fc6	*Scenedesmus obliquus*	1.8
alpha-1 syntrophin PDZ	1qav	*Mus musculus*	1.9
EpsC (extracellular protein secretion C) PDZ	2i6v	*Vibrio cholerae*	1.63
Dvl2 (Dishevelled 2) PDZ	2f0a	*Xenopus laevis*	1.8

Table 4.2: Ten PDZ domains used as a test set for the dynamics-based alignment algorithm.

4.4.3 Analysis of SMS-profiles

Since it was optimized by the matrix alignment algorithm, the DFM-alignment includes only matches of positive SMS values as shown by Figure 4.13A. The SMS-profile has its peaks within β-strands but drops at certain match columns (e.g. in the β1/β2 and β2/β3

loops and at the C-terminal end of α1-helix). This suggests that β-strands of the domains have minor fluctuations that makes them easier to align than the other regions of the proteins.

As discussed before, the dynamic similarity of proteins can also be measured based on a prior (sequence or structural) alignment. In this case, the motion of the subsets of residues defined by the prior alignment is compared. To test this option, the Needleman-Wunsch alignment was used as a prior alignment, which resulted in a dynamic similarity score of -132.8. The optimal similarity score found for this example is 1307.3, and the extreme non-optimality of the sequence based alignment score illustrates that conserved sequence positions can match dynamically dissimilar subsets of residues. Accordingly, 46 per cent of the columns of the collapsed Needleman-Wunsch alignment have negative SMS. The less-matching region (a continuous block of negative SMS values) appears to be the α2-helix, explaining why this region is excluded from the optimized DFM-alignment (Figure 4.13C).

4.4.4 Distribution of fluctuation values

The dynamics-based alignment of rigid substructures is essentially equivalent to their structural alignment problem as rigid residue pairs give low matrix entries in both DFMs. It is therefore important to see whether a dynamics-based alignment is constructed based on only the similarities of low DFM entries.

To investigate if the residues identified to be dynamically similar are also the residues that tend to exhibit small deviations (little fluctuation), the DFM patterns were studied at a coarse-grained level. The DFM matrix values have been discretized into three values: LOW, MEDIUM and HIGH fluctuation. The thresholds of discretization were selected by analysing the total distribution of matrix entries collected from the 10 different DFMs corresponding to the studied PDZ domains. The thresholds were set using the equal-frequency binning method: i.e. 1/3 of the values in this distribution were assigned to LOW, 1/3 of the values were assigned to MEDIUM and 1/3 of the values were assigned to HIGH fluctuation.

Using this discretization scheme, the frequency of LOW/MEDIUM/HIGH fluctuation values in the collapsed DFMs identified by matrix alignment have been studied. For example, as discussed above, the dynamics-based alignment of PSD-95 PDZ3 and nNOS PDZ results in a 77x77 collapsed DFM. The proportion of LOW, MEDIUM and HIGH fluctuation values in the collapsed DFM patterns is 45 per cent, 40 per cent and 15 per cent, respectively. In other words, a large proportion of the DFM entries included in the pattern indeed correspond to low pairwise fluctuations, but still a significant proportion of values correspond to medium and high fluctuations. Of course, the relative proportion of LOW/MEDIUM/HIGH values may be different in the case of different protein pairs.

4.4.5 Dynamics-space of PDZ domains

The dynamic similarity score of the PDZ domain of nNOS and PDZ3 of PSD-95 is S=1307.3, which was converted to a p-value of $9 \cdot 10^{-11}$ using the significance analysis framework described in Methods. Similarly to this highly significant similarity, the alignment algorithm has found significant dynamic similarities between other pairs of PDZ domains. The p-values are summarized in Table 4.3 and can also be presented as a dynamic similarity graph shown in Figure 4.14A, in which the different PDZ domains are represented by the nodes of the graph, and two PDZ domains are connected if they have significantly similar (p-value < 0.05) dynamics.

Strikingly, the dynamic similarity shows differences between different pairs of structures. We can see a cluster of five proteins (1be9, 1qau, 1qav, 1ihj and 2f0a) that are better-connected in the dynamic similarity graph. Out of the 10 possible links between these structures 8 are present in the graph. Two additional links are found between 2f0a↔1x5r and 1ihj↔1fc6. Three structures (2i6v, 1k32 and 1y8t), however, do not have significant dynamic similarity with any other structures. Looking for structural explanations for these differences, the dynamic similarity data have been compared to the DALI Z-scores between the 10 domains (summarized in Table 4.3 and shown as a DALI Z-score graph in Figure 4.14B). In this second graph, two nodes are connected if their DALI Z-score is more than 8.5 (a threshold selected empirically). As expected, each protein pair has significant

	1qau	**1ihj**	**1be9**	**1k32**	**1x5r**	**1y8t**	**1fc6**	**1qav**	**2i6v**
1ihj	$\underline{4 \cdot 10^{-4}}$ **(13)**								
1be9	$\underline{9 \cdot 10^{-11}}$ **(12)**	0.34 **(12.1)**							
1k32	0.62 **(6.7)**	0.76 **(7.0)**	0.74 **(5.8)**						
1x5r	0.32 **(10.1)**	0.32 **(10.7)**	0.60 **(9.8)**	0.39 **(5.2)**					
1y8t	0.61 **(7.7)**	0.47 **(7.9)**	0.58 **(6.4)**	0.62 **(7.1)**	0.36 **(6.2)**				
1fc6	0.08 **(8.0)**	$\underline{8 \cdot 10^{-3}}$ **(8.7)**	0.78 **(7.4)**	0.76 **(6.6)**	0.65 **(7.2)**	0.19 **(9.4)**			
1qav	$\underline{1 \cdot 10^{-8}}$ **(14.7)**	$\underline{2 \cdot 10^{-3}}$ **(14.0)**	$\underline{1 \cdot 10^{-5}}$ **(15.7)**	0.66 **(6.8)**	0.14 **(10.7)**	0.07 **(7.3)**	0.17 **(7.6)**		
2i6v	0.68 **(4.5)**	0.78 **(4.8)**	0.58 **(3.6)**	0.70 **(6.0)**	0.71 **(4.2)**	0.55 **(4.5)**	0.81 **(4.7)**	0.59 **(4.2)**	
2f0a	$\underline{1 \cdot 10^{-4}}$ **(12.5)**	$\underline{3 \cdot 10^{-6}}$ **(12.4)**	0.27 **(11.7)**	0.30 **(6.8)**	$\underline{8 \cdot 10^{-4}}$ **(9.2)**	0.14 **(7.0)**	0.47 **(7.9)**	$\underline{6 \cdot 10^{-6}}$ **(13.2)**	0.72 **(4.2)**

Table 4.3: Dynamic similarity p-values for the pairwise alignments of the 10 PDZ domains. Significant similarity scores ($p < 0.05$) are underlined. The corresponding DALI Z-scores between the same structures are shown in brackets.

structural similarity (all DALI Z-scores are larger than 3.5), but a subset of structures are more similar to each other than to the others. A cluster of six structures (1be9, 1qau, 1qav, 1ihj, 2f0a and 1x5r) appears to be fully connected in the graph, while two additional links are found between 1fc6↔1y8t and 1ihj↔1fc6. Two domains (1k32 and 2i6v) are not linked to any other structures.

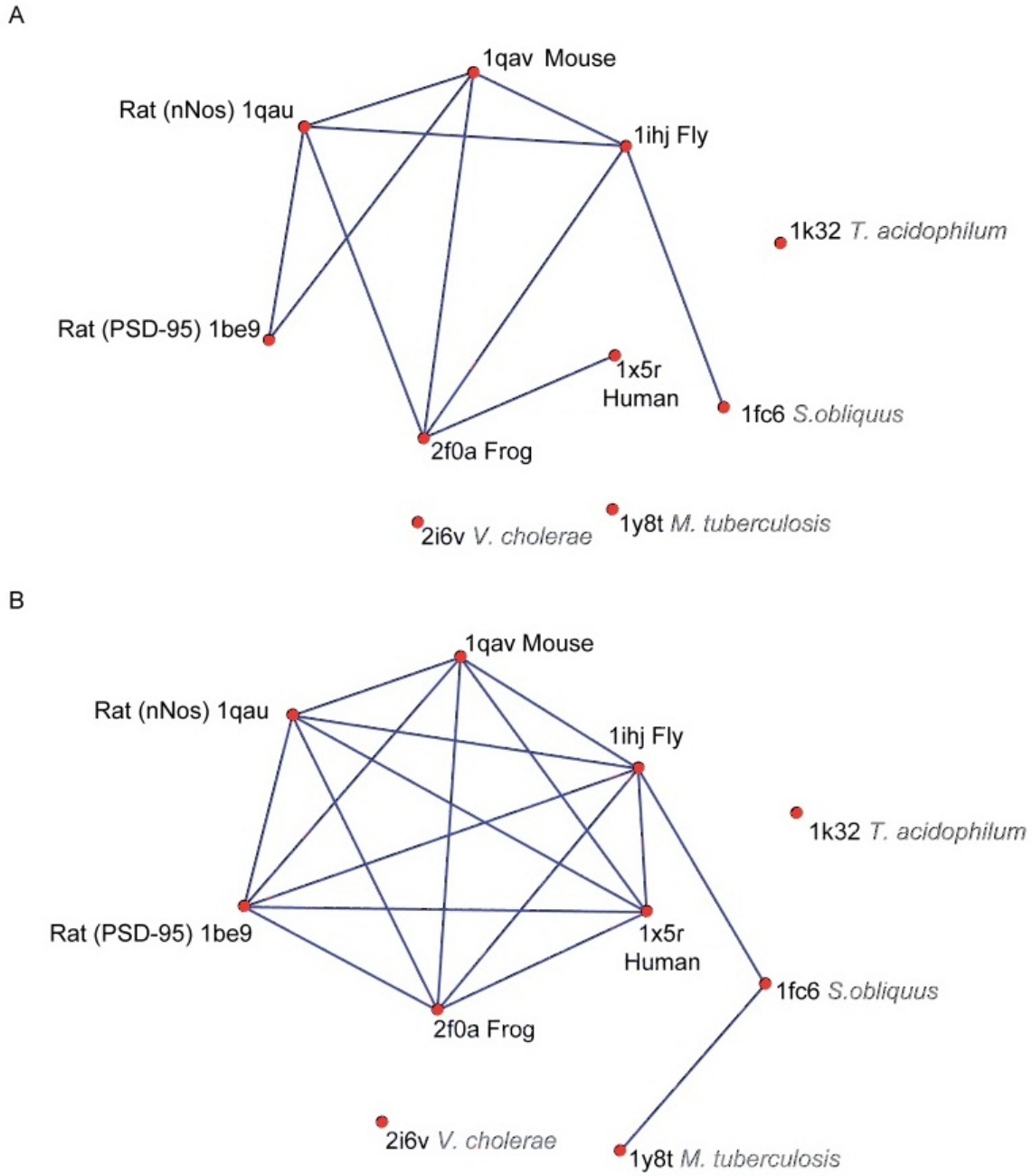

Figure 4.14: **A:** Similarity graph of the 10 PDZ domains derived from DFM alignments. Nodes represent PDZ structures and edges represent $p > 0.05$. **B:** Similarity graph of the same 10 PDZ domains calculated with DALI. Edges represent Z-scores > 8.5.

The almost perfect overlap (with the only exception of 1x5r) between the well-connected clusters in the two graphs suggests a topology-preserving mapping between the structure space and dynamics space of PDZ domains. There appears to be a strong correlation (0.82)

between the raw dynamic similarity scores and DALI Z-scores considering all 45 protein pairs (see Figure 4.15), and a weaker but still strong correlation (0.63) considering only the 35 protein pairs having non-significant dynamic similarities.

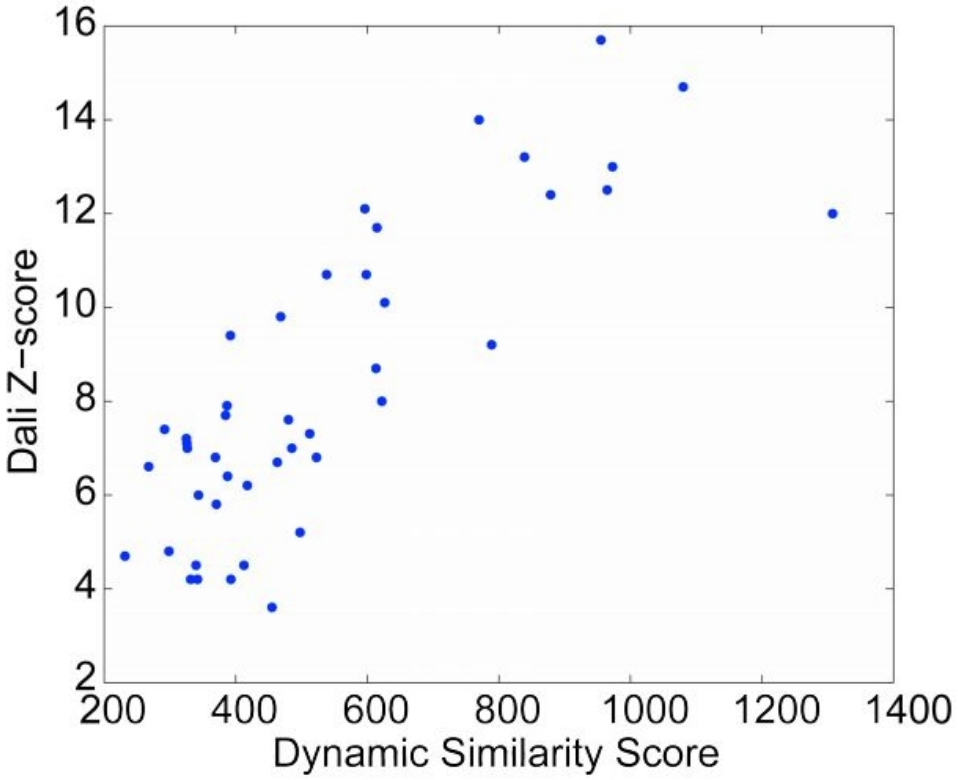

Figure 4.15: Correlation between the dynamic similarity (as computed by the DFM overlap) and the structural similarity as computed by DALI.

Interestingly, all the 6 proteins in the fully connected cluster of the DALI graph (5 of which are well-connected in the dynamics graph too) are from multicellular animals (metazoa), while the other 4 proteins are from unicellular species. The structural difference between PDZ domains from simple and complex organisms is very well-known. First recognized by Liao et al.[337] and exemplified by other authors[338,339], PDZ domains of bacterial and plant origin have a circularly permuted fold compared to the canonical PDZ domain fold found in metazoa. Despite their considerably different architecture, non-metazoan PDZ domains have very similar overall tertiary structure to metazoan PDZ domains. This is indeed reflected by Table 4.3 which shows that metazoan and non-metazoan PDZ do-

mains are significantly similar structures (DALI Z-scores above 2). On the other hand, the fact that the metazoan structures form a distinct cluster in the DALI graph (created with a Z-score threshold of 8.5) highlights the difference between the canonical and the circularly permuted fold.

Putting it all together, the dynamics-based alignment results suggest that the essential structural difference between PDZ domains of metazoan and non-metazoan origin is also reflected by the dissimilarity of their global dynamics. Metazoan PDZ domains appear to be both structurally and dynamically more conserved. However, even within the cluster of the metazoan proteins, there are significant differences in dynamics that can be quantified. Focusing the analysis on binding site residues, the local fluctuation patterns of five metazoan PDZ domains are compared in the next chapter.

4.4.6 Convergence of DFMs

MD simulations are subject to sampling problems. In order to assess whether the simulations have run long enough to provide a reasonable picture of the dynamics the convergence of the DFM patterns has been examined. Five 20 ns simulations of the same protein (PSD-95 PDZ3) have been run using different random seeds for the initial atomic velocities. The similarities of each pairs of DFMs resulted from the different MD runs were measured by the matrix alignment algorithm. Naturally, the five DFMs were not perfectly the same, but the similarity between each pair was highly significant (see Table 4.4).

	Run 1	**Run 2**	**Run 3**	**Run 4**
Run 2	2363.3			
Run 3	3223.7	3029.8		
Run 4	2283.9	1589.4	1715.7	
Run 5	2693.1	2221.1	2692.8	2244.1

Table 4.4: Dynamic similarity scores between five different MD trajectories of PSD-95 PDZ3 starting from different initial velocities. All pairs have highly significant similarities (p-value ≈ 0).

Most importantly, comparing different simulations of the same protein results in much higher similarity scores than the comparison of different PDZ domains. These results lead to the conclusion that the sampling in 20 ns simulations can be sufficient to provide converged DFM patterns for the purpose of a comparative analysis. Clearly, simulating the proteins for longer period of time would further improve the convergence of DFMs.

4.4.7 Correlation matrix vs. DFM

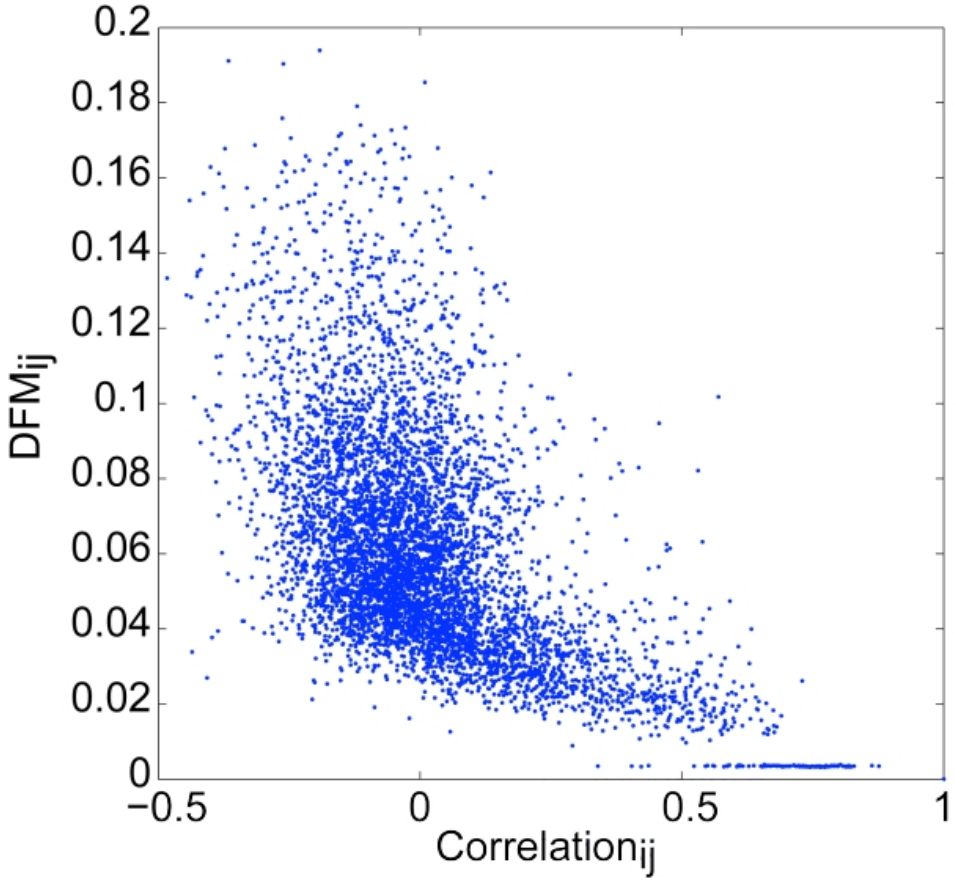

Figure 4.16: DFM versus correlation matrix. Plotting the corresponding entries of the two matrices against each other for all pairs of residues calculated from a 20 ns simulation of PSD-95 PDZ3.

To study the relationship between the dynamic fingerprint matrices and the correlation matrices commonly used in the analysis of proteins dynamics, the two matrices calculated for the same MD simulation have been plotted against each other. That is, for each residue

pair (i,j) the DFM value DFM_{ij} is assigned to the correlation value C_{ij}. For example, Figure 4.16 presents the graph for the 20 ns simulation of PSD-95 PDZ3. The number of points in the graph is the number of different residue pairs in the protein. Note that the correlation vs. DFM plot calculated for other MD simulations appears to be very similar to the example presented here.

For residue pairs of positive correlation, the graph shows a clear tendency between the correlation matrix and DFM entries: the higher the correlation, the lower the pairwise fluctuation (DFM entry). This tendency, however, is not a general rule even for residue pairs of positive correlation, as there are pairs of relatively high (>0.4) correlation and high (>0.07) fluctuation. Moreover, the inverse relationship between correlation and DFM values breaks off at negative (<-0.1) correlations and high (>0.05) DFM values. In these regions of the graph, the two measures appear to be totally uncorrelated. It therefore follows that one cannot accurately predict the fluctuation of two residues which have negative correlation. Similarly, the correlation of two residues cannot be predicted if they have relatively high pairwise fluctuation. Very high (>0.1) pairwise fluctuation values, however, are typically found for pairs of negative correlation. It is important to note that the pairs of neighbouring (covalently bonded) amino acids form a distinct cluster with fluctuation values close to zero and positive (typically >0.6) correlation coefficients.

In summary, it can be concluded that although in a well-defined region of the graph there is a clear inverse tendency between the entries of the two matrices, in other regions they appear to be independent. Therefore the dynamic fingerprint matrix and the correlation matrix provide two different measures for the characterization of protein dynamics.

4.5 Concluding discussions

A novel approach for comparing the backbone dynamics of proteins studied by MD simulations was introduced in this chapter. The method is based on a simple matrix representation of protein motions, the 'Dynamic Fingerprint Matrix' (DFM) which captures the pairwise flexibility of residues. This matrix does not depend on any superposition pro-

cess as it is derived from the distributions of inter-residue distances in the conformational ensemble.

As it is shown by the example of PSD-95 PDZ3, the analysis of the DFM pattern is *per se* useful for characterizing the flexible and rigid regions of a protein. In the literature, the RMSF (root mean square fluctuation) profile is most commonly used to describe the motions of individual residues. By contrast, the DFM pattern describes residue fluctuations in two-dimensions that enables the investigation of the relative mobility of residues. In addition, the Dynamic Fingerprint Matrix can easily be transformed to a one-dimensional profile referred to as the average fluctuation profile (AFP). As demonstrated, the RMSF profile is highly correlated with the average fluctuation profile showing that the DFM captures essentially the same information as RMSF but provides a more detailed description.

While the calculation of RMSF profile is biased by the superposition process and the arbitrary choice of reference structure, the DFM-based average fluctuation profile does not suffer from these issues and is therefore suggested to be a more reliable measure of residue fluctuations. Similarly to this approach, several important quantities that are based on the least square superposition method (e.g. RMSD similarity, coordinate covariance matrix and principal components) could be replaced by unbiased measures that are dependent only on interatomic distances (as discussed in Section 7.4). Such superposition-free analysis is anticipated to be most useful when studying the structure and conformational dynamics of highly flexible proteins for which rigid-body superposition is least reliable.

The pairwise comparison of DFM patterns across different proteins requires a mapping between the residues of these proteins. As discussed in details in Section 2.5.4 and in this chapter, there are two potential solutions for this 'residue matching problem'. The first option is using a prior sequence or structural alignment which defines the submatrices of the DFMs to be compared. The second possibility is aiming to find the best alignment of the two input DFMs that also results in a sequence alignment of the two proteins ('dynamics-based alignment'). The DFM alignment score serves as a measure of global dynamic similarity of the two proteins. On the other hand, this similarity score can also be used to evaluate how well a prior alignment matches dynamically similar regions. All these ideas

were illustrated in this chapter for the PDZ3 domain of PSD-95 versus the PDZ domain of nNOS.

The stochastic optimization algorithm based on a Markov chain Monte Carlo method designed to address the matrix alignment problem was proven to be efficient for the alignment of PDZ domains. The mathematical complexity of the matrix alignment problem is in contrast with the global sequence alignment problem for which the dynamic programming approach (i.e. the Needleman-Wunsch algorithm) guarantees to find the global optimum. Since matrix alignment is an NP-hard problem, search methods that guarantee to find the best solution (e.g. dynamic programming) are not feasible and heuristic strategies (e.g. MCMC) are necessary. Consequently, the search algorithm used here do not guarantee to find the global maximum of the similarity score, especially if many local minima are present in the search space. It is important to note that several potential improvements might be made in the algorithm outlined in this chapter, and different heuristic methods could be tested if they are more efficient in addressing the matrix alignment problem.

The dynamics-based alignment of PSD-95 PDZ3 and nNOS PDZ presented as an example shows that the flexibility patterns of the two proteins are similar enough to enable the algorithm to correctly match the corresponding secondary structural elements of the two PDZ domains. The most important difference between the dynamics-based alignment and the Needleman-Wunsch sequence alignment is the absence of the α2-helix in the former. This observation suggests that the structurally conserved α2-helix which is located next to the peptide binding site has considerably different dynamics in the two PDZ domains.

While the dynamic similarity scores calculated between different MD simulations of the same PDZ domain (PSD-95 PDZ3) are highly significant, the similarity scores between the 10 different PDZ domains show large variation. The results suggest that the dynamics of the four non-metazoan PDZ domains are less conserved. The proposed structural explanation of this observation is that non-metazoan PDZ domains have a circularly permuted fold with regards to the metazoan PDZ fold. While their structures are significantly similar to the metazoan structures according to DALI, this structural similarity is lower than what we can see between the metazoan PDZ domains.

The only difference in sequence between the circularly permuted and canonical fold is that the the C-terminal β-strand ($\beta 6$) in the non-metazoan fold corresponds to the N-terminal β-strand ($\beta 1$) in the metazoan fold. This means that most of the domain sequence is in the same order in the two folds: i.e. the sequence of the $\beta 1$-$\beta 2$-$\alpha 1$-$\beta 3$-$\beta 4$-$\alpha 2$-$\beta 5$ secondary structural elements in the metazoan structures are equivalent to the sequence of $\beta 2$-$\beta 3$-$\alpha 1$-$\beta 4$-$\beta 5$-$\alpha 2$-$\beta 6$ elements in the canonical structures. Although the matrix alignment algorithm is only able to match regions that are in the same sequence order, most parts of the domains, if dynamically similar, could have been aligned. In addition, low similarity scores have been found between the pairs of non-metazoan PDZ domains. These considerations also support the conclusion that the global dynamics of non-metazoan PDZ domains are less constrained.

The results about the dynamic dissimilarity of PDZ domains are somewhat surprising. One would expect that proteins that have as highly conserved tertiary structures as PDZ domains would necessarily have very similar equilibrium dynamics. However, as discussed in details in Section 1.4, currently little is known about how protein dynamics are determined by sequence and structure. Similar three-dimensional structures certainly tend to have similar motions (as demonstrated by ENM studies), but minor sequence and structural changes may result in large differences in dynamics. Interestingly, although high correlation ($r = 0.82$) was found between their structural and dynamic similarity, certain pairs of PDZ domains seem to have diverged global dynamics. Thus it is suggested here that the mapping of PDZ domains between the sequence, structure and dynamics space can be described with the schematic representation shown in Figure 1.16E. Although they have highly conserved 3D-structures, the considerable sequence variation of PDZ domains may explain their dynamic diversity.

The application of the above described methodology to the family of PDZ domains was used here as a pilot study to see whether it was possible to quantify the dynamic similarity in a meaningful way. However, both the DFM representation and the dynamics-based alignment algorithm are generally applicable to any protein families. It is therefore proposed that this approach (or similar, improved versions) could be used as an effective

tool to study the distribution of proteins in the 'dynamics space'. As the method is capable of detecting detailed differences in the dynamics between structures it could also be used to assess the influence of ligand-binding on the dynamics of protein structure.

Related publication:

Münz, M., Lyngsø R., Hein, J. and Biggin, P.C. (2010). Dynamics based alignment of proteins: an alternative approach to quantify dynamic similarity. *BMC Bioinformatics*, 11:188

Chapter 5

Comparative MD study of PDZ domains

5.1 Summary

The goal of Chapter 4 was to introduce a novel methodology for aligning proteins based exclusively on information of their intrinsic motions. The dynamics-based alignment approach had been applied to a set of PDZ domains used as a test case. PDZ domains, however, have highly conserved 3D structures that makes it feasible to compare their dynamics relying on prior sequence or structural alignments that are used to define equivalent residue positions (see detailed explanation in Sections 2.5.4 and 4.3.3). To illustrate this approach and discuss the underlying mechanisms of binding specificity of PDZ domains, this chapter presents a comparative MD study of five PDZ domains: the human Dvl2 (Dishevelled-2) PDZ domain, the human Erbin PDZ domain, the PDZ1 domain of InaD (inactivation no after-potential D protein) from fruit fly, the PDZ7 domain of GRIP1 (glutamate receptor interacting protein 1) from rat and the PDZ2 domain of PTP-BL (protein tyrosine phospatase) from mouse. All-atom MD simulations of 200 ns have been performed for the five apo protein structures. The fluctuations and flexibility properties of their binding sites have been compared after mapping the structures based on a prior multiple sequence alignment. The results summarized here show that despite their high structural similarity, the PDZ binding sites have significantly different equilibrium dynamics. Importantly, the degree of binding pocket flexibility has been found to be closely related to the various characteristics of peptide binding specificity and promiscuity of the five PDZ domains. Overall, these findings suggests that the intrinsic motions of the apo structures play a key role in the distinguishing functional properties of different PDZ domains.

5.2 Introduction

5.2.1 Conformational selection, flexibility, promiscuity and evolvability

A number of structural studies comparing holo and apo forms of proteins have demonstrated that ligand binding is often coupled to conformational changes of the interacting partners.[340–342] The real challenge is, however, to uncover the exact sequence of events resulting in the observed structural changes. Two main models, the induced fit[343] and the conformational selection (or population shift)[344] hypothesis, have been introduced to describe the limiting cases of the complex process of molecular recognition.

According to the induced fit model, ligand binding happens first and the formation of a 'weak complex' is followed by the conformational rearrangement of the protein that results in stronger binding. By contrast, in the conformational selection model, the intrinsic dynamics of the protein lead it to spontaneously transition between a stable unbound and a less stable bound conformation. As the apo protein actually visits the bound state with significant probability, the ligand can bind directly to this conformation shifting the distribution of conformers towards the bound population. As proposed by recent studies, it seems likely that the induced fit and conformational selection mechanisms often act together in ligand recognition.[345,346]

With more and more complete understanding of protein-protein interactions it is increasingly clear that many proteins display functional promiscuity which requires them to be able to interact with multiple partners.[347,348] If the conformational selection mechanism is involved in promiscuous ligand binding, this assumes that the protein needs to visit multiple (often dissimilar) binding conformers capable of binding the different ligands.

The earliest structural evidence of such multispecificity included an X-ray crystallography study of the SPE7 antibody (a monoclonal immunoglobulin E raised against a 2,4-dinitrophenyl hapten) that has been shown to exist in an equilibrium between several binding conformers and is able to bind to two unrelated ligands.[51] An NMR study of apo ubiquitin has identified an ensemble of conformers almost identical to complexes of ubiquitin with 46 different binding partners.[50] An example of promiscuous enzymes is cy-

tochrome P450 which has been shown to adopt a wide range of active site conformations and is able to bind and transform a large diversity of peptides.[349]

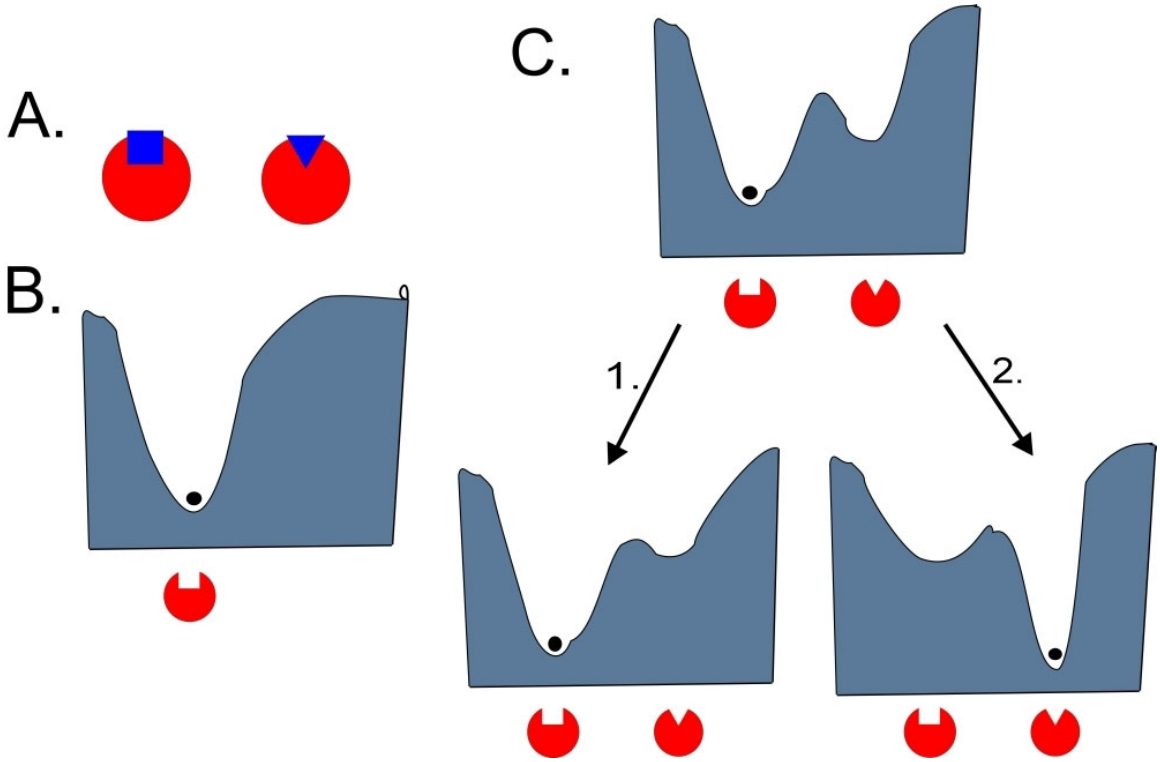

Figure 5.1: Schematic representation of the relationship between flexibility, promiscuity and evolvability. **A.** Two potential conformers capable of binding to two different ligands; **B.** Energy landscape of a rigid protein visiting only one of the two binding conformations: it has restricted specificity for one of the ligand; **C.** Energy landscape of a more flexible protein that visits both binding conformations and is therefore able to promiscuously interact with both partners. The first binding conformer is visited with larger probability. A few mutations may shift the energy landscape (and the distribution of conformations): **1.** rigidification, the higher probability conformer becomes even more dominant, the protein becomes less flexible and consequently more specific; **2.** the originally less probable conformer becomes dominant, the protein may evolve to a new primary function.

As shown by these examples, the intrinsic dynamics of promiscuous proteins let them visit multiple unrelated binding conformers and the property of multispecificity seems to be related to conformational flexibility (see Figure 5.1). Promiscuous proteins that are able to bind to multiple partners through conformational selection need to explore a larger conformational space than those that bind to only a single partner. More rigid binding sites therefore may have restricted specificity with the benefit of higher binding affinity. Indeed, a study of human cytochrome P450 enzymes has found that while a relatively rigid

member of the family (CYP2A6) exhibits narrow substrate specificity, the most flexible member (CYP3A4) is also the most promiscuous one.[350]

Functionally promiscuous proteins could be of key importance for the emergence of new functions in protein evolution (see Figure 5.1). Recent research about the relationship between binding promiscuity, conformational flexibility and evolvability of proteins has been reviewed by Tokuriki et al.[173]. These studies suggest that for proteins that exist in equilibrium between a highly populated native state (interacting with a native ligand) and less populated conformers (binding to alternative partners), mutations can gradually shift the equilibrium towards a promiscuous conformer. This can eventually lead to a new dominant primary function. While mutations may be neutral with regards to the original function (i.e. hardly change the relative occupancy of the native conformer), they may cause significant increase in the occupancy of the alternative conformer.

On the other hand, point mutations that reduce the occupancy of promiscuous conformers may result in a decreased flexibility (rigidification) but increased specificity (and higher affinity) for the native ligand (see Figure 5.1) as for example observed in the process of antibody maturation.[351] Promiscuity may therefore be a common feature of highly evolvable proteins.

5.2.2 What makes PDZ domains specific?

Despite their highly conserved overall fold and binding sites, PDZ domains have been found to have surprisingly diverse binding specificities. Although a series of different classification systems have been proposed aiming to organize PDZ domains based on their preference towards peptide ligands (as discussed in Section 1.5.4), there is no consensus on the best way of classification. The possible reason of this contradiction is that the question of PDZ specificity turned out to be unexpectedly complex as many PDZ domains are able to bind to multiple ligands that belong to different classes of peptide motifs. This property is often referred to as degenerate specificity, multivalent specificity or most commonly, binding promiscuity. In addition, single peptides have been shown to bind to multiple PDZ domains. The complex picture of PDZ-peptide interactions therefore makes it rather

difficult to develop a simple specificity-based classification scheme.

In addition, very little is known about what are the determinants of the specificity and promiscuity of PDZ domains. To address this question, Stiffler et al. have used protein microarrays and quantitative fluorescence polarization to characterize the binding specificity of 157 mouse PDZ domains and found only a weak correlation between the pairwise sequence divergence of PDZ domains and their distances in selectivity space. The fact that overall sequence similarity proved to be a poor predictor of PDZ domain function indicates that the majority of sequence variation in the PDZ family is neutral with regards to peptide-binding selectivity. This also suggests that binding specificity is mostly determined by only a subset of residues that are likely to be located in the binding pocket of the domain.[229].

In order to study the sequence determinants of specific ligand recognition, Tonikian et al. performed mutagenesis at ten binding site positions in the Erbin PDZ domain. As a result, they identified several mutations that altered binding specificity. Since not all of these critical residues were in direct contact with the ligand, Tonikian et al. concluded that specificity of PDZ domains is determined by multiple structural and chemical mechanisms involving both direct interactions and cooperative, long-range effects (see Figure 5.2).[200]

In a recent study, using a combinatorial peptide library and site-directed mutagenesis, Shepherd et al. have found that only four point mutations were enough to switch between the distinct binding specificities of the Tiam1 (T-cell lymphoma invasion and metastasis 1) PDZ and Tiam2 PDZ domains.[352] Gee et al. have come to similar conclusions after performing in-vitro mutagenesis studying the PDZ domains of PSD-95 (postsynaptic density protein 95) and α1-syntrophin. By identifying a few critical sequence positions, they have found that single-amino acid substitutions can alter specificity and affinity of PDZ domains for their ligands.[353]

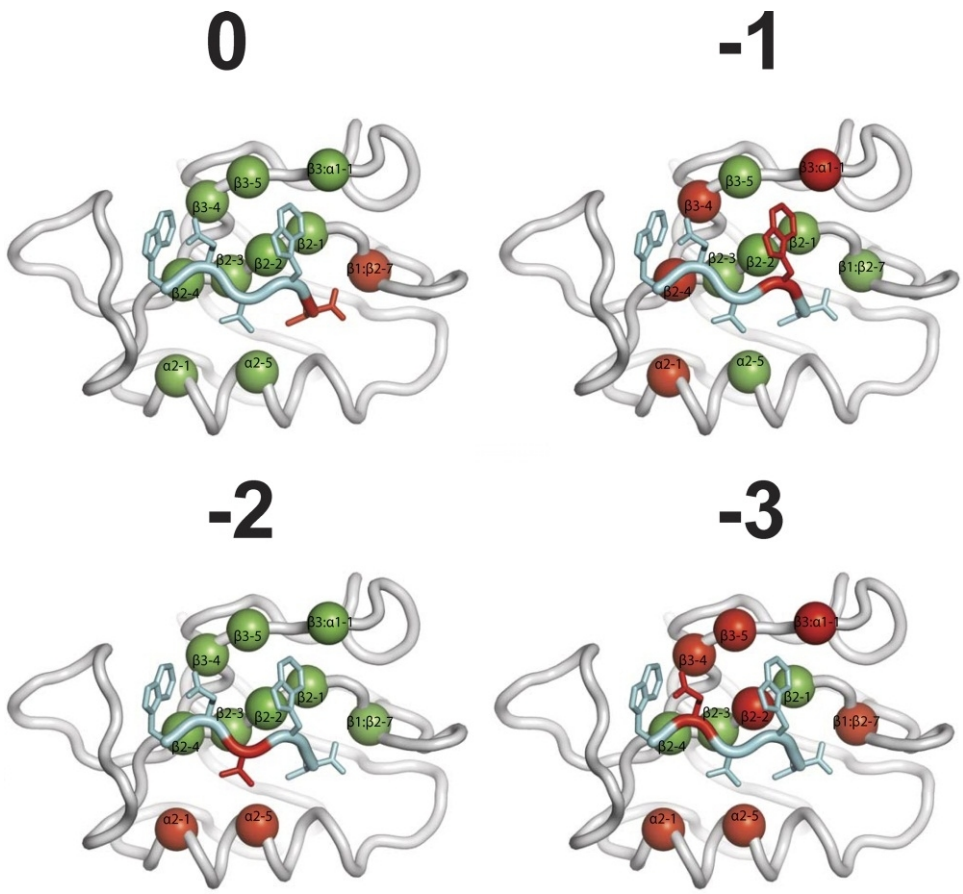

Figure 5.2: Point mutations in Erbin PDZ alter the peptide binding specificity of the PDZ domain. The ten positions tested with mutagenesis by Tonikian et al. are shown as spheres. Those mutated positions that have affected the specificity for the indicated peptide positions 0, -1, -2 and -3 (defined in Section 1.5.4) are highlighted in red. (Image courtesy: Tonikian et al. 2008[200])

The fact that ligand specificity relies on minor sequence modifications, while the chemistry of the binding region and the overall domain fold are rather well conserved, suggests a very favourable flexibility property of the PDZ domain fold.[334] PDZ domains are both versatile and robust because mutations frequently change their specificities without a loss of function.[200] Similar robustness under high mutational pressure has also been observed for other peptide-binding domains (i.e. WW and SH3 domains).[354,355]

On the other hand, a number of other studies[248,249,356,357] have found that the conformational dynamics of PDZ domains may also play crucial role in determining binding specificity. These results suggest that the intrinsic fluctuations of PDZ structures are likely to be

also related to the selectivity for peptide ligands. Recently, Gerek et al. used a modified coarse-grained elastic network model to find characteristic residue fluctuation patterns for PDZ domains belonging to different specificity classes. By clustering these residue fluctuation profiles, they have identified common motion characteristics of Class I and Class II type PDZ domain interactions.[356]

Basdevant et al. performed 20-25 ns molecular dynamics simulations of 12 PDZ domain complexes and used the MM/PBSA (Molecular Mechanics/Poisson-Boltzmann Surface Area) method to analyse electrostatic, non-polar and configurational entropy contributions to the binding free energies. Their results show that the degree to which the dynamics of the peptide ligands are coupled to those of the PDZ domains varies highly. They concluded that complex-specific dynamical or entropic responses may form the basis of the selective recognition of peptides.[357]

It is important to note that different flexible docking strategies have already been proposed to be able to incorporate the effect of binding site flexibility in structure-based drug design studies targeting PDZ domains.[248,249].

The goal of this work was to further investigate the role of conformational dynamics in determining the ligand binding specificity of PDZ domains. In particular, given the possible relationship between flexibility and promiscuity discussed in the previous section, the objective was to understand to what extent the property of multi-specificity of these domains is correlated with the flexibility of their binding pockets.

5.2.3 The 5 PDZ domains compared in this study

Below is an introduction to the five PDZ domains selected for this study. For each PDZ domain-containing protein, a brief description is given about its biological/biochemical function, most important interaction partners and clinical relevance. Distinguishing characteristics of each PDZ domain predicted by experimental studies are highlighted and the main questions addressed are also described in this section.

Dvl2 PDZ (*Homo Sapiens*)

Wnt signalling pathways are crucially implicated in normal development of tissues and organs during embryogenesis and adult tissue maintenance.[358] The highly conserved Wnt genes encode a large family of secreted protein growth factors that bind to Frizzled and LRP receptors on the cell surface. Wnt signals are transmitted via a series of cytoplasmic components to the nucleus and result in altered transcription of Wnt target genes. Controlling cell-cell communication Wnt signalling plays various roles during development such as regulation of cell fate, proliferation, migration, polarity and death.

Wnt signals have been shown to be transmitted by three distinct intracellular cascades: the canonical (Wnt/β-Catenin) pathway and the non-canonical Wnt/Ca^{2+} and Planar Cell Polarity (PCP) pathways.[359] Inappropriate regulation of Wnt signalling is implicated in the development and progression of a variety of human malignant cancers.[358]

Dishevelled (Dvl) proteins are key players in canonical Wnt signalling. They interact with Frizzled receptors to transduce Wnt signals to downstream components of the pathway.[360,361] In the absence of Wnt signals the APC/Axin/CK1/GSK3β destruction complex is able to form promoting the hyperphosphorylation and proteolytic degradation of β-catenin. However, when a Wnt signalling protein binds to the Frizzled/LRP receptor complex, it activates cytoplasmic Dvl that enhances the phosphorylation of the GSK3β protein. As a result, the destruction complex becomes inhibited that leads to the stabilization of cytoplasmic β-catenin. Unphosphorylated β-catenin translocated to the nucleus and interacts with TCF/LEF transcription factors altering gene transcription.[360]

In addition to Frizzled receptors, Dvl proteins interact with several other partners including components of the canonical Wnt pathway (Axin and GBP/Frat), the Wnt/Ca^{2+} pathway ($G\alpha_o/G\alpha_t$) and the PCP pathway (Rac1, Daam1 and strabismus).[361] The fact that Dvl proteins serve as "interaction hubs" suggests that they play key regulatory roles in the signal distribution between the three alternative Wnt pathways.[362]

Dvl proteins are composed of three modules: an N-terminal DIX domain, a central PDZ domain and a C-terminal DEP domain (see Figure 5.3A).[361] The combination of these three protein-protein interaction domains makes Dvl PDZ capable of binding multiple partners

simultaneously. On the other hand, a series of studies have provided evidence for the binding promiscuity of the central PDZ domain which alone serves as an interaction hub with the ability of binding various signalling proteins.[200]

It is the PDZ domain of Dvl that has been found to interact with a C-terminal region of Frizzled receptor. Shown by NMR spectroscopy data, a conserved internal peptide motif at the cytoplasmic tail of Frizzled binds to the conventional binding site of the mouse Dvl1 PDZ domain.[231] The interaction mode with an internal peptide differs significantly from the classical carboxyl-terminal binding mode of PDZ domains (see Section 1.5.3).

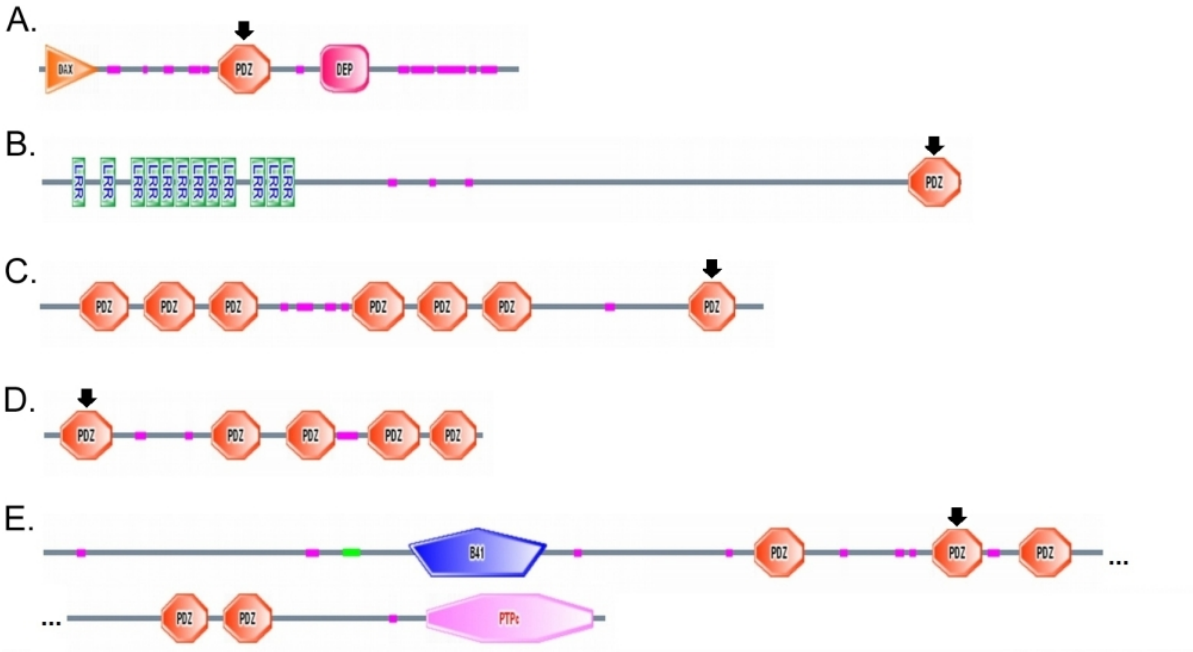

Figure 5.3: Domain architecture of the five PDZ-containing proteins studied here as predicted by the online sequence annotation tool SMART[363]. The five PDZ domains analysed are highlighted with arrows: **A.** Dvl2 (Homo Sapiens); **B.** Erbin (Homo Sapiens); **C.** GRIP1 (Rattus Norvegicus); **D.** InaD (Drosophila Melanogaster); **E.** PTP-BL (Mus Musculus).

In addition, a recent peptide-phage display study showed that the human Dvl2 PDZ domain is able to bind both C-terminal and internal ligands via specific interactions.[364] Four crystal structures of Dvl2 PDZ in complex with peptides representing four distinct ligand families have been solved. One of the structures (pep-C1) exemplifies C-terminal ligand binding, and the other three structures (pep-N1, pep-N2 and pepN3) illustrate dif-

ferent orientations of internal ligand binding. The four representative complexes show that Dvl2 PDZ can recognize a diverse set of peptides and bind them using different binding modes. The main chain conformation of Dvl2 PDZ appears to change significantly as a result of ligand binding, unlike in other PDZ domains, e.g. in Erbin PDZ for which apo and holo conformations are almost identical (see next subsection). In other words, the binding cleft of Dvl2 PDZ has been found to be more flexible than those of canonical PDZ domains. The remarkable versatility of the Dvl PDZ binding pocket explains the promiscuity of the domain towards biological interaction partners.

Given its central role in the Wnt signalling pathways, Dvl PDZ domain has been recognized as a potential drug target. Several studies aimed to develop peptides[364,365] and small molecule inhibitors[243–246] targeting the PDZ domain of Dishevelled to suppress upregulated Wnt signalling in tumour cells.

In this work, the goal of studying the human Dvl2 PDZ domain was to find out how the flexibility of its binding cleft is reflected in the intrinsic dynamics of the domain. Investigating what roles the equilibrium fluctuations play in the versatility of the Dvl2 PDZ binding pocket is not only important for understanding the link between protein dynamics and promiscuity, but could also be relevant for structure-based drug design studies targeting this PDZ domain.

Erbin PDZ (*Homo Sapiens*)

The ERBB protein family is a group of four receptor tyrosine kinases (RTKs) including the epidermal growth factor receptor (EGFR), ERBB2/HER2, ERRB3/HER3 and ERRB4/HER4. [366] These four structurally closely related RTKs are expressed in a variety of tissues of epithelial, mesenchymal and neuronal origins and have essential roles in regulating diverse cellular processes such as proliferation and differentiation.[367]

Binding extracellular growth factor ligands (EGFs) induces the formation of homo- and heterodimers of these receptors stimulating their intrinsic tyrosine kinase activity. This results in autophosporylation of specific tyrosine residues within their cytoplasmic domain activating intracellular signalling cascades.[368] ERBB2 is thought to be an orphan receptor,

i.e. none of the EGF family ligands can activate it, however, it is a preferential heterodimerization partner for other ERBB receptors.[366]

Malfunction of ERBB signalling has been shown to be implicated in a number of neurodegenerative diseases including multiple sclerosis and Alzheimer's disease.[369] Furthermore, overexpression of ERBB2 has been found in breast, ovary, lung and other types of malignant epithelial cancers, often correlating with more aggressive phenotypes and poor prognosis.[370]

Erbin (ERBB2 interacting protein) is a member of the LAP protein family. LAP (leucine-rich repeat and PDZ-containing) proteins have been found to play important roles in the regulation of maintaining the shape and polarity of epithelial and neuronal cells.[371] In particular, Erbin is thought to serve as an adaptor for ERBB2 protein controlling the basolateral localization of ERBB2 receptor in epithelial cells.[366,372,373]

Since Erbin is an important component in ERBB signalling, a number of studies have investigated its implication in cancer; such as the development of basal cell carcinoma[374] and the sensitivity of breast cancer cells to tumour necrosis factors[375].

On the other hand, in addition to ERBB2, Erbin is able to bind a other signalling proteins. C-terminal phage display[376] and yeast two-hybrid screens[377,378] have identified a series of additional interaction partners including p120-catenins. Erbin is able to bind components of the cadherin-catenin cell adhesion complex (e.g. δ-catenin, p0071 and ARVCF) [376] and via its interaction with δ-catenin it has been shown to play a critical role in regulating dentritic morphogenesis in hippocampal neurons[379]. Furthermore, Erbin has also been shown to play various additional roles, for example it inhibits EGF signalling preventing the activation of the Raf-1 kinase by Ras.[380]

Erbin is composed of 16 N-terminal leucine-rich repeats (LRRs) and a C-terminal PDZ domain (see Figure 5.3B). The leucine-rich repeats are involved in mediating the basolateral localization of the protein.[381] On the other hand, it is the PDZ domain of Erbin that interacts with ERBB2 and p120-catenins.

Interestingly, while δ-catenin, p0071 and ARVCF are all class I PDZ ligands, ERBB2 belongs to class II ligands (see canonical classification explained in Section 1.5.4). As it has

dual specificity, Erbin belongs to class V, a family of PDZ domains that are able to bind both class I and class II ligands.[382] Despite its rather promiscuous nature, Erbin PDZ has stringent binding specificity for C-terminal peptides and its interaction partners share a highly conserved PDZ domain-binding motif.[376]

Comparison of class I/II peptide-bound and apo experimental structures of Erbin PDZ has shown no structural change in the $\beta 2/\alpha 2$ binding pocket upon ligand binding and only slight changes in the main chain of the whole domain.[55] The little structural variability observed for Erbin PDZ is in contrast with the highly flexible binding pocket of Dvl2 PDZ which has been shown to undergo significant structural change upon peptide binding (see previous subsection).

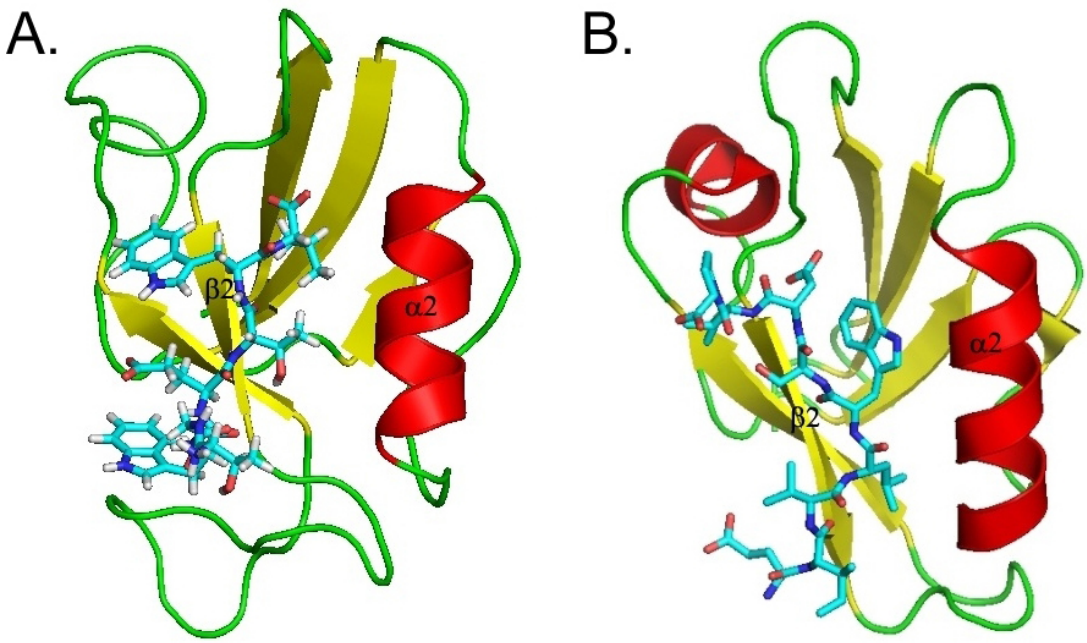

Figure 5.4: Experimental structures show differences between binding modes in PDZ-peptide interactions: **A.** human Erbin PDZ domain binding to a class I ligand exemplifying interactions with C-terminal peptides (PDB: 1n7t); **B.** human Dvl2 PDZ domain binding to an internal ligand exemplifying interactions with internal peptides (PDB: 3cc0).

In this work, the aim of studying the human Erbin PDZ domain was to find out whether the relative rigidity of this PDZ structure was also reflected in its equilibrium dynamics.

In particular, the goal was to see if its binding groove shows considerably different fluctuations compared to the flexible Dvl2 PDZ domain binding pocket. Comparison of these two PDZ domains may help to understand the relationship between flexibility and binding promiscuity.

GRIP1 PDZ7 (*Rattus Norvegicus*)

GRIP1 (glutamate receptor interacting protein 1) plays an important role as a scaffold for the assembly of multiprotein signalling complexes and as a mediator of trafficking of its interaction partners in neurons.[207,215,383] GRIP1 is composed of 7 PDZ domains (see Figure 5.3C).[383] PDZ4-6 are known to interact with the COOH-terminal of the GluR2/3 subunits of AMPA receptors.[207] In addition, GRIP1 was shown to bind various signalling and cytoskeletal proteins, such as EphB receptor tyrosine kinase, ephrinB ligands, alpha-liprin scaffolding protein, proteoglycan NG2, Fras1 and 2, kinesin-1/KIF5 microtubule motor protein, matrix metalloproteinase 5 and GRASP-1 (GRIP1-associated protein 1).[384] This suggest that by organizing macromolecular complexes GRIP1 has a role of linking AMPA receptors and other membrane proteins to the cytoskeleton, participating in membrane trafficking and signalling pathways.[384]

Given its central role in neuronal signalling, it is likely that GRIP1 will be found to be associated with a number of neurological disorders. For example, a recent study has identified five rare missense variants within or near the PDZ4-6 genomic region exclusively in autistic patients. Two variants that are correlated with a severe deficit in social interactions observed in autism were shown to have altered binding with GluR2/3 resulting in faster recycling and increased surface distribution of GluR2 in neurons.[385]

In this work GRIP1 PDZ7, the PDZ domain that interacts with GRASP-1, a Ras guanine-nucleotide exchange factor regulating synaptic distribution of AMPA receptors[383], has been studied. Interestingly, the binding of GRIP1 PDZ7 to GRASP-1 is unusual compared to all other known PDZ domain-ligand interactions. As discussed in Section 1.5.3, in PDZ domains the peptide binding groove is usually formed between their α2-helix, β2-strand and the carboxylate-binding loop. By contrast, the NMR structure of GRIP1 PDZ7 (PDB:

1m5z) shows that the traditional binding pocket formed by $\alpha 2$ and $\beta 2$ adopts a closed conformation and is probably unable to interact with a carboxyl peptide.[383] However, unlike other PDZ domains, GRIP1 PDZ7 has a large solvent exposed hydrophobic surface formed by the $\beta 5$ strand, $\alpha 2$ helix and the loop connecting these two secondary structural elements. It has been found that GRASP-1 binds to GRIP1 PDZ7 via this hydrophobic surface distinct from the conventional ligand-binding $\alpha 2/\beta 2$-groove.[383] The hypothesis that GRIP1 PDZ7 is unable to bind carboxyl peptides via its $\alpha 2/\beta 2$-groove has also been confirmed by a yeast two-hybrid screening in which no interaction partner with classical PDZ-binding carboxyl termini were found.[386]

Name	PDB	Organism	Size	Binding pocket	Characteristics
Dvl2 PDZ	3cbx	*Homo Sapiens*	105	$\beta 2 : [280Ile, 284Gly]$ $\alpha 2 : [330Asn, 339Asp]$	flexible; large structural variability
Erbin PDZ	2h3l	*Homo Sapiens*	103	$\beta 2 : [334Phe, 338Gly]$ $\alpha 1 : [388His, 397Thr]$	rigid; little structural variability
Grip1 PDZ7	1m5z	*Rattus Norvegicus*	91	$\beta 2 : [38Phe, 42Asp]$ $\alpha 2 : [84Cys, 93Glu]$	closed; unable to bind C-terminal peptides
InaD PDZ1	1ihj	*Drosophila Melanogaster*	98	$\beta 2 : [30Ile, 34Arg]$ $\alpha 2 : [84Glu, 93Glu]$	capable of different binding modes
PTP-BL PDZ2	1gm1	*Mus Musculus*	94	$\beta 2 : [27Ile, 31Gly]$ $\alpha 2 : [78His, 87Asn]$	induced-fit binding mechanism

Table 5.1: Summary of the 5 PDZ domains used in this study. The table highlights the sequence regions corresponding to the $\alpha 2$ helix (or $\alpha 1$ helix in Erbin PDZ) and the $\beta 2$ strand. (See Methods for details of the multiple sequence alignment analysis used for defining binding site residues). Notable characteristics of the five binding pockets predicted by experimental studies are also listed.

Although the solution structure shows that the GRIP1 PDZ7 $\alpha 2/\beta 2$-groove is significantly smaller than those of other PDZ domains, this does not exclude the possibility that it can undergo a conformational change opening the binding pocket. For example, such conformational change has been observed in the PDZ domain of LARG (Leukemia-associated Rho guanine nucleotide exchange factor).[55] LARG PDZ has very similar three-dimensional

structure to GRIP1 PDZ7 (with an αC RMSD dissimilarity of only 1.4 Å in the secondary structural elements) and its binding groove is also closed in the apo state. However, as shown by a ligand-bound structure, the base of LARG PDZ $\alpha 2/\beta 2$-pocket can open up which increases the accessible surface area providing better binding for peptides.[55]

It is not clear if the $\alpha 2/\beta 2$-groove of GRIP1 PDZ7 is also able to undergo such transition from its closed state to an open state capable of carboxyl peptide binding. The goal of investigating GRIP1 PDZ7 was to address this question by studying the intrinsic fluctuations of the domain.

InaD PDZ1 (*Drosophila Melanogaster*)

Photoreceptor cells are specialized neurons found in the eye's retina that are sensitive to light and capable of converting the signals of absorbed photons to changes of membrane potential. This process called phototransduction is mediated by a G-protein-coupled cascade and is among the best understood signalling pathways in biology.[387] In particular, the phototransduction of the fruit fly is a well-studied model system. It is the fastest known G-protein signalling cascade: photon absorption results in membrane depolarization in about 20 milliseconds.[388]

The phototransduction cascade of Drosophila is initiated with the activation of the light-sensitive protein rhodopsin which is composed of a protein (opsin) and a covalently linked photoreactive chromophore. Upon absorption of a photon, the chromophore is isomerized from a 11-cis to an all-trans configuration resulting in a conformational change of rhodopsin.[389] The active form of rhodopsin (referred to as metarhodopsin) triggers the G_q protein to release its alpha subunit which in turn activates the PLCβ (phospholipase Cβ) enzyme also known as NorpA. As a result, the active PLCβ enzyme hydrolyzes PIP2 (phosphatidylinositol (4,5)-bisphosphate), a phospholipid located in the plasma membrane. The products of this reaction are IP3 (inositol triphosphate) and DAG (diacylglycerol) that function as secondary messenger molecules.[389]

While IP3 diffuses through the cell, DAG stays inside the membrane and causes the opening of the cation-selective ion channels TRP (transient receptor potential) and TRPL

(TRP-like). The resulting Ca^{2+} influx leads to the depolarization of the photoreceptor cell. [389] On the other hand, calcium entry also mediates the deactivation of the phototransduction cascade by triggering a negative feedback pathway which involves a series of other proteins including eye-specific PKC (protein kinase C), CaM (calmodulin) and NinaC (unconventional myosin III). [390]

A probable explanation of the remarkable high speed of Drosophila phototransduction is that most components of the cascade are clustered together and form a large signalling complex (also referred to as the signalplex or transducisome). Since the interaction partners are located in close proximity to each other, it significantly increases the rates of interaction between them thus maximizing the speed of signalling. [208]

InaD (Inactivation-no-afterpotential D) protein plays and essential role in the formation of the signalplex as it functions as an adaptor that brings together the components of the phototransduction pathway into a large macromolecular complex. [208,391] InaD is composed of five PDZ domains (see Figure 5.3D), all of which have been shown to interact with various components of the phototransduction cascade. [208,391]

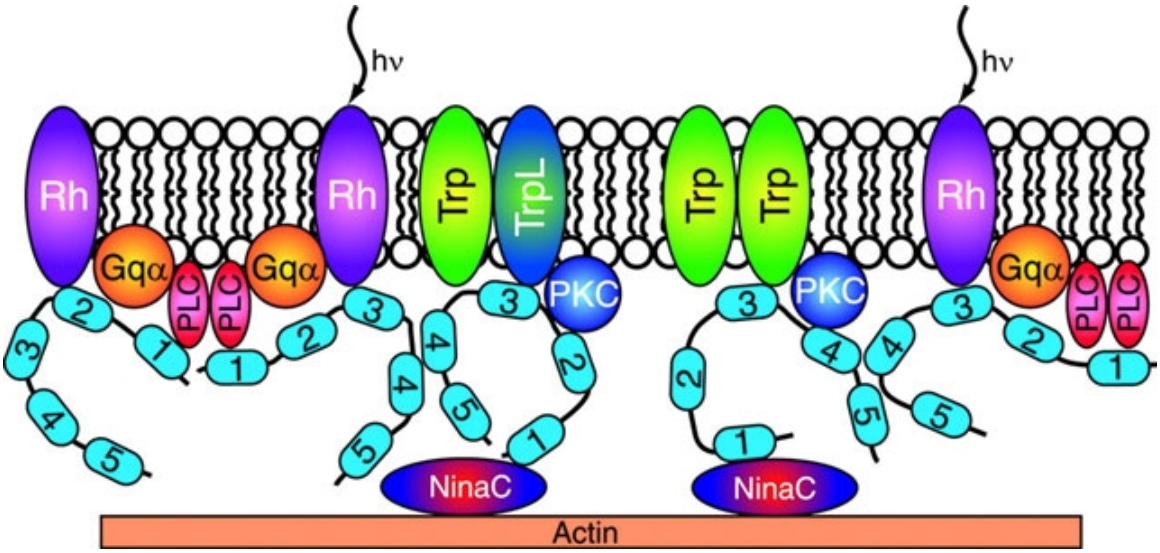

Figure 5.5: Schematic illustration showing how the InaD protein facilitates the assembly of the signalplex at the plasma membrane in Drosophila phototransduction. The five PDZ domains of InaD interact with different components of the phototransduction cascade serving as a scaffold for the signalling complex. InaD PDZ1 also binds to the unconventional myosin NinaC anchoring the signalplex to the actin cytoskeleton. (Image courtesy: Kimple et al. 2001) [392]

For example, the TRP ion channel has been found to bind to the PDZ3 domain of InaD via an internal sequence in the cytoplasmic tail of the receptor.[208] PLCβ binds to the InaD PDZ1 domain with its C-terminus and is also able to interact with the PDZ5 domain via an internal sequence.[393–395] The eye-specific PKC has been shown to bind to both PDZ2 and PDZ4 of InaD.[208] NinaC can interact with the PDZ1 domain.[396] CaM shows affinity for a region between PDZ1 and PDZ2 domains.[395] Furthermore, rhodopsin and TRPL have also been shown to interact directly or indirectly with PDZ3 and/or PDZ4[395], while the activated G_q alpha subunit has been found to associate with InaD via binding PLCβ[397].

Based on the above-listed interactions it has been suggested that InaD plays a key role in the subcellular targeting of the components of the phototransduction cascade in Drosophila and also serves as a scaffold for the assembly of the signalling complex at the plasma membrane. In addition, InaD has been shown to be able to multimerize via its PDZ3 and PDZ4 domains[395] forming homopolymers that function as even larger intracellular scaffolds.

The N-terminal PDZ domain of InaD studied in this work plays a crucial role in linking the PLCβ (NorpA) protein to the signalplex. The crystal structure of its complex with a C-terminal peptide of NorpA (PDB: 1ihj) shows in many aspects similar binding mode as seen in a number of other PDZ domains, however, important differences have also been found. For example, an intermolecular disulfide bond is formed between 31Cys of InaD PDZ1 and (-1)Cys of the NorpA peptide that is unique in PDZ-peptide interactions. As confirmed by further in vitro and in vivo studies, this disulfide bond is required for high affinity interaction.[392]

On the other hand, InaD PDZ1 has been shown to bind to the unconventional myosin NinaC.[396] Since NinaC interacts with actin[398], InaD PDZ1 is likely to have a role of linking the phototransduction signalplex to the actin cytoskleton through NinaC. Using yeast two-hybrid assay, Wes et al. tested different segments of the C-terminal region of NinaC for their ability to interact with InaD PDZ1. Their results show that while the C-terminal 21 residues of NinaC are sufficient to bind to PDZ1, the C-terminal 15 residues are not enough.[396] This observation suggests that InaD PDZ1 may interact with NinaC in a differ-

ent mode than it does with NorpA.[392] However, no experimental structure of the complex of InaD PDZ1 and NinaC peptide is currently available.

In this study, the objective was to find out if it is possible to use MD simulation data to predict the promiscuous nature of the InaD PDZ1 binding site. Is the fact that InaD PDZ1 is able to interact with peptides using different binding modes reflected in its conformational flexibility?

	Erbin PDZ	**GRIP1 PDZ7**	**InaD PDZ1**	**PTP-BL PDZ2**
Dvl2 PDZ	24.3 % 1.8 Å (1.1 Å)	19.8 % 2.4 Å (1.4 Å)	19.4 % 2.0 Å (1.7 Å)	30.1 % 2.1 Å (1.4 Å)
Erbin PDZ	—	26.4 % 3.5 Å (0.8 Å)	24.5 % 2.6 Å (1.0 Å)	25.5 % 3.2 Å (0.6 Å)
GRIP1 PDZ7	—	—	29.7 % 1.4 Å (1.1 Å)	28.6 % 1.0 Å (0.5 Å)
InaD PDZ1	—	—	—	23.4 % 1.3 Å (0.9 Å)

Table 5.2: Sequence and structural similarity of the five PDZ domains used in this study. Pairwise sequence identity, RMSD based on residues belonging to secondary structural elements and RMSD between binding site residues only (values shown in brackets). Equivalent positions of the five PDZ structures are defined by their multiple sequence alignment (see details in Methods).

PTP-BL PDZ2 (*Mus Musculus*)

PTP-BL (PTP-Basophil-like) is a large (270 kDa) mouse non-receptor protein tyrosine phosphatase that is expressed during development and found in various adult tissues such as pancreas, kidney, muscle, liver, brain and lung.[399,400] It is composed of multiple functional modules: an N-terminal KIND domain, a FERM domain, five different PDZ domains and

a C-terminal catalytic tyrosine phosphatase domain (see Figure 5.3E).

The five PDZ domains of PTP-BL and its human homologue, PTP-Bas have been shown to interact with various cellular partners including the tumour suppressor protein APC (adenomatosis polyposis coli), the LIM-domain containing proteins RIL and Trip6, the death (apoptosis) receptor hFas, a neurotrophin receptor $p75^{NTR}$ and Ephrin B proteins, ligands of the Eph receptors.[401]

Despite the wide range of interactions in which PTP-BL/Bas has been found to be involved, the precise biochemical/biological function of the protein is still little understood. The fact that PTP-BL/Bas contains multiple functional domains and the large number of its splice variants[401,402] suggest a functional versatility of the protein.[401] This notion is indeed confirmed by the diversity of cellular processes in which PTP-BL/Bas has been found to be implicated including organization of the cell cortical actin cytoskeleton, regulation of intracellular vesicular transport, cytokinesis and Fas-mediated apoptosis in human cells.[401]

The second PDZ domain of PTP-BL/Bas is among the most studied PDZ domains. It has been shown to bind to peptides that belong to the class I, II and III family.[232] One of the reasons of the popularity of the PTP-BL/Bas PDZ2 domain is its allosteric nature that makes it an ideal model system to study intramolecular signalling in proteins.

A series of studies have revealed the existence of intramolecular allosteric communication pathways within a single PTP-BL/Bas PDZ2 domain.[222,403] Using NMR relaxation methods to study ^2H-based methyl side-chain dynamics, significant changes have been detected in side-chain dynamics parameters upon binding a peptide ligand. Altered side-chain dynamics has also been observed remote (>9 Å) from the binding site.[222] This indicates that the binding pocket of PDZ2 is linked to distal sites located at the opposite side of the domain via a network of dynamically coupled residues that are involved in long-range allosteric signal transduction.

The allosteric pathways of PDZ2 identified in NMR relaxation studies correlate well with those positions found to be statistically coupled based on an evolutionary analysis of the PDZ family.[403] Interestingly, screening a peptide δ phage display library has shown

that the ability of mouse PDZ2 to bind to class III-type ligands is regulated by the presence or absence of PDZ1. In a full PTP-BL protein, when PDZ2 is not isolated, PDZ1 can interact with PDZ2 at a surface opposite the binding pocket and mediate the specificity of PDZ2 binding pocket through long-range allosteric communication.[232]

As discussed above, PTP-BL has been demonstrated to interact with APC, a protein linked to familiar and sporadic human colorectal cancers. A yeast two-hybrid study has found that PTP-BL PDZ2 binds to the extreme C-terminus of APC.[402] Since the interaction between the APC protein and PTP-BL may indirectly modulate the tyrosine phosphorylation levels of associated proteins (e.g. β-catenin), it might play a major role in regulating cell division, adhesion and migration and therefore could be of clinical importance.[402]

Comparison of the solution structure of mouse PTP-BL PDZ2 in complex with the 10 C-terminal residues of APC and the ligand-free structure has revealed several binding-dependent structural changes.[404] In particular, superposition between bound and ligand-free PDZ2 domain has shown a subtle but detectable change in the orientation of the α2 helix relative to the β2 strand while the overall position of L1 loop is also shifted. Such reorientation of the α2 helix upon peptide binding has also been observed in some other PDZ domains.[405–407]

By contrast, in an other study, high resolution crystal structures and residual dipolar couplings (RDC) data of human PTP-Bas PDZ2 domain in complex with a C-terminal peptide of guanine nucleotide exchange factor (RA-GEF-2) showed no significant binding-dependent structural change of the domain.[17] A possible reason of the discrepancies between the NMR and crystal structures are the use of different methodologies for structure determination. On the other hand, there may also be a real difference between the peptide binding mechanism of mouse and human PDZ2 domains which differ by only 6 residues (two of which are conservative mutations).

An important question is whether the structural change between the free and APC peptide-bound mouse PTP-BL PDZ2 domain is resulted from conformational selection or induced fit mechanism. A recent binding kinetics study using continuous-flow fluorometry has concluded that peptide-binding of mouse PDZ2 domain follows a two-step

induced-fit mechanism as the observed rate constant of the reaction between the PDZ domain and its target peptide has a non-linear dependence both on the concentrations of PDZ2 and the peptide.[198] Although in these kinetic experiments a target peptide mimicking the last six residues of the RA-GEF-2 protein has been used, it belongs to the same class of recognition motifs (i.e. class I) as the APC peptide used for structural studies.

Similarly, a kinetics study of the second PDZ domain of SAP97 (Synapse-associated Protein 97) binding to a peptide derived from the C-terminus of the HPV-18 (Human papillomavirus 18) E6 protein suggested an underlying induced fit mechanism.[408] Despite these examples, however, little is known about whether the lock-key, induced fit or conformational selection model is best to describe most PDZ-peptide interactions.

In this work, the goal was to see whether MD simulation of apo mouse PTP-BL PDZ2 is useful to confirm the induced fit binding of the domain to the APC peptide. In addition, can MD simulations help to decide if the other four PDZ domains follow lock-key, induced fit or conformational selection mechanism when binding to their peptide ligands?

Additional notes

As shown by the overview above, all five PDZ domains studied in this work play crucial roles in signalling cascades by serving as intracellular scaffolds for the formation of supramolecular complexes essential for efficient signal processing (see Figure 5.5). As a result, all five PDZ domains (or their human homologs) are of clinical interest due to their central roles in disease pathways.

Four of these PDZ domains (Dvl2 PDZ, Erbin PDZ, InaD PDZ1 and PTP-BL PDZ2) are promiscuous in the sense that they are able to interact with multiple partners. However, while for example Dvl2 PDZ is capable of interacting with peptides using different binding modes (binding both C-terminal or internal peptides), Erbin PDZ is able to interact only with very similar peptides in rather similar binding modes. (Figure 5.4 illustrates the difference between C-terminal and internal peptide binding modes.)

We can therefore formulate a stronger definition of binding promiscuity: i.e. the ability to interact with multiple ligands that require the binding pocket to adopt considerably

different shapes. In this sense Dvl2 PDZ is promiscuous and Erbin PDZ is not. If conformational selection plays a role in the recognition of peptides, the above-defined property of promiscuity must correlate with intrinsic conformational flexibility since the binding pocket needs to visit all different shapes required for binding multiple ligands.

The five PDZ domains studied here are summarized in Table 5.1 which also contains a list of the characteristic features of their binding pockets based on experimental results. It is intriguing that these five PDZ domains have such diverse specificity properties despite their structurally conserved folds and binding sites. (Their pairwise sequence and structural similarities are given in Table 5.2). The goal of studying these domains was to see whether their intrinsic dynamics can explain their different characteristics.

5.3 Methods

5.3.1 Measures of structural similarity

Let A and B denote two proteins that consist of N_A and N_B residues, respectively. In this study, residues are represented by their α-carbon atoms. An alignment between the two structures defines a mapping between the two sets of residues. Let N denote the number of aligned residue pairs (after removing positions aligned to gaps). The two sets of aligned residues are described by the NxN distance matrices of their α-carbon atoms denoted by d^A and d^B: i.e. the matrix entry d^A_{ij} is the distance of α-carbon atoms i and j in structure A.

Difference distance matrix

The difference distance matrix δ between structure A and B is defined as

$$\delta(A,B)_{ij} := d^A_{ij} - d^B_{ij} \tag{5.1}$$

Positive entries in this matrix indicate pairs of atoms of larger distance in structure A than in structure B. This matrix can be used to characterize the location and extent of structural differences between two different proteins or two conformations of the same protein.

dRMSD dissimilarity

The dRMSD (distance root mean square deviation) measure of dissimilarity between the two structures is defined as

$$dRMSD(A,B) := \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(d_{ij}^{A}-d_{ij}^{B})^2} = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta(A,B)_{ij}^{2}} \tag{5.2}$$

This measure was used instead of the standard RMSD dissimilarity because dRMSD is not dependent on structural superposition.

5.3.2 Characterising conformational dynamics

Let $\mathbf{S} = \{S_1, S_2, ..., S_K\}$ denote an ensemble of conformations of a protein represented by its α-carbon atoms. Let the number of its residues be N.

Fluctuation matrix

An NxN matrix referred to as the F fluctuation matrix describing the extent of pairwise fluctuation of α-carbon atoms was introduced in this study. Matrix F contains the variances of the distance of each α-carbon pair, where the variance is calculated over the whole ensemble. It is precisely defined as

$$F(\mathbf{S})_{ij} := Var(d_{ij}^{S_k}) = \frac{1}{K}\sum_{k=1}^{K}(d_{ij}^{S_k}-\overline{d_{ij}})^2 \tag{5.3}$$

where $\overline{d_{ij}} = \frac{1}{K}\sum_{k=1}^{K} d_{ij}^{S_k}$ is the mean distance of α-carbon atoms i and j in the ensemble. Note that the fluctuation matrix F is basically equivalent to the Dynamics Fingerprint Matrix (DFM) used in Chapter 4; the only difference is that DFM used the standard deviation instead of the variance of distances.

Flexibility matrix

Although variance describes the spread of a distance distribution characterizing the relative fluctuation of two atoms, it is not always informative about how much the distance

between two atoms can change. Even if the distance of two atoms significantly deviates from their mean distance in some conformations, the variance may still be low provided that most of the variation are around the mean.

To measure the pairwise flexibility of two atoms (i.e. the maximal difference of their distance in the ensemble), the flexibility matrix denoted as X is introduced. Matrix X describes the range of distance distribution for each pair of atoms:

$$X(\mathbf{S})_{ij} := \max_k(d_{ij}^{S_k}) - \min_k(d_{ij}^{S_k}) \tag{5.4}$$

Note that the above definitions of F and X matrices allow that two pairs of atoms which have equal pairwise fluctuation can have considerably different pairwise flexibility.

A measure of overall fluctuation

While the F matrix contains pairwise atomic fluctuation values, a measure of the overall fluctuation of the whole structure (or a subset of residues) was also introduced. This overall fluctuation measure denoted by Θ was defined as the root mean square of dRMSD dissimilarity of each structure with regards the mean distance matrix calculated for the whole **S** ensemble.

In other words, Θ is a measure for the size of conformational space the protein explores in the ensemble. It is easy to see that the above definition is equivalent to the root mean square of the entries of F fluctuation matrix calculated for the same conformational ensemble. The precise definition of overall fluctuation is therefore

$$\Theta(\mathbf{S}) := \sqrt{\frac{1}{K}\sum_{k=1}^{K} dRMSD(S_k, \overline{S})^2} = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} F_{ij}} \tag{5.5}$$

where $\overline{S}$ is the mean distance matrix of the ensemble.

5.3.3 Molecular Dynamics simulations

All-atom 200 ns MD simulations were performed for the five apo PDZ domains summarized in Table 5.1. The simulations were run with the GROMACS MD simulation suite using the same settings as described in Section 4.3.1. All five proteins were simulated under the same conditions that makes the direct comparison of their dynamics feasible.

Simulation snapshots were saved at every 5 ps for analysis: a total number of 40000 frames were used from each PDZ domain trajectory. The above described fluctuation and flexibility measures were calculated based on the total set of 40000 frames.

5.3.4 Definition of binding site residues

By contrast the strategy used in Chapter 4, where dynamics data was used to create pairwise alignments (referred to as dynamics-based alignments), this study has relied on the opposite, more common strategy: prior sequence alignments were used to define equivalent positions in the 5 PDZ domains and the dynamics of the mapped residues were compared. (See a more detailed explanation of the two different comparative strategies in Section 2.5.4.) Since PDZ domains have highly conserved 3D structures, their prior sequence or structural alignments give an adequate mapping between their residue positions.

Multiple Sequence Alignment

A multiple sequence alignment (MSA) was created for the 5 PDZ sequences with ClustalW2, a general purpose MSA software using a progressive alignment construction method. (Thompson et al. 1994, Larkin et al. 2007) See the resulting multiple alignment in Figure 5.6. Although this alignment was constructed based on the primary sequences of the five proteins, it also defines a proper multiple structural alignment as shown by the low pairwise RMSD values of mapped residues (see Table 5.2). The average pairwise RMSD of the secondary structural elements of the five aligned structures is 2.1 Å indicating that the domain folds are well-conserved.

The multiple sequence alignment was used to define the subsets of binding site resi-

dues in the five domains. As discussed in Section 1.5.3, the binding groove of PDZ domains is located between the β2 strand and the α2 helix. Two sequence regions were therefore selected in the MSA that correspond to the conserved structural elements of the β2 strand and α2 helix (or α1 helix, in Erbin PDZ). The β2 region (5 residue positions) and the α2 region (10 residue positions) are highlighted in Figure 5.6 and are also shown in Table 5.1. The five binding pockets are structurally highly conserved as demonstrated by their low pairwise RMSD values (see Table 5.2). (The average pairwise RMSD of the five binding pockets is 1.1 Å .)

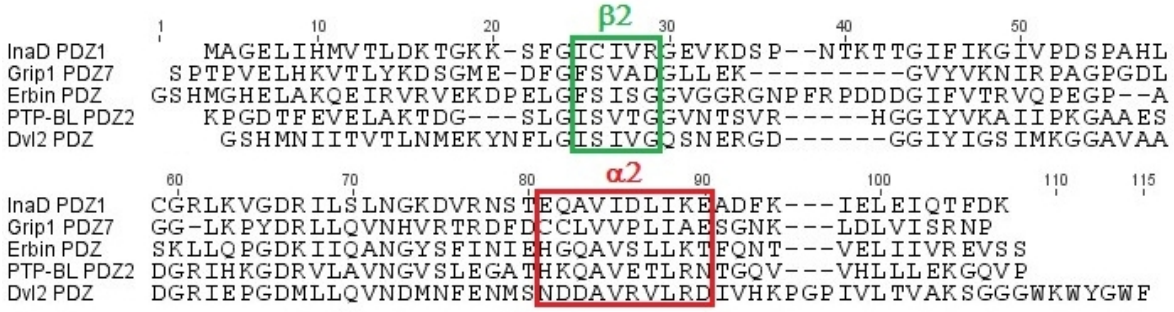

Figure 5.6: Multiple sequence alignment of the five PDZ domain sequences created with ClustalW2. The two selected regions corresponding to binding pocket residues in β2-strand and α2-helix are highlighted.

Binding pocket patterns

As the comparative analysis was focused on the five binding pockets, the local structural and dynamic characteristics were studied only for the binding site residues. In particular, the relative flexibility of the β2-strand and the α2-helix with regards to each other were analysed. The definitions of binding site patterns used in this study are defined here.

Let $I^{(\beta 2)}$ and $I^{(\alpha 2)}$ be the index sets of the two selected β2 and α2 sequence regions, respectively. Note that $|I^{(\beta 2)}| = 5$ and $|I^{(\alpha 2)}| = 10$. Furthermore, let M represent any of the above-mentioned NxN matrices (i.e. d, δ, F or X). The binding pocket pattern P based on the M matrix is then defined as

$$P_{ij} := M_{I^{(\beta 2)}(i) I^{(\alpha 2)}(j)} \tag{5.6}$$

P is therefore a 5x10 submatrix of the M matrix containing only those entries that describe the relations of binding site residues. For example, if M is the F fluctuation matrix, the binding pocket pattern describes the pairwise fluctuation values of the 5 β2-residues with regards to the 10 α2-residues. These local patterns calculated for the five PDZ domains can be directly compared as they characterize equivalent residues defined by the multiple sequence alignment.

5.3.5 Conformational clustering

k-mean clustering

MD simulation trajectory snapshots were clustered with k-mean cluster analysis, a simple unsupervised learning algorithm. (MacQueen 1967, Hartigan and Wong 1975). The method can be used for partitioning N data points (here, protein conformations) into k disjoint subsets (or clusters) denoted by $C_1, C_2, ..., C_k$. The parameter k is fixed a priori. The goal of the algorithm is to find the optimal partitioning of conformations to minimize the within-cluster sum of squares (WCSS):

$$WCSS := \sum_{i=1}^{k} \sum_{x_j \in C_i} dRMSD(x_j, \overline{C_i}) \tag{5.7}$$

where the dRMSD measure is used to capture the similarity of conformations and $\overline{C_i}$ is the mean distance matrix of cluster i.

Since k is an arbitrary parameter, the goodness of clustering results was estimated using the Silhouette Index cluster validity measure (see below) (Rousseeuw, 1987). The optimal k-value that provided the highest overall Silhouette Index was selected.

Silhouette Index

Once the conformational ensemble is clustered, the following Silhouette Index measure is calculated for each conformation:

$$S(i) := \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5.8}$$

where $a(i)$ is the average dRMSD dissimilarity of conformation i to all other conformations in the same cluster and $b(i)$ is the minimum of average dRMSD dissimilarities of conformation i to all other clusters.

The silhouette index is between -1 and 1: if S(i) it is close to 1, it means, the conformation is well-clustered; if S(i) is close to 0, it means the conformation could be assigned to another cluster as well; if S(i) is close to -1, it means the conformation is misclassified.

The goodness of clustering result was then measured by the overall average silhouette index S_{over} which is simply the average of S(i) for all conformations in the ensemble:

$$S_{over} := \frac{1}{N} \sum_{i=1}^{N} S(i) \tag{5.9}$$

5.3.6 Classical multidimensional scaling

As described in Section 2.2.2, Multidimensional Scaling (MDS) (also known as Principal Coordinates Analysis) is a dimensionality reduction method often used to visualize high-dimensional data on a two-dimensional map. (Borg and Groenen 2005) The input of the method is a dissimilarity matrix that contains distances (dissimilarities) between pairs of objects calculated in a high-dimensional space. The output is a configuration of points embedded into lower (ideally, two or three)-dimensions. In Classical Multidimensional Scaling (CMDS) (also referred to as Torgerson-Gower scaling) (Torgerson 1952), the goal is that the Euclidean distances between the outputted points should approximately reproduce the original dissimilarity matrix. This is achieved by minimizing a cost function using matrix eigendecomposition. The CMDS method was used in this study to project conformational ensembles onto a 2-dimensional map for the purpose of visualization. The dRMSD dissimilarity matrix of conformations was taken as the input of the algorithm.

5.3.7 Neighbouring conformers

In order to study the difference between induced fit and conformational selection binding, a simple definition is introduced to measure how similar conformations are sampled in an apo simulation to a given experimental ligand-bound structure. Let $S^{(k)}$ denote the set of k

most similar conformations (neighbouring conformers) with regards to a reference experimental structure E (ranked based on the dRMSD dissimilarity measure). The following $Q^{(k)}$ value is defined as the average dRMSD dissimilarity of conformations in $S^{(k)}$ with regards to structure E:

$$Q^{(k)} := \frac{1}{k} \sum_{x \in S^{(k)}} dRMSD(x, E) \tag{5.10}$$

In this study the quantities $Q^{(1)}$, $Q^{(10)}$, $Q^{(100)}$ and $Q^{(200)}$ were used to characterize the similarity of the most similar, 10 most similar, 100 most similar and 200 most similar conformations to an experimental ligand-bound structure of interest.

Furthermore, the set $S^{(100)}$ was used to characterize the exact locations of structural differences between the 100 most similar conformations and the reference conformer. For this purpose, the mean absolute difference distance matrix is calculated between each conformation in $S^{(100)}$ and the reference structure:

$$\Delta := \frac{1}{100} \sum_{x \in S^{(100)}} |\delta(x, E)| \tag{5.11}$$

Entries of the Δ matrix that are close to zero indicate pairs of atoms which have similar distance in the conformers of $S^{(100)}$ than in the ligand-bound structure.

5.4 Results

5.4.1 Diverse flexibility properties of the 5 binding pockets

To compare the inherent flexibility of the five PDZ binding pockets, the Θ overall fluctuation measure was calculated for the five conformational ensembles of the 200 ns MD simulation trajectories (40000 snapshots for each domain). The Θ fluctuation values of the five binding pockets (i.e. the five sets of binding site residues defined by the multiple sequence alignment) are summarized in Table 5.3. As discussed in Methods, the Θ measure shows the size of conformational space the binding pocket explores in the simulation.

In order to assess the significance of these results, a statistical framework has been developed. To this end, each of the five 200 ns MD simulations were cut into 1 ns non-

Name	$\Theta_{Binding\ pocket}$
InaD PDZ1	0.077
Dvl2 PDZ	0.071
Grip1 PDZ7	0.06
PTP-BL PDZ2	0.047
Erbin PDZ	0.038

Table 5.3: Overall fluctuation measure calculated for the five PDZ binding sites based on the conformational ensembles of the 200 ns MD trajectories.

overlapping, consecutive segments. For each PDZ domains, every possible pairs of 1 ns segments were compared by determining the absolute difference of the overall binding site fluctuation (Θ) measure calculated for the two different segments. These absolute difference values have been collected in a distribution that was used as a reference background distribution to describe how well Θ was converged.

Since even for the same PDZ domain, the values of Θ calculated for different MD conformational ensembles are never exactly the same, it should be assessed whether the differences we see between different PDZ domains are significant or within thermal noise. Therefore the spread of the reference background distribution was analysed to calculate p-values of the absolute Θ differences found between different PDZ domains.

Our results showed that the differences detected between the overall binding site fluctuations of different PDZ domains are statistically highly significant. For example, the Θ difference between InaD PDZ1 and GRIP1 PDZ7 has a p-value of 0.0212, while the significance of the difference found between InaD PDZ1 and PTP-BL PDZ2 is a p-value of 0.002. Additional examples are the differences between Dvl2 PDZ and PTP-BL PDZ2 or Erbin PDZ that have p-values of 0.0045 and 0.0015, respectively, or the difference between GRIP1 PDZ7 and Erbin PDZ that has a p-value of 0.022.

As the above is estimated from 1 ns simulation segments, this could be considered a lower limit as longer simulation times (for example from the whole 200 ns) period are likely to exhibit greater convergence.

Interestingly, despite the high structural similarity of the five binding sites (Table 5.2), one can see large differences in the extent of their intrinsic fluctuation. We find that InaD PDZ1 and Dvl2 PDZ have the most flexible binding pockets, while the binding site of Erbin PDZ is the most rigid of those of the five PDZ domains. The Θ value of Dvl2 PDZ is almost twice as large as of Erbin PDZ. These results are in good agreements with the conclusions of experimental studies which have found that Erbin PDZ binding site is rigid showing little structural variability while Dvl2 PDZ binding site is flexible showing large structural variability. The results suggest that the rigidity/flexibility of these binding sites demonstrated in other studies by comparison of apo and holo crystal structures can be explained by the intrinsic dynamics of the apo proteins.

The reason of the large overall fluctuation value of InaD PDZ1 is further investigated below. It is also interesting to note that Grip1 PDZ7 has larger Θ value than Erbin PDZ despite the fact that the base of its binding pocket is closed. The reason of this is also explained below.

5.4.2 Erbin PDZ and Dvl2 PDZ: rigid vs. flexible binding site

Fluctuation and flexibility patterns

As we see, the binding pocket of Dvl2 PDZ is significantly more flexible than the binding pocket of Erbin PDZ. In order to perform a more detailed comparison of the dynamics of the two PDZ domains, the binding pocket patterns based on the fluctuation (F) and flexibility (X) matrices were calculated. As described in Methods, binding pocket patterns are 5x10 submatrices characterizing pairwise relations of binding site residues. Figure 5.7 shows the patterns of fluctuation and flexibility for Erbin PDZ and Dvl2 PDZ domains.

We can see that the fluctuation patterns of the two domains (Figure 5.7A and B) are remarkably different. The Erbin PDZ pattern shows low pairwise fluctuation values of the binding site residues. Only residues 334Phe and 335Ser (at the N-terminal of β2-strand) are slightly mobile with regards to 396Lys (located at the C-terminal of α1-helix). Overall, the binding pocket of Erbin PDZ forms a rather rigid structure.

By contrast, the Dvl2 PDZ pattern shows large pairwise fluctuations between the binding site residues. In particular, the residues 14Ile and 15Ser (located at the N-terminal end of β2-strand) have prominent fluctuation with regards to the entire α2-helix. The largest mobility is observed between 14Ile-15Ser with regards to 68Val-69Arg (located at the middle of α2-helix). The considerable fluctuation of 14Ile and 15Ser relative to the α2-helix indicates that it is the top portion of the binding groove that undergoes an opening-closing motion. Similar conformational changes of the binding site have been reported for other PDZ domains.[55,404] Overall, the binding pocket of Dvl2 PDZ is much more flexible than that of Erbin PDZ.

The flexibility pattern of Erbin PDZ (Figure 5.7D) shows that the top portion of the Erbin PDZ binding site is more flexible than suggested by the fluctuation pattern alone. For example, the large pairwise flexibility value between 335Ser (located at the N-terminal end of β2-strand) and 396Lys (located at the C-terminal end of α1-helix) indicates that the distance of the two residues changes within a wide interval. Their distance distribution shows that the top part of the binding site occasionally opens up considerably more, but these extreme conformers are very infrequent and the binding site behaves as a rigid structure most of the time. Similarly, in the case of the Dvl2 PDZ domain, we see residue pairs that have large pairwise flexibility despite low pairwise fluctuation (e.g. 18Gly with regards to 68Val). These examples illustrate that fluctuation and flexibility patterns provide complementary measures for studying protein dynamics.

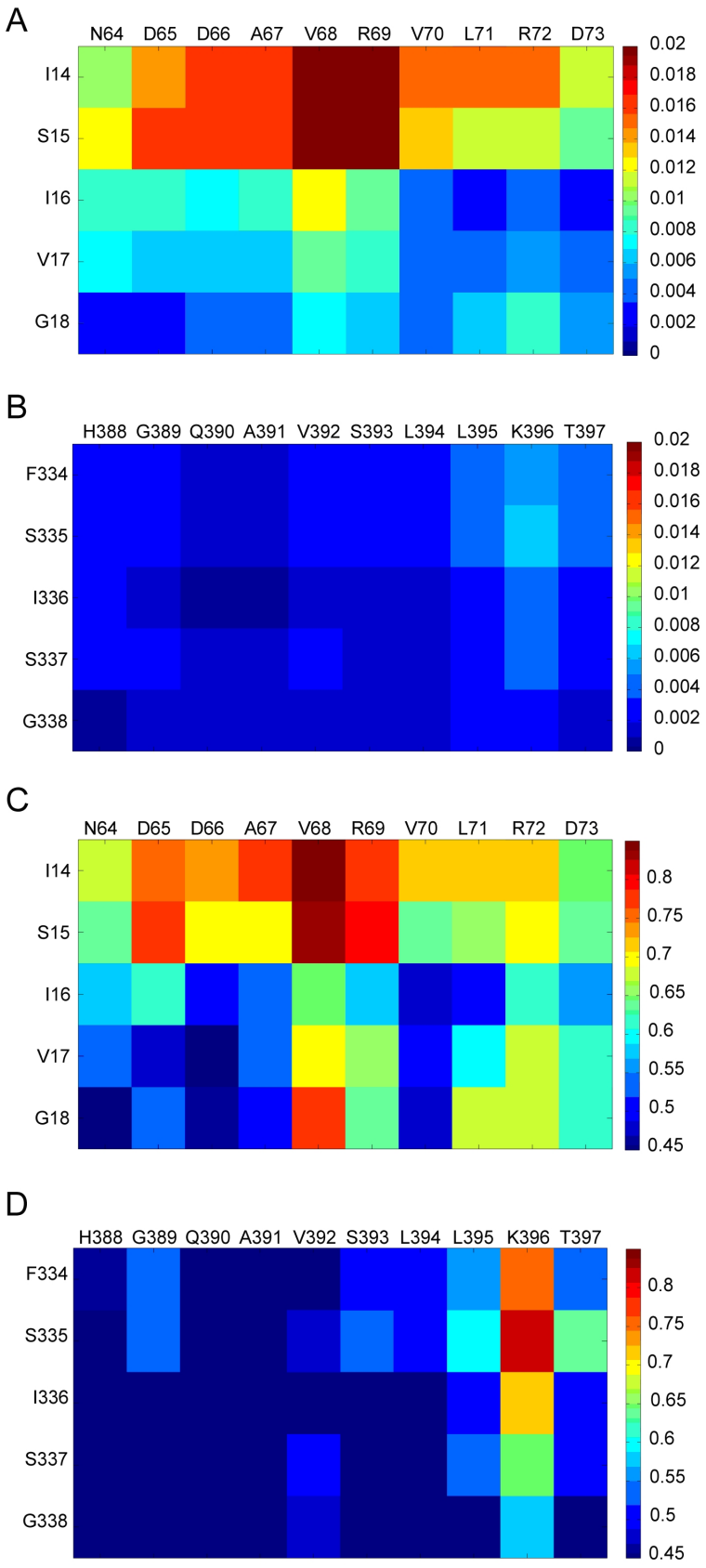

Figure 5.7: Comparison of the fluctuation and flexibility binding pocket patterns of Dvl2 PDZ and Erbin PDZ: **A.** Fluctuation (F) matrix pattern of the Dvl2 PDZ binding site; **B.** Fluctuation (F) matrix pattern of the Erbin PDZ binding site; **C.** Flexibility (X) matrix pattern of the Dvl2 PDZ binding site; **D.** Flexibility (X) matrix pattern of the Erbin PDZ binding site.

Flexibility pattern of experimental structures

To understand what roles the intrinsic dynamics of the Dvl2 PDZ domain plays in ligand binding, the fluctuation and flexibility properties of the binding pocket have been compared to the structural differences seen between experimental apo and ligand-bound structures. For this purpose, an experimental conformational ensemble was assembled including a crystal structure of the apo Dvl PDZ domain (PDB: 2rey) and four crystal structures of different ligand-bound conformations (PDB: 3cbx, 3cby, 3cbz and 3cc0). One of the four ligand bound-structures (referred to as pep-C1) represents C-terminal ligand family, while the other three structures (referred to as pep-N1, pep-N2 and pep-N3) represent three different internal peptide ligand families (Zhang et al 2009).

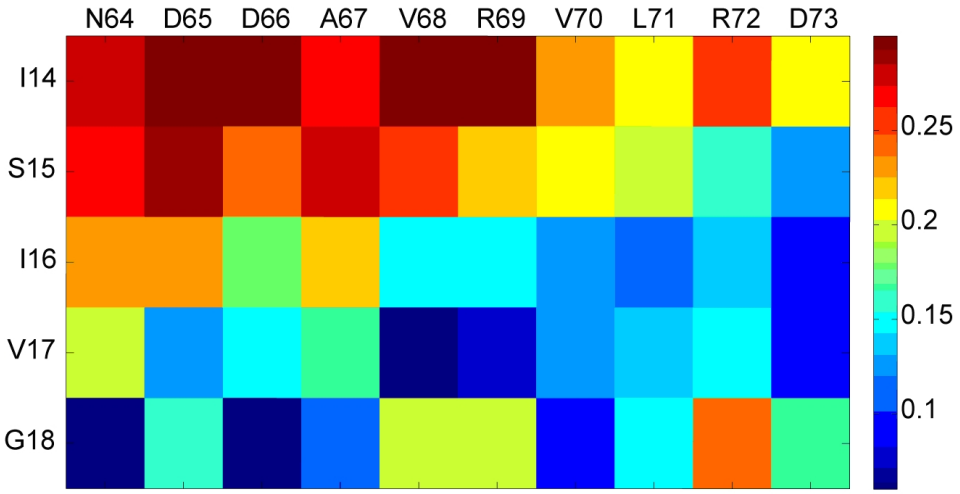

Figure 5.8: Flexibility matrix calculated based on the experimental ensemble of Dvl2 PDZ. The used ensemble consists of five crystal structures: an apo (PDB: 2rey) and 4 ligand-bound (PDB: 3cbx, 3cby, 3cbz and 3cc0) structures.

The flexibility matrix (X) pattern of the binding pocket was calculated based on the experimental ensemble (see Figure 5.8) to characterize the differences between the crystal structures. This pattern shows us which binding pocket residue pairs have the largest relative displacement between the apo and ligand-bound structures. This experimental flexibility pattern is markedly similar to the fluctuation and flexibility patterns which are based on the apo MD simulation (Figure 5.7A and C). The correlation with the fluctuation

and flexibility patterns of the simulation are 0.74 and 0.68, respectively.

Shown by Figure 5.8, in the experimental ensemble the largest displacements are seen for residues 14Ile and 15Ser with regards to the α2-helix. As discussed in the previous section, the most prominent motions were observed for the same two residues relative to the α2-helix. The similarity of intrinsic conformational dynamics and structural deformations observed upon ligand-binding suggest that the spontaneous fluctuations of the apo structure do play some role in peptide recognition. The intrinsic dynamics may be important either for actually visiting the bound conformations or could just be an indicator for that the structure is predisposed for conformational changes induced by the ligands.

Cluster analysis and multidimensional scaling

To study how the MD snapshots are distributed in the conformational space, the apo Dvl2 PDZ simulation ensemble was clustered and was projected to a 2-dimensional map using the classical multidimensional scaling (MDS) method. Every 10th snapshots (taken with 50 ps frequency from the trajectory) were included in the analysis. The input of clustering and MDS was therefore a 3981x3981 matrix containing the pairwise dRMSD dissimilarity values of the 3981 conformers.

Groups of similar conformers were identified with the k-mean cluster analysis algorithm and clustering results were validated with the silhouette index measure. The optimal number of clusters corresponding to the maximal overall average silhouette index ($S_{over} = 0.411$) was found to be 2. The two large conformational clusters identified that contain 1512 and 2469 conformers, will be referred to as Cluster 1 and 2, respectively.

Figure 5.9 summarizes the results of cluster analysis and multidimensional scaling. Figure 5.9A presents the dRMSD dissimilarity matrix of the 3981 conformers rearranged according to the clustering results (i.e. conformers belonging to the same cluster are arranged next to each other in the matrix). We can see low dissimilarity values between conformations that belong to the same cluster and high dissimilarity values between conformations belonging to different clusters.

Figure 5.9B shows the results of multidimensional scaling: conformations represented

by dots are visualized on a 2-dimensional map in which similar conformers are located close to each. The colors of the dots represent the corresponding silhouette index values: the higher this value is, the more we are assured that the conformer had been labelled to the correct cluster. We can see that while conformations in the cores of the clusters have high silhouette indices, conformers found at the interface of the two clusters have low silhouette values as they could be classified to the other cluster too. The rearranged dRMSD matrix, the MDS map and the overall average silhouette index indicates that the conformational ensemble is split to two distinct, contiguous clusters.

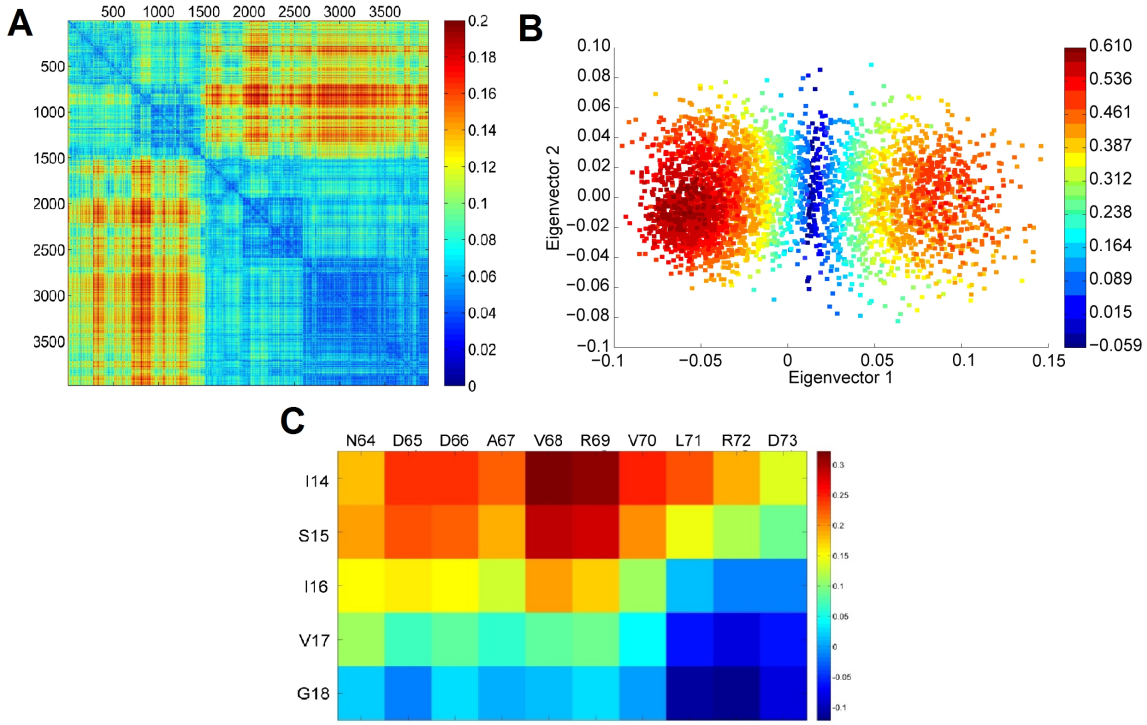

Figure 5.9: Result of cluster analysis of the Dvl2 PDZ conformational ensemble consisting of 3981 MD simulation snapshots. **A:** dRMSD dissimilarity matrix of the 3981 conformers rearranged in a way that rows and columns corresponding to conformers that belong to the same cluster are next to each other in the matrix; **B:** Multidimensional Scaling results visualizing the conformational ensemble on a 2-dimensional map. Colors represent silhouette index values. (The overall average silhouette index is 0.411); **C:** Difference distance matrix (δ) calculated between the representative medoid conformations of the two clusters.

Finally, Figure 5.9C shows the difference distance matrix (δ) calculated between the representative conformations (medoids) of the two clusters. We can see that the largest

structural differences between the two cluster medoids are the deviations of 14Ile and 15Ser with regards to the α2-helix. In other words, the motion of 14Ile and 15Ser relative to the α2-helix (indicated by the fluctuation matrix pattern; Figure 5.7A) corresponds to the transition of the protein from one conformational cluster to another. These results suggest that the conformational dynamics of the Dvl2 PDZ binding pocket is determined by two distinct basins in the energy landscape.

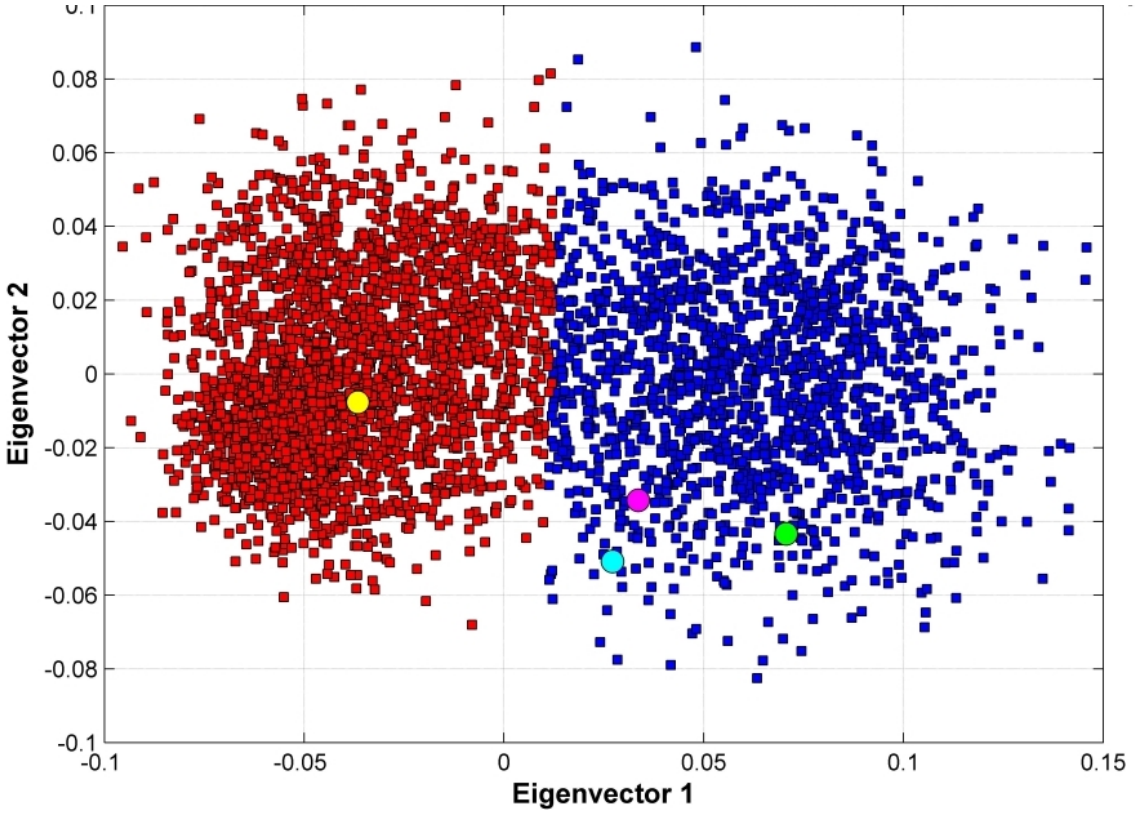

Figure 5.10: Multidimensional Scaling of the extended conformational ensemble of the Dvl2 PDZ domain containing the 3981 MD simulation snapshots and the 4 experimental ligand-bound conformers (pep-C1, pep-N1, pep-N2 and pep-N3). Blue dots represent conformations that belong to Cluster 1, red dots represent conformations that belong to Cluster 2. The 4 crystal structures: pep-C1, pep-N1, pep-N2 and pep-N3 are shown in magenta, cyan, yellow and green color, respectively.

To further investigate the relationship between the intrinsic dynamics and ligand binding of Dvl2 PDZ, the location of the four ligand-bound crystal structures in the conformational space explored by the apo MD simulations was studied. For this purpose, the multidimensional scaling was recalculated for an extended conformational ensemble that

contained the 3981 MD simulation snapshots and the 4 experimental ligand-bound conformers discussed in the previous subsection (pep-C1, pep-N1, pep-N2 and pep-N3). The resulting 2-dimensional map showing the MD simulation ensemble and the 4 crystal structures together is presented in Figure 5.10. Conformations belonging to Cluster 1 and 2 are shown in blue and red color, respectively. The four experimental structures are highlighted in different colors (pep-C1: magenta; pep-N1: cyan; pep-N2: yellow and pep-N3: green).

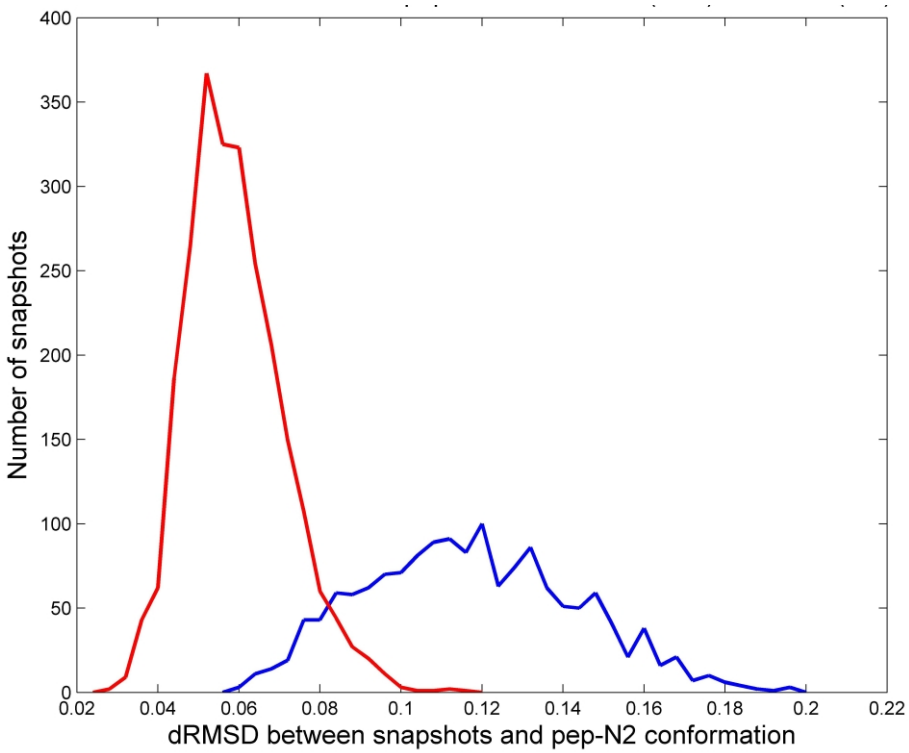

Figure 5.11: Distribution of dRMSD dissimilarity values between MD simulation snapshots in the two different clusters and the pep-N2 ligand-bound conformation. The red curve depicts the distribution corresponding to Cluster 2, while the blue curve depicts the distribution of Cluster 1.

We can see that the MDS analysis has placed pep-C1, pep-N1 and pep-N3 conformers in Cluster 1, while pep-N2 has been allocated to the middle of Cluster 2. In other words, all simulation snapshots that are similar to the pep-C1, pep-N1 and pep-N3 ligand-bound conformations appear to belong to Cluster 1, while snapshots similar to the pep-N2 conformer belong to Cluster 2. These results are also confirmed by the distributions of dRMSD dissimilarity values between the snapshots in the two different clusters and the pep-N2

and pep-N3 conformations (see Figure 5.11 and 5.12).

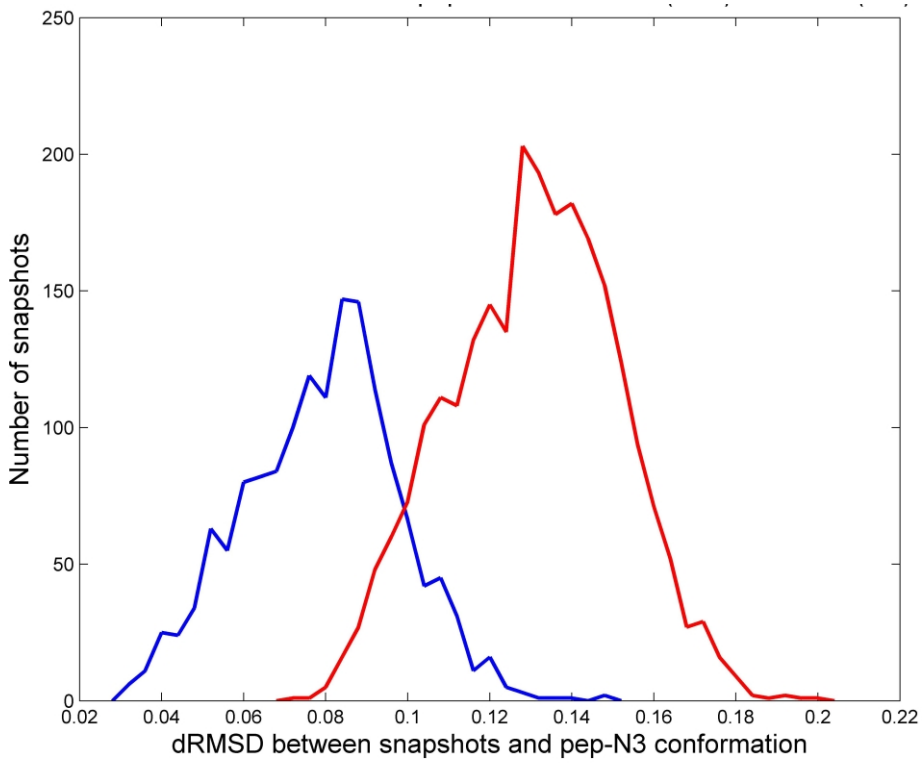

Figure 5.12: Distribution of dRMSD dissimilarity values between MD simulation snapshots in the two different clusters and the pep-N3 ligand-bound conformation. The blue curve depicts the distribution corresponding to Cluster 1, while the red curve depicts the distribution of Cluster 2.

Two main conclusions can be made based on the MDS results. Firstly, binding pocket conformations similar to each of the four experimental ligand-bound structures are visited in the MD simulation. Secondly, the relatively large conformational space explored by the apo Dvl2 PDZ binding pocket (i.e. compared to the size of conformational space visited by the Erbin PDZ binding site) is required for incorporating all four experimental binding site conformers. In other words, it is likely that visiting Cluster 1 is necessary for the domain to form the pep-C1, pep-N1 and pep-N3 complexes, while visiting Cluster 2 is required for forming the pep-N2 complex.

Figure 5.13 shows the superposition of the binding sites of pep-N2 and pep-N3 conformers with regards to the centers (medoid conformations) of the two clusters. The pep-N3 binding site is markedly more similar to the center of Cluster 1 than to the medoid

of Cluster 2. On the other hand, the pep-N2 binding pocket align much better with the medoid of Cluster 2 than with the center of Cluster 1. In addition, Figure 5.13 shows the main structural differences between the two cluster medoids that is already discussed above: the displacement of 14Ile and 15Ser (at the N-terminal end of β2-strand) relative to α2-helix. Since the same structural differences are observed between the pep-N2 and pep-N3 binding sites, the intrinsic fluctuations of 14Ile and 15Ser seem to be necessary for accessing both conformations.

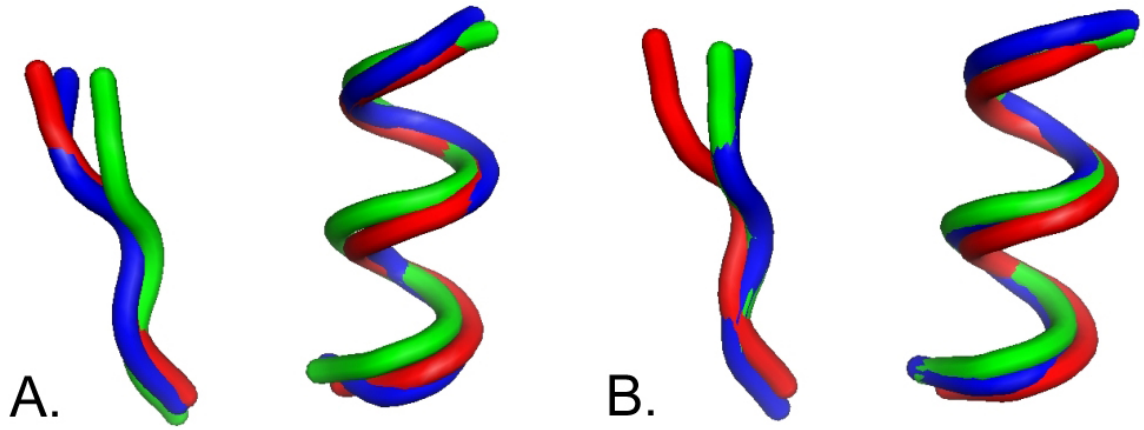

Figure 5.13: Comparison of the binding sites of the ligand-bound conformations pep-N2 (shown in green) and pep-N3 (shown in red) to the medoid conformations of the two clusters. **A:** The two experimental binding sites are superposed to the binding pocket of the medoid conformation of Cluster 1 (shown in blue); **B:** The two experimental binding sites are superposed to the binding pocket of the medoid conformation of Cluster 2 (shown in blue).

To summarize, these results suggest that the high conformational flexibility of the Dvl2 PDZ binding site is closely related to its binding promiscuity. Its binding pocket explores significantly larger conformational space than the binding pocket of Erbin PDZ. The flexibility enables the domain to visit binding site conformers similar to a number of different ligand-bound conformations. Consequently, the Dvl2 PDZ domain can be characterised by the "strong" definition of promiscuity (described in Introduction) and its intrinsic dynamics seems to be the key for its promiscuous nature.

Although these results suggest that conformational selection is essential in peptide recognition of the Dvl2 PDZ domain, it is still an open question to what extent the induced

fit mechanism is involved as fine-tuning in the ligand binding process.

By contrast, the Erbin PDZ domain has a relatively rigid binding site. Figure 5.14 shows the 2-dimensional map resulted from Multidimensional Scaling of 3981 MD snapshots of the Erbin PDZ binding pocket extended by 2 experimental ligand-bound conformations which correspond to complexes with class I and class II peptides (PDB: 1n7t and 1mfg). (Note that the scales of MDS maps created for Erbin PDZ and Dvl2 PDZ are different, so cluster sizes are not comparable between Figure 5.10 and Figure 5.14.)

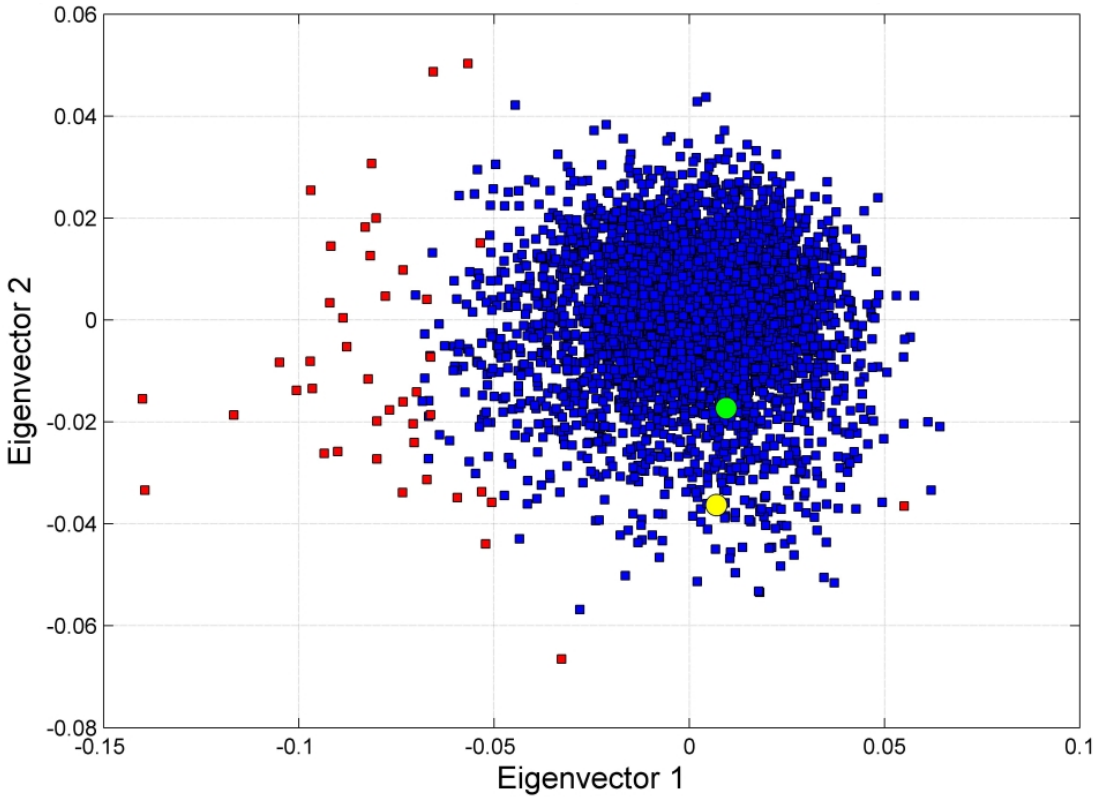

Figure 5.14: Multidimensional Scaling of the extended conformational ensemble of the Erbin PDZ domain containing the 3981 MD simulation snapshots and 2 experimental ligand-bound conformers (complexes with class I and class II peptides). Red dots represent outlier conformations that have dRMSD dissimilarity larger or equal than 0.8 Å from the medoid conformer. The two complexes with class I and class II peptides are shown in green and yellow color, respectively.

The MDS map of Erbin PDZ shows that the simulation ensemble forms a single conformational cluster, however, some conformers are distant from this main cluster. Outlier conformations (defined as snapshots with dRMSD dissimilarity larger or equal than 0.8 Å

from the medoid conformer) are highlighted in red color. The existence of these outlier conformers explain why the flexibility pattern of the Erbin PDZ binding site (Figure 5.7D) shows that the top portion of the binding pocket is much more flexible than suggested by the fluctuation pattern (Figure 5.7B).

We can also see that the MDS analysis has placed the two experimental peptide-bound conformations in the main cluster. In particular, the complex with the class I peptide is located close to the cluster center, while the complex with the class II peptide is placed closer to the edge of the cluster. Both ligand-bound conformers, however, are located within the limited conformational space explored by the binding pocket, suggesting that the conformational selection mechanism may be involved in ligand binding. As discussed in Introduction, the Erbin PDZ domain is promiscuous in the sense that it is able to bind to multiple peptides, however, it interacts with its partners using essentially the same binding mode. It therefore does not satisfy the "strong" definition of promiscuity defined here.

As the shape of binding site is almost the same regardless the ligand bound to it, the apo Erbin PDZ domain does not need to be flexible in order to visit these conformations. On the other hand, its rigidity could be beneficial as it may optimize its specificity towards ligands, similarly to antibodies for which rigidification has been shown to result in increased affinity (Thorpe and Brooks 2007) (see Introduction).

5.4.3 Analysis of InaD PDZ1, PTP-BL PDZ2 and GRIP1 PDZ7

The results highlighting interesting features of the binding sites of the 3 other PDZ domains studied here (InaD PDZ1, PTP-BL PDZ2 and GRIP1 PDZ7) are discussed in this subsection. Figure 5.15 shows the comparison of the binding pocket fluctuation and flexibility patterns of these PDZ domains. As in Figure 5.7, we see large differences between their fluctuation and flexibility properties. Note that the patterns in Figure 5.7 and Figure 5.15 are comparable as they use the same color scale.

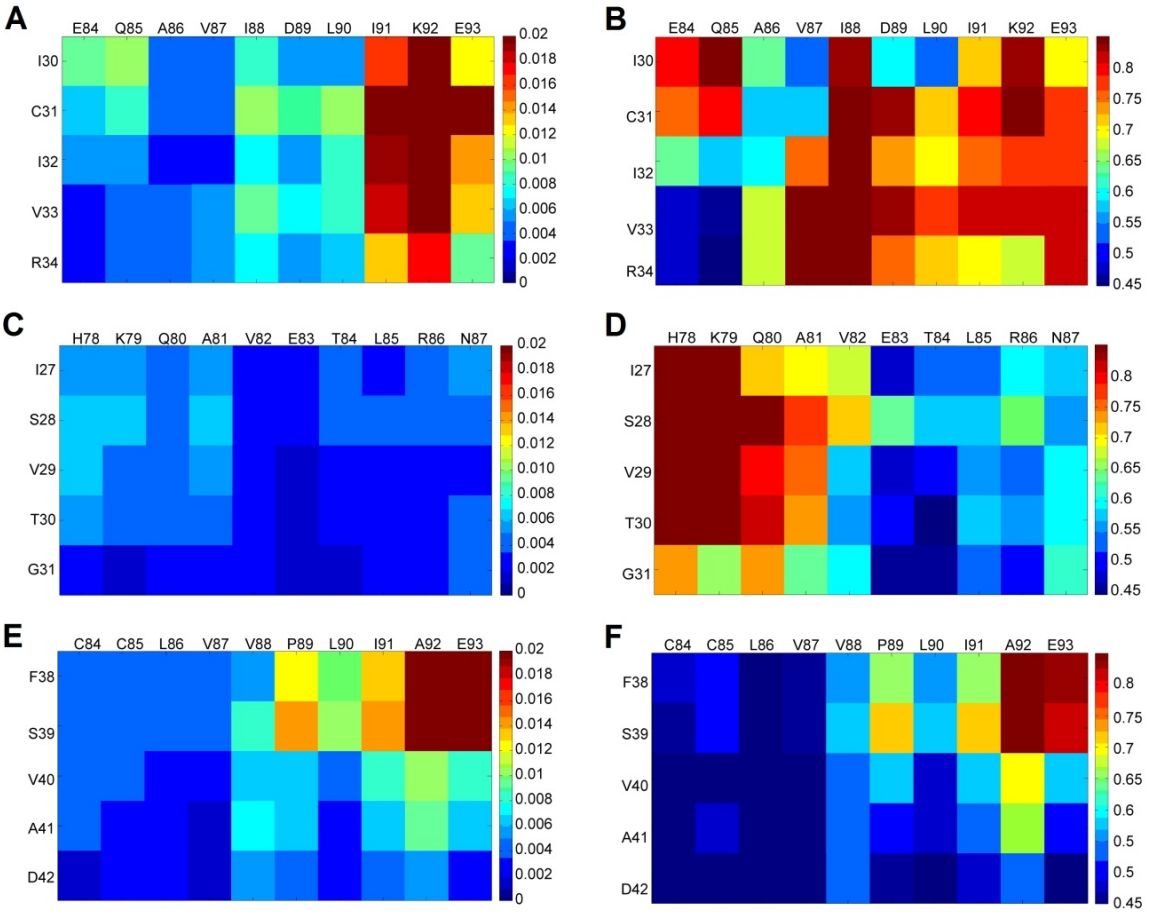

Figure 5.15: Comparison of the fluctuation and flexibility binding pocket patterns of InaD PDZ1, PTP-BL PDZ2 and Grip1 PDZ7: **A.** Fluctuation (F) matrix pattern of the InaD PDZ1 binding site; **B.** Flexibility (X) matrix pattern of the InaD PDZ1 binding site; **C.** Fluctuation (F) matrix pattern of the PTP-BL PDZ2 binding site; **D.** Flexibility (X) matrix pattern of the PTP-BL PDZ2 binding site; **E.** Fluctuation (F) matrix pattern of the Grip1 PDZ7 binding site; **F.** Flexibility (X) matrix pattern of the Grip1 PDZ7 binding site.

InaD PDZ1: flexibility as a predictor of promiscuity

As presented in Table 5.3, InaD PDZ1 has the most flexible binding site of the five compared PDZ domains. Figure 5.15A shows that the part of InaD PDZ1 domain that fluctuates the most is three residues at the C-terminal end of the α2-helix (91Ile, 92Lys and 93Glu) with regards to the entire β2-strand. On the other hand, the flexibility matrix pattern (Figure 5.15B) also shows that much larger region (almost the whole binding pocket) is very flexible.

Clustering the MD conformational ensemble of the apo domain (3981 snapshots of the 200 ns trajectory) identified two main clusters that are referred to Cluster 1 and 2. The two clusters contain 2114 and 1867 conformations, respectively. (The overall average silhouette index is 0.43.) Figure 5.16 presents the results of multidimensional scaling of the 3981 snapshots and the known experimental structure of the PDZ domain in complex with the NorpA peptide (PDB: 1ihj). Conformations belonging to Cluster 1 and 2 are visualized in blue and red colors, respectively. The figure shows that the division of the conformational ensemble into two clusters is not as clear as in the case of Dvl2 PDZ binding pocket.

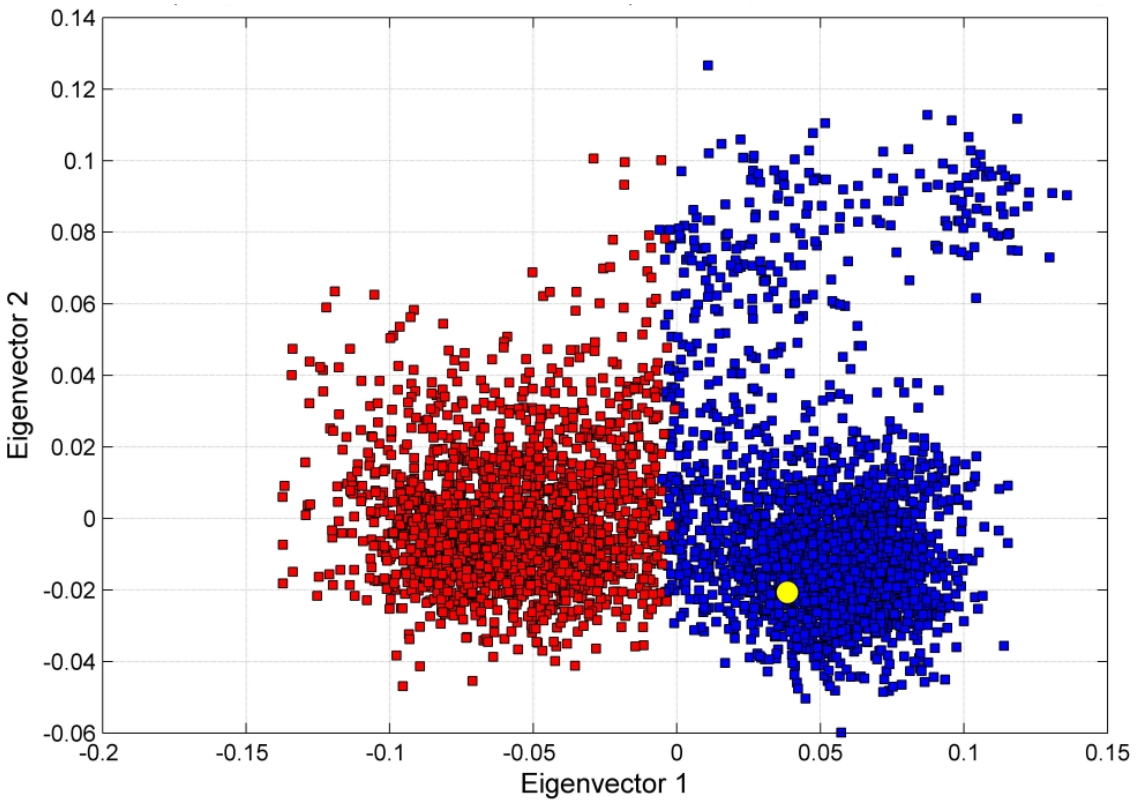

Figure 5.16: Multidimensional Scaling of the extended conformational ensemble of the InaD PDZ1 domain containing the 3981 MD simulation snapshots and the experimental ligand-bound conformer (complex with the NorpA peptide) which is shown in yellow. Blue dots represent conformations that belong to Cluster 1, red dots represent conformations that belong to Cluster 2.

One can see that the MDS analysis has placed the NorpA-bound conformer in Cluster 1. In other words, all simulation snapshots that are similar to the NorpA peptide-bound conformation appear to belong to Cluster 1. Therefore, exploring the conformational space

of Cluster 1 could be important for the InaD PDZ1 binding pocket to be able to form a complex with the NorpA peptide.

On the other hand, the reason of exploring Cluster 2 is not immediately clear. As we have seen for the Dvl2 PDZ domain, the conformational flexibility of the binding pocket is likely to be related to its promiscuity. The Dvl2 PDZ binding pocket also explores a relatively large conformational space consisting of two different conformational clusters that have been found to incorporate multiple ligand-bound conformers corresponding to different partners and different binding modes. This could also be the case for the InaD PDZ1 binding site. As discussed in Introduction, besides the NorpA peptide, the InaD PDZ1 domain has been shown to bind to the unconventional myosin NinaC. Moreover, experimental results suggest that InaD PDZ1 may interact with NinaC in a different mode than it does with NorpA (i.e. binding to an internal peptide sequence motif of NinaC). This hypothesis, however, has not been directly confirmed since no experimental structure of the complex of InaD PDZ1 and NinaC peptide is currently available.

If the hypothesis based on experimental data is true (i.e. InaD PDZ1 indeed interacts with NinaC in a different binding mode than with NorpA) and the conformational selection mechanism plays a prominent role in this interaction, the InaD PDZ binding pocket needs to be relatively flexible exploring the conformational space between the different ligand-bound conformers. This is exactly what we see in the results of MDS analysis. While visiting Cluster 1 may be important for the binding pocket to interact with the NorpA peptide, visiting Cluster 2 may be essential for interacting with the NinaC peptide.

To summarize, the InaD PDZ1 domain is likely to satisfy the "strong" definition of promiscuity as it can probably bind to different partners using considerably different binding modes. We see again that this property is correlated with the conformational flexibility of the binding pocket, such as in the case of the Dvl2 PDZ domain. Based on the results of multidimensional scaling, it is predicted here that visiting the conformational space corresponding to Cluster 2 is necessary for the InaD PDZ1 binding site to be able to interact with NinaC. In other words, it is predicted that the ligand-bound conformer of the complex of

InaD PDZ1 with NinaC would be placed in Cluster 2 by the MDS analysis. This hypothesis could easily be tested once the structure of the complex is solved experimentally.

Furthermore, these results also support the earlier anticipation of Kimple et al. and Wes et al. that InaD PDZ1 binds to NorpA and NinaC using different binding modes.[392,396] Based on the fluctuation pattern in Figure 5.15A, the main structural difference between the NorpA and NinaC peptide-bound conformers is predicted to be the shift of the C-terminal end of the α2-helix (91Ile, 92Lys and 93Glu) with regards to the β2-strand.

PTP-BL PDZ2: peptide binding by induced fit

The fluctuation pattern of PTP-BL PDZ2 (Figure 5.15C) shows that this domain has a considerably rigid binding site, similarly to Erbin PDZ domain. However, the flexibility pattern (Figure 5.15D) reveals that the N-terminal end of α2-helix is flexible (even if does not fluctuate much) with regards to the β2-strand. The result of multidimensional scaling of the 3981 simulation snapshots and the experimental ligand-bound conformer corresponding to a complex with the APC peptide (PDB: 1vj6) is presented in Figure 5.17. The majority of conformations appear to be distributed within a single compact cluster, however a large number of outlier conformations are also observed that are considerably different from those belonging to the main cluster. Outlier conformations defined as snapshots with dRMSD dissimilarity larger or equal than 0.9 Å are highlighted in red color. These outlier conformers explain why the flexibility pattern of the PTP-BL PDZ2 binding site shows that the N-terminal end of α2-helix is more flexible with regards to the β2-strand than suggested by the fluctuation binding pocket pattern.

Although the MDS analysis has placed the experimental ligand-bound conformer in the main conformational cluster, this does not exclude the possibility that induced fit plays essential role in the binding process. As discussed in Introduction, experimental evidences suggest that PTP-BL PDZ2 binds to the APC peptide through an induced fit mechanism. In order to investigate this, the structural differences between the APC-bound conformation and the most similar (neighbouring) conformers sampled in the apo MD simulation have been characterised using the $Q^{(1)}$, $Q^{(10)}$, $Q^{(100)}$ and $Q^{(200)}$ values defined in Methodology.

These measures describe the similarity of the most similar, 10 most similar, 100 most similar and 200 most similar binding pocket conformations to the experimental ligand-bound structure of interest.

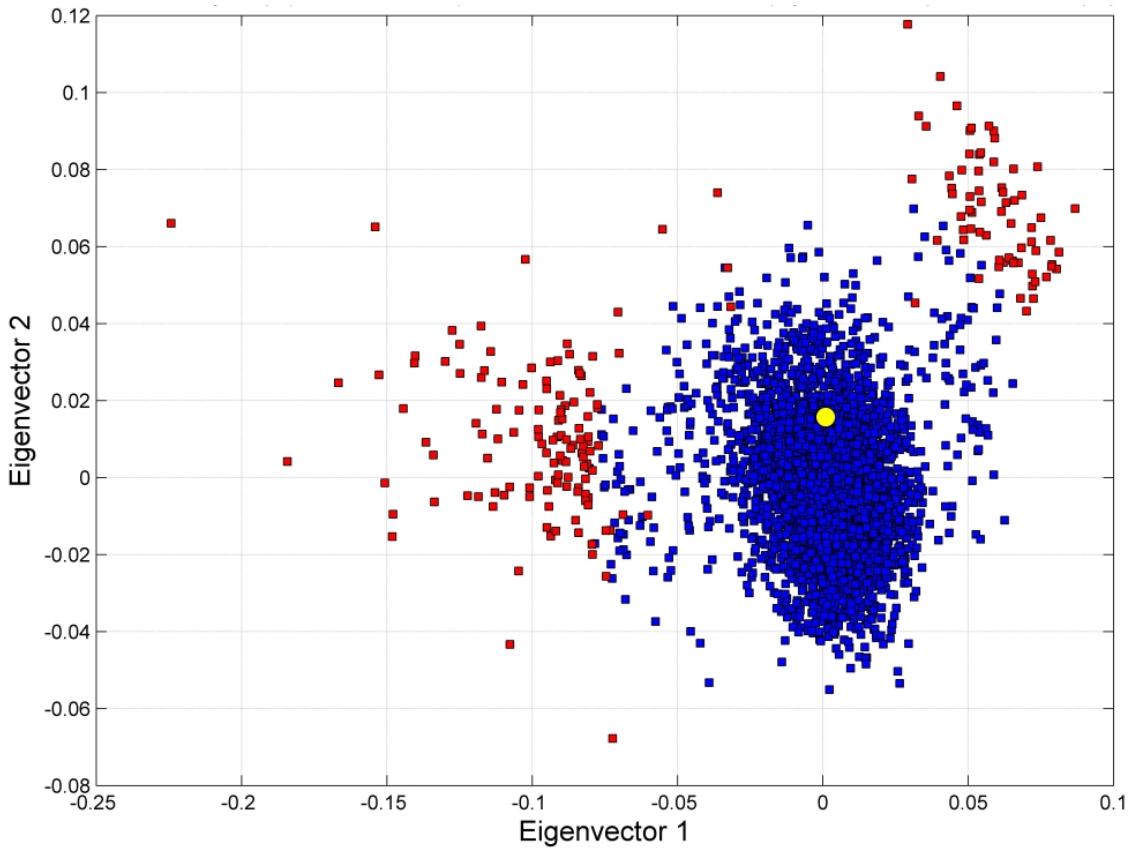

Figure 5.17: Multidimensional Scaling of the extended conformational ensemble of the PTP-BL PDZ2 domain containing the 3981 MD simulation snapshots and the experimental ligand-bound conformer corresponding to a complex of PTP-BL PDZ2 with the APC peptide which is shown in yellow. Red dots represent outlier conformations that have dRMSD dissimilarity larger or equal than 0.9 Å from the medoid conformer.

Table 5.4 gives a summary of the $Q^{(1)}$, $Q^{(10)}$, $Q^{(100)}$ and $Q^{(200)}$ values calculated for each peptide-bound conformation of each of the five PDZ domains simulated in this study. We see that, out of the five PDZ domains, the complex of PTP-BL PDZ2 domain with the APC peptide is the less similar to the apo MD simulation ensemble. It has the highest $Q^{(1)}$ and $Q^{(10)}$ values (0.37 Å and 0.39 Å which represent the average dRMSD dissimilarity between the ligand-bound conformer and the most similar and ten most similar simulation snap-

shots, respectively. In other words, the structure of the ligand bound PTP-BL PDZ2 binding site is the most distant from the conformational space sampled in the MD simulations. By contrast, for example, the complex of InaD PDZ1 with the NorpA peptide has significantly lower $Q^{(1)}$ and $Q^{(10)}$ values (0.15 Å and 0.18 Å indicating that this ligand-bound binding site is accessed much closer in the apo MD simulation.

Complex	$Q^{(1)}$	$Q^{(10)}$	$Q^{(100)}$	$Q^{(200)}$
Erbin PDZ : Class I peptide	0.17 Å	0.18 Å	0.21 Å	0.22 Å
Erbin PDZ : Class II peptide	0.21 Å	0.26 Å	0.31 Å	0.33 Å
Dvl PDZ : pep-C1 peptide	0.29 Å	0.32 Å	0.36 Å	0.41 Å
Dvl PDZ : pep-N1 peptide	0.3 Å	0.36 Å	0.44 Å	0.48 Å
Dvl PDZ : pep-N2 peptide	0.19 Å	0.25 Å	0.31 Å	0.33 Å
Dvl PDZ : pep-N3 peptide	0.28 Å	0.3 Å	0.34 Å	0.36 Å
InaD PDZ1 : NorpA peptide	0.15 Å	0.18 Å	0.21 Å	0.23 Å
PTP-BL PDZ2 : APC peptide	**0.37Å**	**0.39 Å**	**0.43 Å**	**0.44 Å**

Table 5.4: Mean dRMSD dissimilarity between the ligand-bound conformations and the most similar, 10 most similar, 100 most similar and 200 most similar snapshots of the apo MD simulations (see more detailed explanation of the $Q^{(1)}$, $Q^{(10)}$, $Q^{(100)}$ and $Q^{(200)}$ measures in Methods).

Due to the possibly incomplete sampling of these simulations, we are unable to tell if the apo structures get any closer to the peptide-bound conformations in reality. However, based on the data presented here, we can say that out of the five PDZ domains studied, PTP-BL PDZ2 is the most likely to involve induced fit mechanism when binding to the APC peptide. Figure 5.18. shows the mean absolute difference distance matrix (Δ) pattern calculated between the peptide-bound structure and the 100 most similar snapshots. We can see that the largest deviations are found in the distances between 28Ser and 85Leu and between 29Val and 86Arg. The Δ pattern suggests that these two inter-residue distances are altered the largest extent upon binding to the APC peptide.

Although these results support the conclusion of experimental studies that induced fit is involved in the binding of PTP-BL PDZ2 to APC, it also seems likely that the induced fit

mechanism may play important role in other PDZ-peptide interactions as well. Conformational selection and induced fit both appear to be essential in the binding of PDZ domains to their peptides. Firstly, conformational selection seems to be an essential mechanism for PDZ domains to visit regions of the conformational space that are close to different ligand-bound states. Visiting these regions are probably necessary for the formation of initial complexes. Secondly, once an initial complex is formed, the induced fit mechanism, as a fine-tuning step, could lead to minor changes in the shape of the binding pocket stabilizing the PDZ-peptide complex.

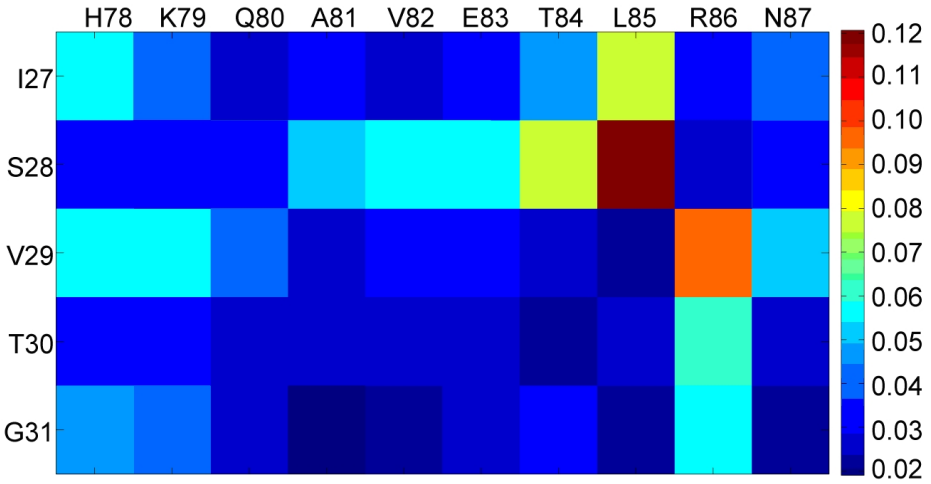

Figure 5.18: Mean absolute difference distance matrix (Δ) pattern calculated between the APC peptide-bound conformation of PTP-BL PDZ2 and the 100 most similar MD simulation snapshots.

GRIP1 PDZ7: a closed carboxyl peptide binding pocket

As discussed in Introduction, the $\alpha 2/\beta 2$ binding pocket of the GRIP1 PDZ7 domain adopts a closed conformation and is probably unable to interact with a carboxyl peptide. This prediction, however, is based on the solution structure of the domain which shows that GRIP1 PDZ7 has a significantly smaller carboxyl peptide binding site than other PDZ domains. On the other hand, some other PDZ domains are known to have similarly closed binding pockets that are however able to open up in order to incorporate a peptide ligand (e.g. the LARG PDZ domain: see Introduction). Here the conformational ensemble generated by the MD simulation of GRIP1 PDZ7 was used to study whether the apo PDZ domain is able

to undergo such conformational transition opening its carboxyl peptide binding site.

The fluctuation and flexibility patterns of the GRIP1 PDZ7 binding pocket (Figure 5.15E and F) show that the N-terminal end of β2-strand has notable fluctuation with regards to the C-terminal end of α2-helix. On the other hand, the patterns also show that the C-terminal end of β2-strand has little mobility with regards to the N-terminal end of α2-helix. Since the bottom of the binding pocket is located between the C-terminal end of β2-strand and the N-terminal end of α2-helix, their low relative fluctuation suggests that the base of the binding site does not open significantly.

In order to study this question in more details, the distance between the C-terminal residue of β2-strand and the N-terminal residue of α2-helix is used to characterize to what extent the base part of the binding pocket is open. The equivalent residue pairs whose distance is calculated are 338Gly and 388His in Erbin PDZ, 18Gly and 64Asn in Dvl2 PDZ, 34Arg and 84Glu in InaD PDZ1, 31Gly and 78His in PTP-BL PDZ2 and 41Ala and 84Cys in GRIP1 PDZ7. Figure 5.19 presents the distance distribution based on the total set of MD simulation snapshots except the first 1 ns (39801 conformations) for each PDZ domain.

A number of interesting observations can be made comparing the five distance distributions. First of all, the curves characterizing Erbin PDZ and PTP-BL PDZ2 are almost identical (both are approximately Gaussian functions with a mean of 0.628 nm and 0.64 nm, and standard deviation of 0.029 nm and 0.049 nm, respectively) indicating that the base part of the binding groove of these two PDZ domains behave in a very similar fashion. The distribution of the InaD PDZ1 domain, however, has larger spread (a standard deviation of 0.058 nm) , but the mean distance is about the same (0.635 nm) as in Erbin PDZ and PTP-BL PDZ2. Interestingly, the distance distribution of Dvl2 PDZ is a superposition of two Gaussian distributions (with a mean of 0.6 nm and a standard deviation of 0.058 nm). However, the location of one of the two superposed Gaussian curves agrees well with the distributions observed for Erbin PDZ and PTP-BL PDZ2.

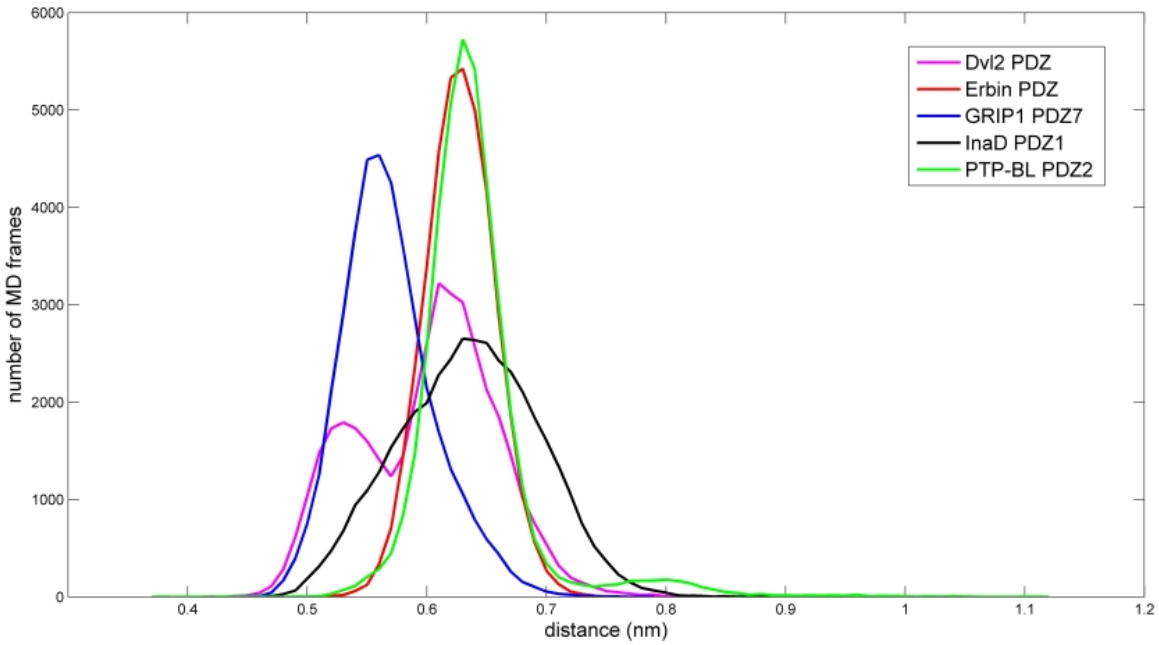

Figure 5.19: Distance distribution of the N-terminal residue of α2-helix and the C-terminal residue of β2-strand showing the extent the bottom of binding groove is open. The distributions is calculated for each PDZ domain based on 39801 simulation snapshots spanning 199 ns simulation time.

Most importantly, the distance distribution of GRIP1 PDZ7 (which can be approximated well as a single Gaussian distribution) is significantly shifted relative to the other four distributions. It has a mean of only 0.568 nm and a standard deviation of 0.039 nm. While the distance distributions of Erbin PDZ, PTP-BL PDZ2, Dvl2 PDZ and InaD PDZ1 domains have considerable overlap within the [0.6 nm; 0.7 nm] interval, the distribution of GRIP1 PDZ7 has a much smaller overlap with this region. In other words, the probability that the base part of the binding pocket is open with an extent larger than 0.6 nm is considerably low in the case of the GRIP1 PDZ7 domain but is high in the four other PDZ domains.

These results indicate that, as also shown by the experimental structures, the bottom of the binding groove of the GRIP1 PDZ7 domain is closed and it remains closed in the course of the 200 ns MD simulation unlike in other PDZ binding sites. This unique property of GRIP1 PDZ7 is probably the reason why this PDZ domain has been found to be unable to bind to carboxyl peptides. However, since the 200 ns MD simulation may still not pro-

vide sufficient conformational sampling of GRIP1 PDZ7, in order to determine whether its binding pocket really does not open spontaneously, one might need to use enhanced sampling MD simulation methods such as adaptive umbrella sampling or replica exchange MD (REMD) that would allow better exploration of the accessible conformational space of the apo domain.

5.5 Concluding discussions

The intrinsic dynamics of the binding sites of five PDZ domains have been compared in this chapter, based on 200 ns all-atom molecular dynamics simulations of the apo structures. The equivalent residues of the five binding pockets have been defined using a multiple sequence alignment of the PDZ domains. Despite the remarkable structural similarity of the five PDZ folds and binding sites, their fluctuation and flexibility properties have been found to be surprisingly different. Furthermore, the differences of their mobility correlate well with differences of their functional properties. Therefore, the intrinsic dynamics of the binding site seems to be a good predictor of functional characteristics.

The binding sites of InaD PDZ1 and Dvl2 PDZ are the most flexible of those of the five PDZ domains and this high degree of flexibility is likely to be necessary for them to be able to interact with multiple partners using significantly different binding modes, a property referred to "strong promiscuity" in this chapter. The Erbin PDZ domain, by contrast, has a rigid binding site and while it is also promiscuous, it interacts with very similar peptides using very similar binding modes. Besides the detailed characterisation of dynamics of PDZ domains, these results reveal a possibly generally important link between binding site flexibility and promiscuity also discussed in other studies.[173,350]

The MD simulation confirms that GRIP1 PDZ7 has a closed canonical binding site which is consequently unable to accommodate carboxyl peptides. The binding pocket does not appear to undergo a transition from its closed state to an open state in the course of the 200 ns trajectory. These results agree with the experimental observations that GRIP1 PDZ7 cannot interact with carboxyl ligands.

Currently there is no experimental structure available of the complex of InaD PDZ1 with the NinaC peptide. Based on the results presented in this chapter, it is predicted that InaD PDZ1 interacts with NinaC in a significantly different binding mode than it does with NorpA. This hypothesis could be tested experimentally and is a good example for that data of dynamics can be used to make predictions about the binding promiscuity of proteins.

Finally, the results about PTP-BL PDZ2 have revealed that the conformational space explored by the apo protein is the most different from the APC peptide-bound conformation compared to the other PDZ-peptide complexes. These results, in accordance with experimental data, suggests that the induced fit mechanism may be crucially involved in the binding of PTP-BL PDZ2 to the APC peptide. On the other hand, both the induced fit and the conformational selection mechanism seems to be important for PDZ domains to interact with various peptides.

While PDZ domains are structurally conserved, their sequences are highly diverged. As shown by Table 5.2, the average pairwise sequence identity of the five PDZ domains studied is 25.17 %. Consequently, even if the slight dissimilarities of their 3D structures cannot explain the large differences of their dynamics, the dissimilarities of their sequences could provide suitable explanation.

The conservation of their 3D structures and the divergence of their sequences indicate that the PDZ fold has high designability measured as the number of sequences that belong to the same fold. In other words, evolution had a large sequence space to explore for developing PDZ domains of different characteristics. The rigid binding sites of some PDZ domains might be optimized for interactions of reduced specificity but high affinity, while flexible binding sites could be optimal for promiscuous ligand binding. As discussed by Tokuriki et al.[173], a few mutations may change the conformational space visited by the protein, making it more rigid or more flexible altering its binding specificity properties (as for example observed in antibody maturation).

The results presented in this chapter also highlight how important it is for structure-based drug design studies to consider the flexibility of the target when searching for potent

inhibitors against PDZ domains. As the dynamics of the binding pocket seems to be a key factor determining the ability of a PDZ domain to interact with a small molecule, in cases of many PDZ domains, the flexibility of the binding site should also be taken into account in addition to the mobility of the ligand as it has already been done in some flexible docking studies.[248,249]

Related publication:

Münz, M., Hein, J. and Biggin, P.C. (2012). The role of flexibility and conformational selection in the binding promiscuity of PDZ domains. *PLoS Comp Biol*, 8(11):e1002749

Chapter 6

Network analysis of mouse PTP-BL PDZ2

6.1 Summary

The most important conclusion of Chapter 5 was that the binding specificity and promiscuity of PDZ domains are closely related to their intrinsic motions. However, recent studies have found that the peptide specificity of certain PDZ domains is not a constant property, but is also allosterically regulated. Some evidences support that dynamics could play a major role in mediating allosteric communication between distant sites of a single PDZ domain. The goal of the present study discussed in this chapter was to identify optimal intramolecular signalling pathways in the mouse PTP-BL PDZ2 domain that could be implicated in its allosteric regulation. A weighted network representation (proposed by Sethi et al.[309]) of the PDZ2 domain has been created based on a 200 ns molecular dynamics simulation of the apo structure. Using network analysis tools, three major optimal intramolecular pathways have been identified that overlap well with those pathways found in previous NMR and MD studies[222,223]. Two of the three pathways identified here appear to connect the peptide binding pocket to a distal surface which had been found to serve as a binding site for the PDZ1 domain[232]. Since the interaction between PDZ1 and PDZ2 domains had been shown to modulate the peptide binding specificity of PDZ2, the identified communication pathways are likely to be involved in allosteric signal transmission between the distal surface and the peptide binding site. The network analysis study presented in this chapter offers a complementary approach to previous research and was able to highlight multiple communication pathways predicted only by different studies.

6.2 Introduction

6.2.1 Dynamically driven allostery of proteins

Allostery, the process by which remote sites in a protein are energetically coupled, is an efficient and widely used mechanism to regulate protein activity.[409] Intramolecular signalling enables the communication of spatially distant sites via long-range propagation of information. In an allosteric protein, usually binding of a ligand to a regulatory or effector site results in a change at other functional sites, for example, altering the affinity for other binding partners.

Although evidences of allostery has been found for a large number of proteins, still little is know about the underlying processes that connect remote residues. According to the classical ('mechanical') view developed in the recent decades, allosteric regulation is mediated by a series of discrete conformational changes.[410–412] In this notion ligand binding actually alters protein structure and these structural changes at distant sites lead to the observed functional changes.

However, an early model and theoretical analysis presented by Cooper and Dryden [413] has shown that ligand-induced changes in dynamics can mediate allosteric communication between distinct binding sites, even in the absence of a macromolecular conformational change. Indeed, there is increasing evidence for the role of conformational dynamics in intramolecular communication.[222,414–419] The role of dynamics in allosteric regulation has been reviewed by Kern and Zuiderweg[420] and discussed in other papers[421–423]. Moreover, a few studies have found that allostery can be mediated exclusively by transmitted changes in protein dynamics.[60,222,409,424]

Some of the more important evidence of dynamics-mediated allostery has been presented by Popovych et al.[60] (also discussed in Section 1.2.3 and 1.4.4) who used NMR and isothermal titration calorimetry (ITC) to study the role of structure and dynamics in the negatively cooperative binding of cAMP to the dimeric catabolite activator protein (CAP). They have found that when the first cAMP binds to one subunit of a CAP dimer, it has no effect on the conformation of the other subunit. However, binding of the first cAMP

molecule partially enhances the dynamics of both the liganded and unliganded subunit, while subsequent binding of the second cAMP greatly suppresses protein motions. Therefore in this case allosteric communication is not accompanied with structural changes but is transmitted exclusively by changes in dynamics. These results have been confirmed by a further study.[409]

In an NMR study of the second PDZ domain of human protein tyrosine phosphatase PTP-BAS, Fuentes et al. have found that ligand binding results in large change of side chain motions[222]. Importantly, ligand-induced changes in dynamics was not limited to the binding site residues, but have also been observed at sites remote from the peptide binding pocket. However, at the locations where long-range changes in side-chain dynamics have been detected, no significant structural changes have been observed, indicating that the PTP-BL/BAS PDZ2 domain is also capable of intramolecular communication mediated exclusively by dynamics.

6.2.2 Allostery and signalling pathways in PTP-BL PDZ2

The possibility of dynamics-mediated intramolecular communication in the PTP-BL/BAS PDZ domain is of particular interest because the specificity of the domain has been shown to be allosterically regulated.[232] Van Den Berk et al. have used a random C-terminal peptide λ phage display library to study the binding preference of the five PDZ domains in the mouse PTP-BL protein. Testing several combinations of PDZ domains, they have found that while a separate PDZ2 domain is able to interact with class III peptides, a protein segment spanning the first two PDZ domains (PDZ1+2) failed to bind the same ligands.

Next, van Den Berk et al. tried to understand how the presence of PDZ1 domain regulates the binding specificity of PDZ2 domain. They have shown that the $\sim$ 200 amino acid long spacer region separating the two PDZ domains did not play a role in the allosteric communication, instead PDZ1 directly binds to PDZ2. The interaction between PDZ1 and PDZ2 domains has been characterized in an NMR experiment by titrating unlabeled PDZ1 into a sample containing ^{15}N-labeled PDZ2. Upon addition of PDZ1, major shifts have been observed for signals of PDZ2 residues in the α1-helix, β1-strand and the

C-terminal end of β6-strand that together form a surface opposite the peptide binding groove. In addition, residues at the top of the binding pocket important in peptide binding (e.g. 16Gly) have also found to be perturbed indicating the long-range allosteric effects of the PDZ1-PDZ2 interaction.

Although these results show that the binding of PDZ1 to the α1/β1/β6-interface on PDZ2 allosterically modulates the peptide specificity of PDZ2, the exact underlying mechanism is not known. As discussed in the previous subsection, the communication of the PDZ2 binding pocket with residues that are located remote from the binding site is mediated by the coupled dynamics of the system.

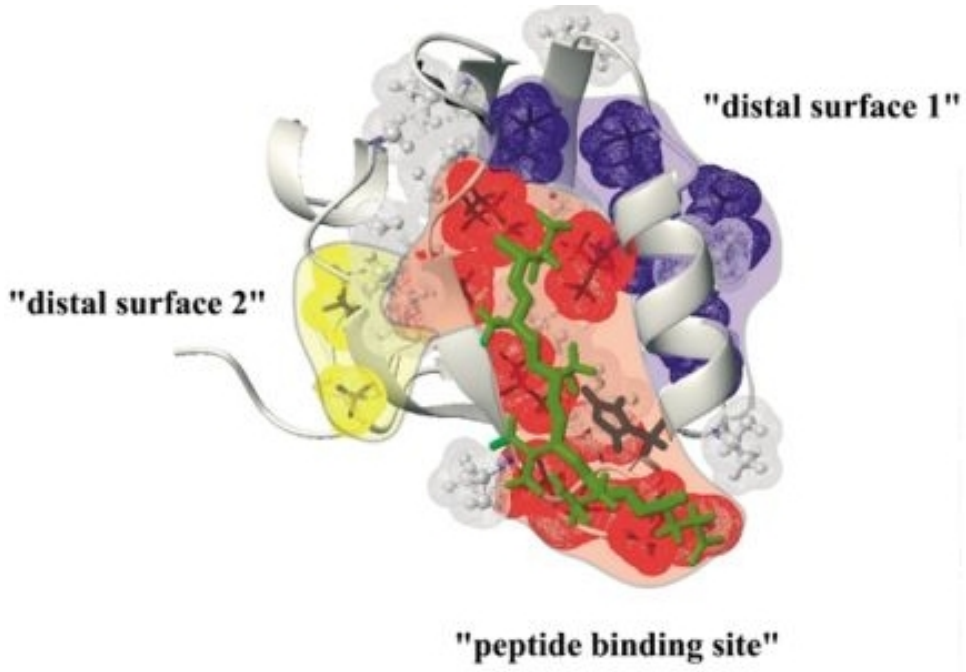

Figure 6.1: The peptide binding site and the two distal sites identified by Fuentes et al. are shown on a 3D structure of the human PDZ2 domain bound to a peptide. The binding site and distal surfaces 1 and 2 are highlighted in red, blue and yellow, respectively. The RA-GEF2 peptide ligand bound to the PDZ domain is shown in green. (Image Courtesy: Fuentes et al. 2004)

In particular, Fuentes et al. have identified two distal sites that are dynamically linked to the binding site in human PDZ2 (see Figure 6.1). The first region (referred to as "distal surface 1") contains residues in the N-terminal end of β6 and the anti-parallel β-sheet element formed by β4 and β5-strands. This region has been found to be linked to the

peptide-binding site through residue 75Leu located in the α2-helix. The second region (referred to as "distal surface 2") is located adjacent to the α1-helix and is linked to the binding pocket through 17Ile found at the N-terminal end of the β2-strand.

Kong and Karpus have performed MD simulations of both the apo and ligand-bound structures of the human PDZ2 domain. Using interaction correlation analysis[223], they have found two continuous interaction pathways connecting the ligand binding pocket with distant parts of the domain.

The first pathway they have identified (referred to as "Pathway I") starts from the binding site residue 17Ile (at the N-terminal end of the β2-strand) and extends along the axis of α1-helix. This pathway is similar to the one found by Fuentes et al. that connects the binding pocket with "distal surface 2". In addition, "Pathway I" agrees well with the pathway predicted by Lockless et al. in their statistical analysis of multiple sequence alignment (MSA) of PDZ domain sequences.[403] By contrast, the second pathway predicted by Kong and Karplus (referred to as "Pathway II") has not been identified in previous studies. It starts from the binding site residues 19Val and 20Thr (located on β2-strand) and runs perpendicularly across β2, β3, β4, β6 and β1-strands. Interestingly, some residues in "Pathway I" have been found to be coupled with residues in "Pathway II".

6.2.3 Network analysis of protein structures

Being complex systems, proteins can be effectively modelled and analysed using the concept of networks.[425] Network studies usually represent the 3D structures of proteins as unweighted or weighted undirected graphs (sometimes called 'protein structure networks') in which nodes correspond to residues and the links between them describe residue interactions (contacts).[425–427] Network analysis methods are now increasingly used to study different features of proteins such as flexibility and folding[428–430], structural similarity[122,431], recurring structural patterns[432,433], allosteric regulation[309,434] and stability[435,436]. Some studies have aimed to understand the general topological features of residue networks such as their degree distribution[437,438], small-worldness[437–440] and modularity[441,442]. Other studies have developed methods to identify key nodes or modules in the network that cor-

respond to functionally relevant sites or residues in the protein.[432,434,443]

However, the network representations used in most studies are constructed based on static protein structures, while proteins are essentially dynamic. If one is interested in properties of the protein that are closely related to its conformational dynamics (e.g. allosteric signalling), the fluctuation of the structure must also be taken into account. The key problem here is how to integrate the ensemble of networks that corresponds to the ensemble of conformations visited by the protein.

As a possible solution, Sethi et al.[309] have introduced a weighted network (referred to as the dynamical network) based on the contact graph of the protein. In this network each link was assigned a weight which was set to the absolute value of pairwise correlation between the two residues it connected (represented by their α-carbon atoms). In other words, the weight of a link described to what extent the dynamics of the two adjacent residues were coupled to each other. Since the pairwise residue correlation values were calculated based on an ensemble of conformations, this network representation incorporates information about conformational dynamics of the protein. (See the exact definition of dynamical networks in Section 6.3.2.)

Sethi et al. have successfully applied the concept of dynamical networks to study allosteric signalling in the bacterial glutamyl-tRNA synthetase and an archaeal leucyl-tRNA synthetase complexes. The networks have been created based on 20 ns molecular dynamics simulations of the tRNA:protein complexes. According to their interpretation, the link weights represent the probability of information transfer between two adjacent monomers. Using network analysis methods (i.e. shortest path, betweenness centrality, characteristic path length and community analysis) they have identified key residues and optimal and suboptimal communication pathways.

In this study, the concept of dynamical networks proposed by Sethi et al. has been applied to study allosteric signalling in the mouse PTP-BL PDZ2 domain (which has 95% sequence identity with the human PTP-BAS PDZ2 domain). The dynamical network representing the PDZ domain is created based on a 200 ns MD simulation of the apo structure. Network analysis methods are used to find optimal pathways in the weighted network

that could be involved in the allosteric communication of the domain.

6.3 Methods

6.3.1 Molecular Dynamics simulation

A 200 ns all-atom molecular dynamics trajectory of the apo mouse PTP-BL PDZ2 domain was analysed in this study. Note that the same simulation was used in the previous chapter in the comparative MD analysis. The relevant parameters of the MD simulation are described in details in Section 4.3.1. Snapshots were taken from the trajectory with a 50 ps time step; the ensemble used for calculating the residue contact and correlation matrices and constructing the dynamical network included 4000 simulation snapshots.

6.3.2 Construction of the residue network

The dynamical network representing the PTP-BL PDZ2 domain was created with the same method used by Sethi et al.[309], also discussed in Chapter 3. Each residue in the protein is represented by a node in the network. As in the original implementation of Sethi et al., two nodes were connected if the closest heavy atoms of the corresponding residues were within 4.5 Å of each other for at least 75% of the simulation snapshots (i.e. in at least 3000 frames). The contact matrix based on the large majority of trajectory frames is a more robust description of residue connectivities than a contact matrix calculated for a single simulation snapshot.

The only difference of the implementation used here was that network weights were derived from an NxN correlation matrix C which was calculated for the βC atoms (or αC atoms for glycine residues) as the goal was to study the coupled motions of side-chains instead of backbone atoms.

The C_{ij} pairwise correlation values represent the strength of dynamic coupling between neighbouring residues and were converted to the negative logarithm of their absolute value ($-log|C_{ij}|$) referred to as the length of the link between nodes i and j. While the C_{ij} correlation is interpreted as the probability of information transfer between the two

residues, the length of the link means how 'distant' two neighbouring nodes are on a communication pathway. Network analysis measures defined below were used to analyse the weighted graph to identify the most optimal communication pathways across the PDZ domain.

6.3.3 Network analysis measures

In graph theory, a path in a network is defined as a sequence of nodes such that from each of its nodes there is a link to the next node in the sequence. In a weighted network, the length of a path is the sum of the weights of the constituent links in the path. As a special case, in an unweighted graph the length of a path is simply the number of its links.

Shortest path distances

The goal of the shortest path problem is to find the path between given two nodes in the network that corresponds to the minimal path length between them. The shortest path length between two nodes is also referred to as their shortest path distance or geodesic distance.

For the dynamical network analysed in this study, the shortest path between two residues is interpreted as the most optimal communication pathway between them: i.e. the pathway through which signals are transmitted with maximal probability.

The Dijkstra's algorithm[444] was used in this study to derive the shortest paths for each pair of residues from the topology of the weighted network. The matrix storing the shortest path lengths of each pair of residues in the network will be referred to as D^0.

Betweenness centrality

The importance of a particular node in the communication of the network can be estimated by counting the number of shortest paths between each pair of nodes going across the node

of interest. The betweenness centrality[445] of a given node n is therefore defined as

$$C_B(n) := \frac{2}{(n-2)(n-1)} \sum_{s \neq n \neq t \in V} \frac{\sigma_{st}(n)}{\sigma_{st}} \tag{6.1}$$

where V is the total set of nodes in the network, σ_{st} is the number of shortest paths from node s to node t and $\sigma_{st}(n)$ is the number of shortest paths from s to t that pass through node n. $C_B(n)$ has a value between 0 and 1.

Residues of high betweenness centrality are in key position in the dynamical network and are probably crucial in the allosteric communication of the domain.

The betweenness centrality of a link **L** can be defined similarly:

$$C_B(\mathbf{L}) := \frac{2}{n(n-1)} \sum_{s \neq t \in V} \frac{\sigma_{st}(\mathbf{L})}{\sigma_{st}} \tag{6.2}$$

where $\sigma_{st}(\mathbf{L})$ is the number of shortest paths from s to t that pass through link **L**.

Characteristic Path Length analysis

Another way to study the importance of a residue is to test how the connectedness of other residues changes upon removing the given node from the graph. The Characteristic Path Length (CPL)[309] is a quantity characterizing the overall interconnectedness of the graph and is defined as the average shortest path length in the network:

$$CPL := \frac{1}{N_{pairs}} \sum_i \sum_j D^0_{ij} \tag{6.3}$$

Let $CPL^{(k)}$ denote the characteristic path length calculated in the network after removing node k with all its links. Furthermore, let $\Delta CPL^{(k)}$ be the increase of CPL upon the removal of node k:

$$\Delta CPL^{(k)} := CPL^{(k)} - CPL \tag{6.4}$$

Large $\Delta CPL^{(k)}$ value means that the node is in an important position in the graph as its removal has a major effect on the communication capabilities of the network.

As a normalization, the $\Delta CPL^{(k)}$ profile was converted to Z-scores using the following formula:

$$Z_k := \frac{\Delta CPL^{(k)} - < \Delta CPL^{(k)} >}{\sigma_{\Delta CPL^{(k)}}} \tag{6.5}$$

where $< \Delta CPL^{(k)} >$ is the mean and $\sigma_{\Delta CPL^{(k)}}$ is the standard deviation of $\Delta CPL^{(k)}$ values.

6.4 Results

6.4.1 Topology of the dynamical network

Figure 6.2 shows the dynamical network generated for the second PDZ domain of mouse PTP-BL based on a 200 ns MD simulation. The network was visualized using the Kamada-Kawai algorithm[311] which aims to find an optimal layout by assigning forces to the links in the graph and trying to find the positions of nodes that minimizes the total potential energy of the system.

The PTP-BL PDZ2 network is composed of 89 nodes and a total number of 331 links. The average node degree (i.e. average number of adjacent nodes) is 7.4 indicating the dynamics of a typical residue is coupled to the motion of several other neighbouring residues with which it is in direct contact.

The highest-degree nodes (hubs) of the network include residue 34Val which has 13 neighbours and 4Phe, 53Asp and 63Leu which have 12 neighbours. These nodes are indeed located at the core of the network layout. On the other hand, residues 1Gly, 25Thr, 26Ser and 78Thr have the smallest degree (≤ 3) and are therefore the most isolated. Residue 1Gly, for example, is located at the N-terminal of the domain, while residues 25Thr and 26Ser are found in the extended flexible $\beta 2/\beta 3$ loop at the base of the binding pocket (Walma et al. 2002). Note that this loop is also visualized by the Kamada-Kawai algorithm as a separated part of the network (Figure 6.2C).

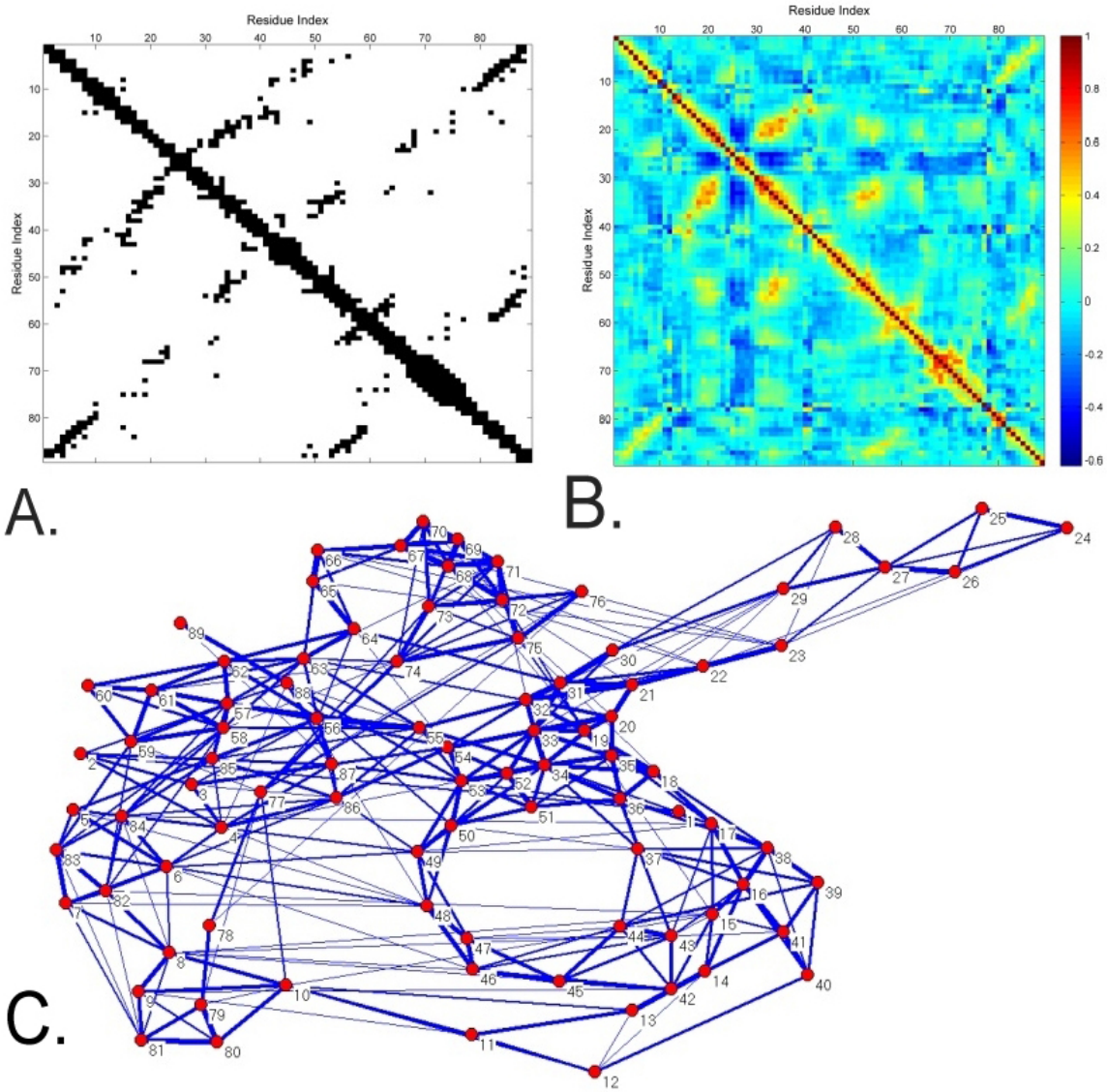

Figure 6.2: Steps of constructing the dynamical network of mouse PTP-BL PDZ2 domain: **A.** Contact matrix of the domain based on the 200 ns trajectory. **B.** Correlation matrix used as weight matrix of the network. **C.** The resulting weighted network visualized by the Kamada and Kawai force directed layout algorithm[311] using the network visualization program Pajek[310]. Edge widths are proportional to link weights.

The average link weight in the network is 0.32. Weights are visualized with different line widths in Figure 6.2. They vary considerably: the strongest link in the graph connecting 67Thr and 68His has a weight of 0.74, while the weakest link connecting 20Thr and 68His has a weight of 0.0034. This variation suggests that the shortest paths calcu-

lated for the weighted network are significantly different compared to those shortest paths calculated for the unweighted version of the network.

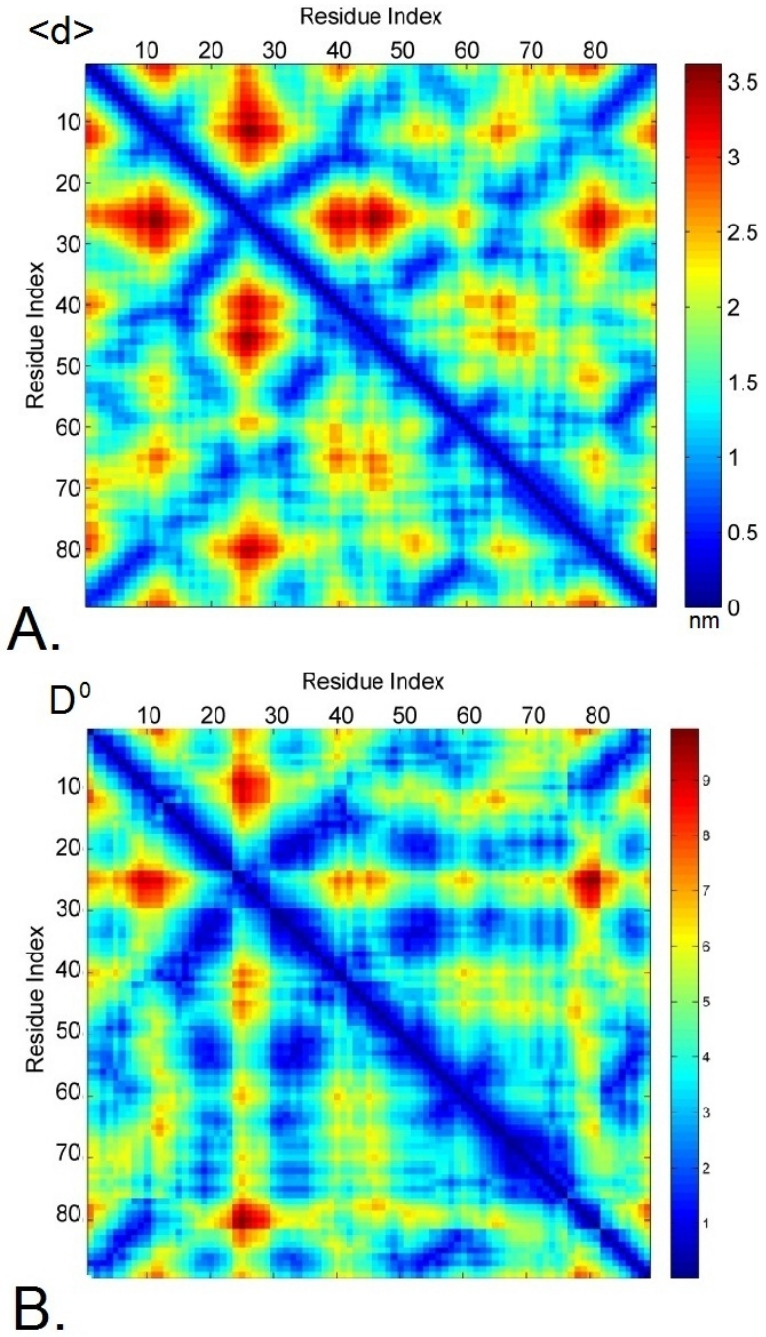

Figure 6.3: Color-coded mean distance matrix (**A.**) and shortest path length matrix (D^0) (**B.**). Blue regions correspond to residue pairs separated by short spatial distance and connected by short paths in the network. The characteristic path length of the network is 3.83.

6.4.2 Shortest path length matrix

Figure 6.3B shows the color-coded D^0 shortest path length matrix calculated by the Dijkstra's algorithm. The matrix contains the shortest path length for each pair of nodes in the network. The mean shortest path distance (characteristic path length) of the graph is 3.83. Although the D^0 matrix and the mean distance matrix of the MD trajectory (Figure 6.3A) look similar, detailed comparison of the two matrices has shown that certain residue pairs which have large spatial distance are connected by short paths in the dynamical network. For example, two remote residues, 19Val and 87Glu (that have a mean distance of 11.3 Å) are connected with shorter path ($D^0_{19,87} = 2.479$) than the pair of 40Lys and 44Glu ($D^0_{40,44} = 2.5$), two residues located relatively close in space (a mean distance of 5.4 Å).

Residue 19Val is located in the binding site (β2 strand) while 87Glu is found at the opposite side of the domain (C-terminal end of β6 strand). The considerably short path we see between these two distant residues may explain the coupling of their dynamics discovered by Kong and Karplus[446]. In that study 87Glu has been predicted to be located on an allosteric communication pathway connecting the binding pocket with a distal site of the domain.

6.4.3 Identifying central residues

To identify key residues in the protein that may be crucially implicated in long-range allosteric signal propagation, the betweenness centrality profile was calculated (see in Figure 6.4). The profile is shown for both the unweighted version of the network (when all link weights are set to 1) and the weighted dynamical network. As discussed in Methods, nodes of high betweenness centrality are likely to play important roles in intramolecular signalling pathways.

The betweenness centrality profiles calculated for the unweighted and the weighted networks are slightly different (i.e. a correlation of 0.66) which shows that assigning dynamics-based weights to links in the graph alters the optimal pathways in the network.

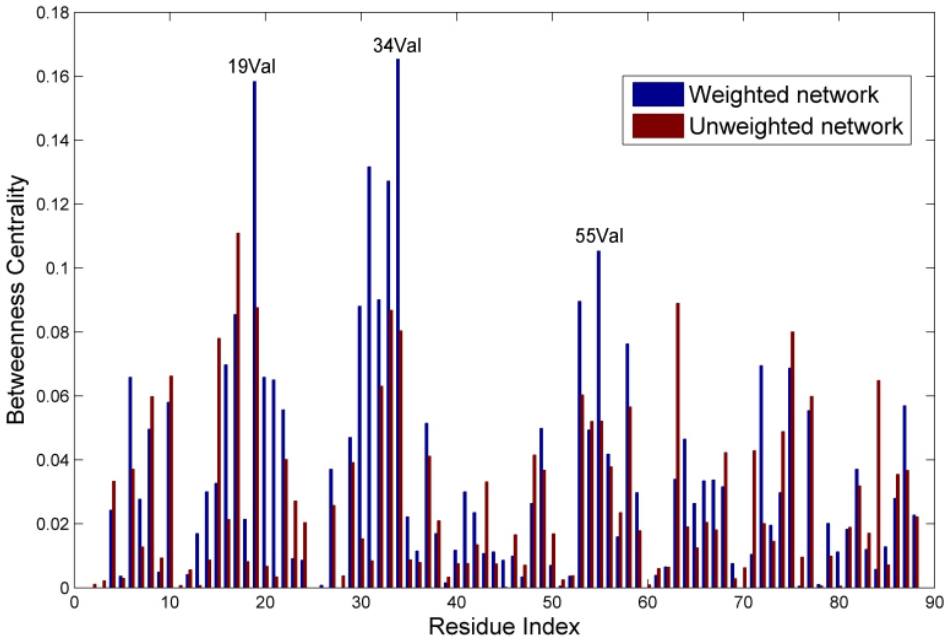

Figure 6.4: Node betweenness centrality profiles calculated for the unweighted and weighted network. Residues of high betweenness centrality in the weighted network are highlighted.

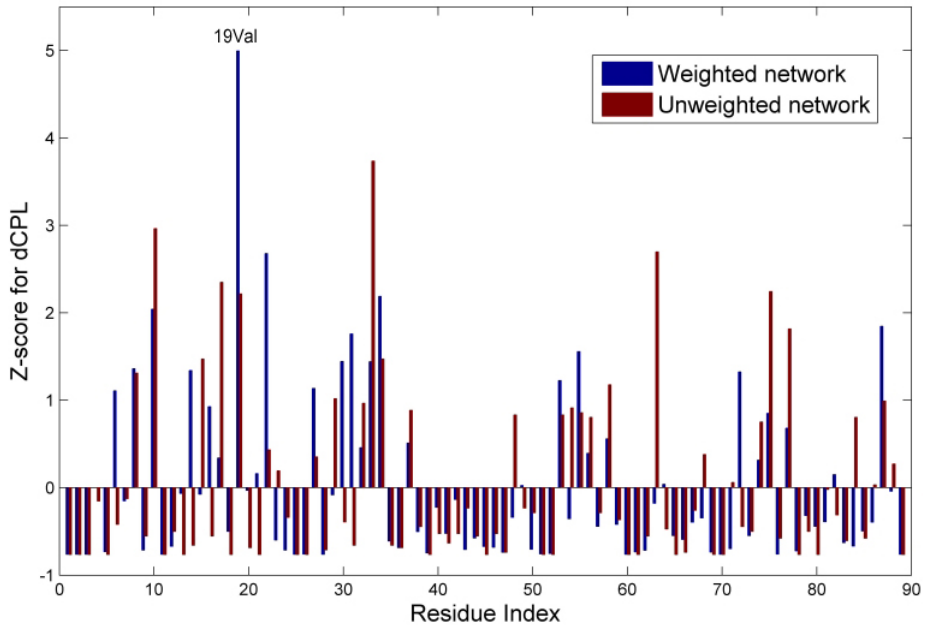

Figure 6.5: Z-score profile based on Characteristic Path Length (CPL) analysis of the unweighted and weighted network. The residue of highest Z-score in weighted network (19Val) is highlighted.

In the weighted network, the residues of highest betweenness centrality are 19Val, 31Gly, 33Tyr, 34Val and 55Val. Residue 19Val is located in the middle of β2 strand, 31Gly, 33Tyr and 34Val are found on the β3 strand while 55Val is in the middle of the β4 strand (see the positions of 19Val, 34Val and 55Val on the 3D-structure in Figure 6.6).

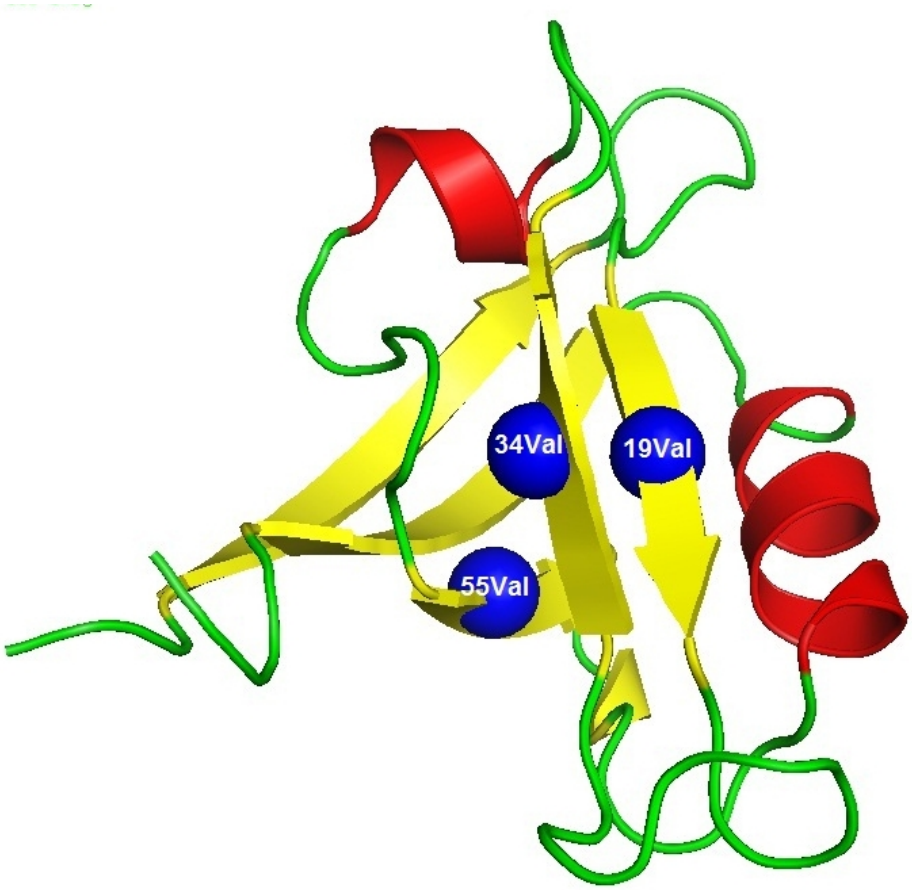

Figure 6.6: Solution NMR structure of PTP-BL PDZ2 (PDB: 1gm1) and the positions of the three Valine residues found to be highly central in the dynamical network of the PDZ domain.

Furthermore, the Z-score profile based on characteristic path length (CPL) analysis was also calculated and is presented in Figure 6.5. Again, the profiles of the weighted and unweighted networks are similar, but there are considerable differences (i.e. a correlation of 0.58). In the weighted network, 19Val has far the largest Z-score suggesting that this residue is crucial in connecting different parts of the domain. Other residues that have high Z-score include 10Lys, 22Gly, 34Val and 87Glu.

6.4.4 Identifying key links and communication pathways

In order to identify crucial links that may be important in allosteric communication, the betweenness centrality of each link in the network was calculated and are summarized in the matrix shown in Figure 6.7.

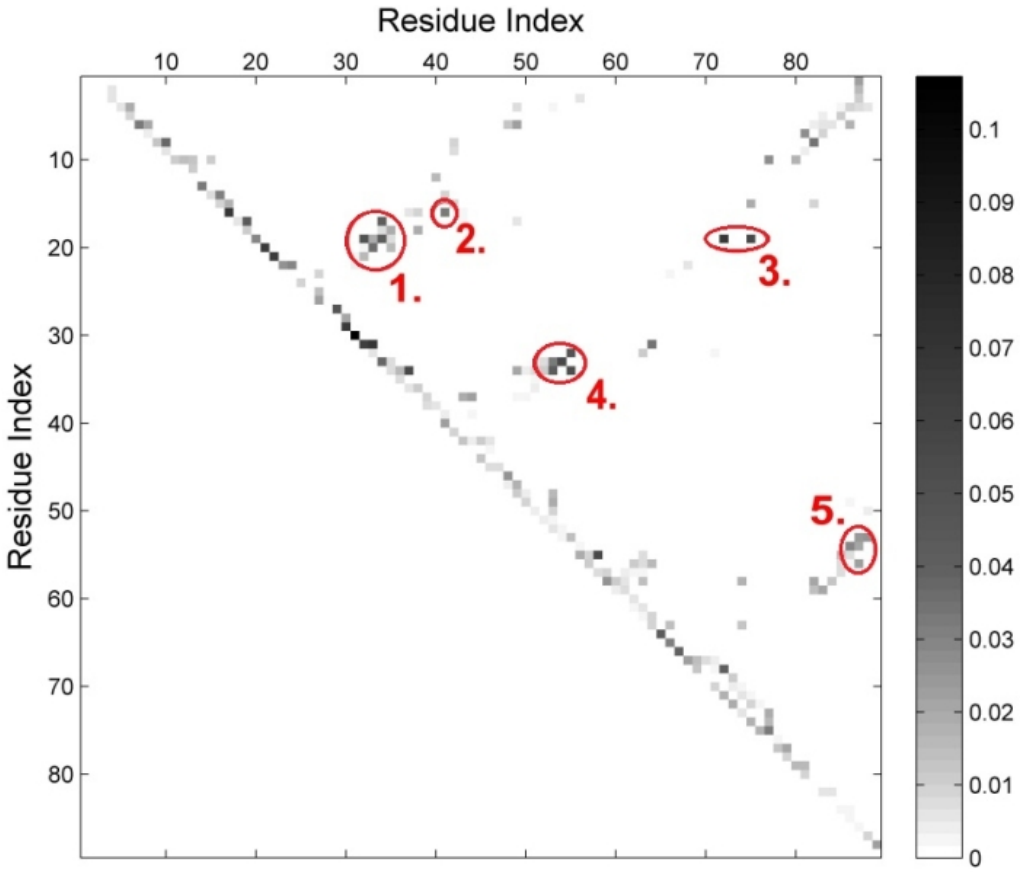

Figure 6.7: Matrix containing the betweenness centrality of each link in the network. Matrix entry i,j is the $C_B(L)$ betweenness centrality of the link between node i and j (if any). Five regions of high betweenness centrality are highlighted: 1. links between β2-strand and β3-strand; 2. link between 16Gly and 41Gly; 3. links between β2-strand and α2-helix; 4. links between β3-strand and β4-strand; 5. links between β4-strand and the C-terminal end of β6-strand.

The link betweenness centrality matrix reveals several links that appear to be essential for the interconnectedness of the dynamical network. These crucial connections include links between the middle of β2-strand (19Val, 20Thr) and the middle of β3-strand (32Ile, 33Tyr, 34Val). We can see two important links between the β2-strand (19Val) and α2-helix (72Val, 75Leu). Links between the the middle of β3-strand (32Ile, 33Tyr, 34Val) and the

N-terminal half of β4-strand (53Asp, 54Arg, 55Val) are important to highlight. A link between 16Gly and 41Gly also seems to be critical. Finally, important links connecting the N-terminal end of β4-strand (53Asp, 54Arg) and the C-terminal end of β6-strand (86Leu, 87Glu, 88Lys) are revealed. These examples of high link betweenness centrality are also highlighted in Figure 6.7.

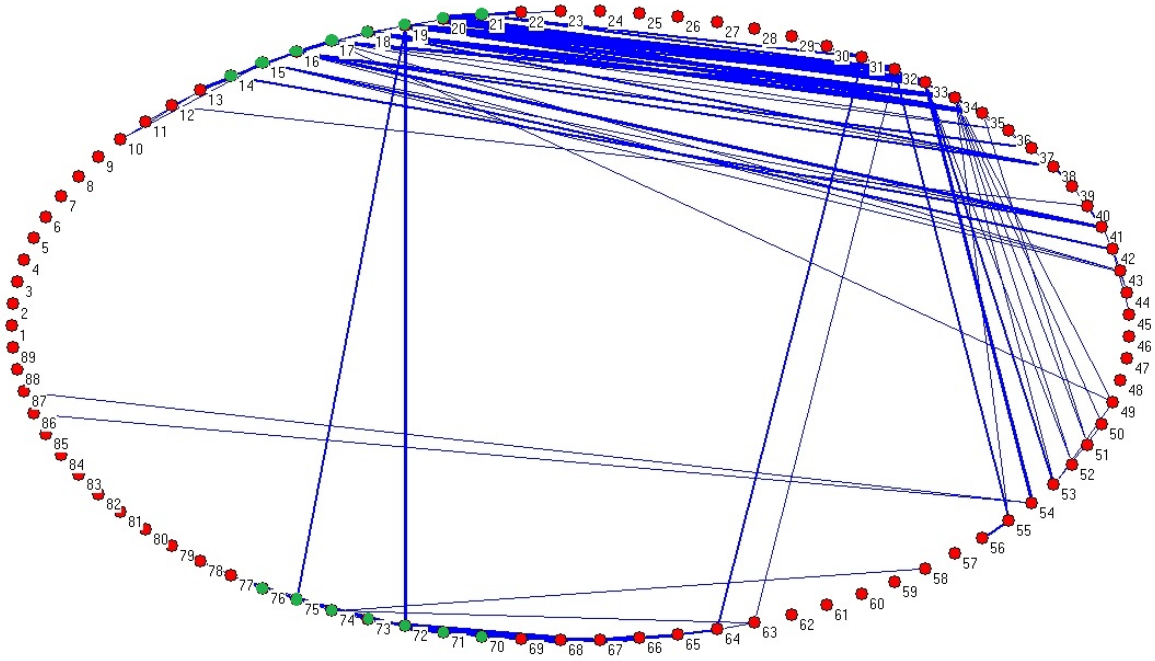

Figure 6.8: Shortest paths between the binding pocket and the rest of the domain are highlighted in a circular network layout. Nodes of green color represent binding site residues. Short paths (with a length less than 2.5) connecting binding site residues to non-binding site residues are shown. Edge widths are proportional to link betweenness centralities.

These results together with the node betweenness centrality and characteristic path length analysis profiles suggest that there is a key pathway passing through the β2, β3, β4 and β6 strands, connecting the binding site with the C-terminal end of β6-strand. This pathway is the reason of the low shortest path distance of 19Val and 87Glu discussed above. Residues predicted to be central (19Val, 34Val and 55Val) are located on the identified pathway. On the other hand, 19Val is connected to residues of the α2-helix as well via links of high betweenness centrality. This suggests that the communication between the β2-strand and the α2-helix is also mediated by the residue 19Val. Finally, the highly cen-

tral link between 16Gly and 41Gly indicates that there may be an other important pathway connecting the binding pocket and the α1-helix.

To further investigate how the binding pocket is linked to other parts of the domain, the key shortest paths connecting the binding site residues (i.e. 14Ser-21Gly and 70Gln-76Arg) with non-binding site residues are visualized on a circular layout of the network (Figure 6.8). For convenience, only those shortest paths that have a length less than 2.5 are highlighted. Using this diagram one can identify possible pathways by systematically mapping the optimal paths between each pair of end points which are connected by short distances in the graph (a cut-off of 2.5 was used here) and separated by a spatial distance in the 3D-structure larger than a given threshold. This analysis provides a set of pathways connecting the binding site with distal sites that have large spatial distance but low shortest path distance in the dynamical network. The identified pathways found between different pairs of residues may overlap and can therefore be clustered into a smaller number of important pathways. Applying this analysis, we can see an important pathway starting from residues of the β2-strand (19Val, 20Thr), passing through the β3 strand (33Tyr) and β4 strand (54Arg) and linking into the C-terminal end of β6-strand (86Leu, 87Glu). Another key pathway seems to connect residues of the β1/β2 loop (14Ser-16Gly) and the α1-helix (41Gly-45Ser). One can also observe the connection between β2-strand (19Val) and α2-helix (72Val, 75Leu) discussed above. In addition, an optimal pathway is found to connect the α2-helix (70Gln-75Leu) with the β5/α2-loop (63Leu-69Lys).

To summarize, betweenness centrality analysis has been used to identify links in the network that are likely to be crucial for dynamics-mediated allosteric communication of the PDZ domain. These links appear to be involved in a few major pathways connecting the binding site with other parts of the domain. In particular, three pathways are highlighted here that could play important roles in long-range communication. The pathway between β2-strand and the C-terminal end of β6-strand (which will be referred to as "Pathway A"). The pathway connecting the β1/β2 loop and α1-helix (which will be referred to as "Pathway B"). Finally, the pathway which connects α2-helix with the β5/α2-loop (which will be referred to as "Pathway C"). These three major pathways are mapped on the 3-

dimensional structure of the PTP-BL PDZ2 domain (see Figure 6.9).

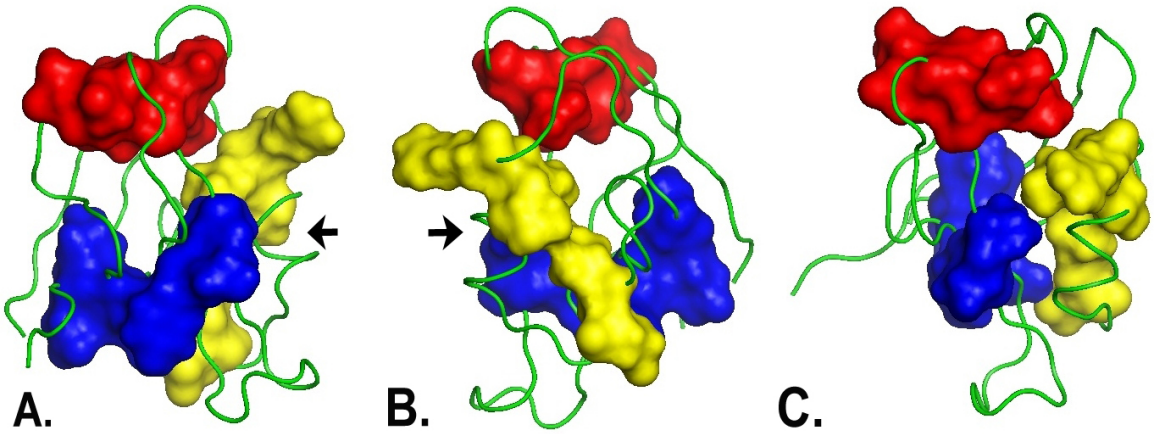

Figure 6.9: The three major optimal communication pathways identified are mapped on the crystal structure of the PTP-BL PDZ2 domain (PDB: 1gm1). "Pathway A", "Pathway B" and "Pathway C" are highlighted in blue, red and yellow, respectively. The same structure is shown in three different orientations: the location of the peptide binding site is marked in **A.** and **B.**, while **C.** shows the front view of the peptide binding pocket.

These results agree strikingly well with the conclusions of other studies that aimed to find allosteric communication pathways in the PTP-BL/BAS PDZ2 domain (summarized in Introduction). "Pathway A" and "Pathway B" identified here are remarkably similar to the two intramolecular pathways referred to as "Pathway II" and "Pathway I" found by Kong and Karplus in the human PDZ2[446]. In addition, the region connected to the binding pocket by "Pathway C" overlaps with "distal surface 1" identified by Fuentes et al.[222]. On the other hand, the region connected to the binding pocket by "Pathway B" is close to "distal surface 2" found in the NMR study. Hence, the three unique pathways found in previous studies have all been predicted by this network analysis method.

6.4.5 How weights affect shortest path distances

To provide an additional example to show that dynamics-based weighting alters the result of analysis, Figure 6.10 shows a comparison of the weighted and the unweighted shortest path lengths between the binding site residue 18Ser and the rest of the network.

For a set of residues of equal unweighted shortest path distance from 18Ser, we can

observe large variation in their weighted shortest path distances. For example, while both 8Leu and 87Glu are 3 link away from 18Ser, their weighted shortest path lengths are considerably different: 5.217 for 8Leu and 2.578 for 87Glu. Therefore the probability of information transfer is significantly larger between 18Ser and 87Glu than between 18Ser and 8Leu. As also discussed earlier, using the dynamically weighted network instead of the unweighted network (based on the contact matrix) gives considerably different results about the optimal pathways in the graph. In other words, the MD simulation data used for estimating the dynamic couplings between residue pairs were essential for inferring the optimal communication pathways that are in good overlap with those found in other studies.

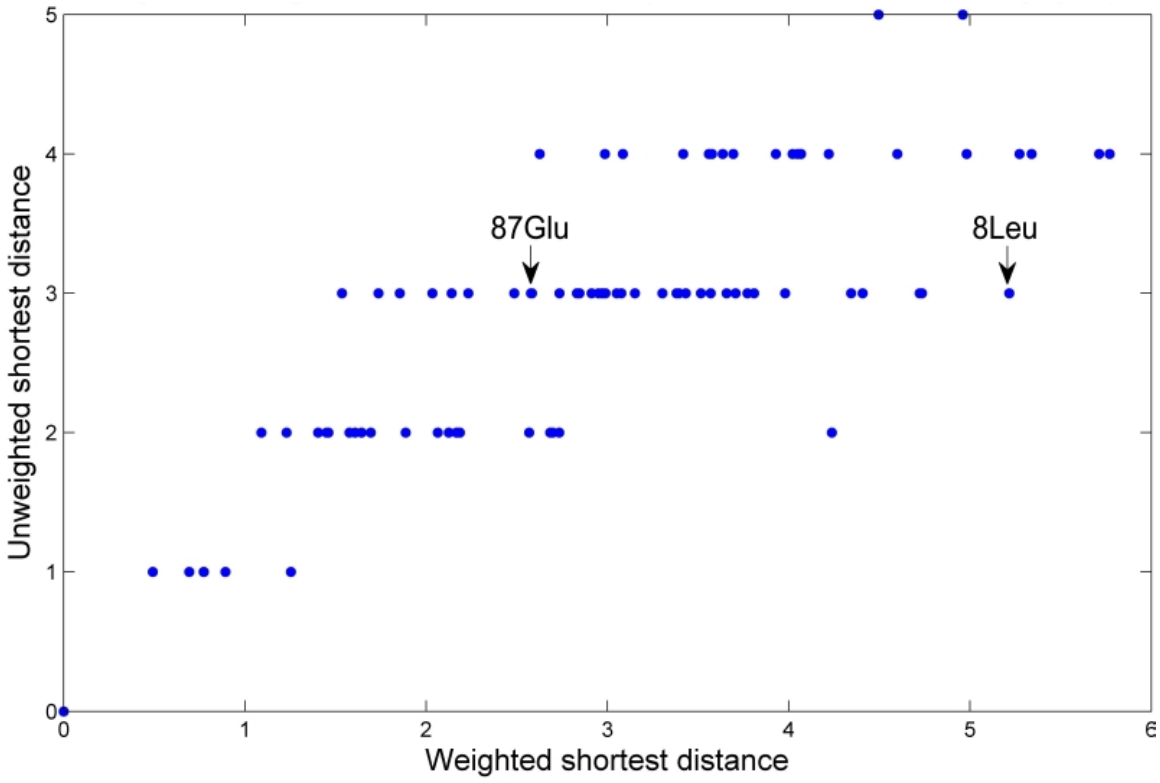

Figure 6.10: Comparison of the weighted and unweighted shortest path lengths from 18Ser to all other residues in the network. The shortest path distances to two residues (8Leu and 87Glu) are highlighted. While 18Ser has an unweighted shortest path distance of 3 to both residues, the weighted shortest path lengths are significantly different (5.217 to 8Leu and 2.578 to 87Glu).

6.5 Concluding discussion

In this study, standard network analysis methods have been used to identify optimal intramolecular signalling pathways in the mouse PTP-BL PDZ2 domain, a protein that is known to be capable of long-range allosteric communication. A network representing the PDZ domain has been created based on both structural and simulation data. Firstly, the trajectory contact matrix describing robust inter-residue contacts has been used for defining the topology of the network. Secondly, the link weights in the network have been calculated based on a 200 ns MD simulation of the protein with the method proposed by Sethi et al.[309] The dynamical network provides a simple representation to summarize residue contacts and dynamic couplings observed in a conformational ensemble. These results show that, as expected, the optimal communication pathways are different in the dynamically weighted network than in the unweighted (contact-based) network.

Using betweenness centrality and characteristic path length analysis, a number of central residues have been identified in the dynamical network including 19Val, 34Val and 55Val. (34Val also serves as a hub in the network.) These three residues found on different secondary structural elements (β2-strand, β3-strand and β4-strand, respectively) are predicted to play key roles in allosteric communication. Analysis of the link betweenness centrality matrix revealed that these residues are located on an optimal pathway that originates from the binding pocket and ends at the C-terminal end of β6-strand ("Pathway A").

Focusing the study on the optimal pathways in the network that connect the binding site with other parts of the domain showed that besides "Pathway A" there are two other important paths: a pathway connecting the β1/β2 loop and α1-helix ("Pathway B") and the pathway between the α2-helix and the β5/α2-loop ("Pathway C"). In addition, a link that seems to be crucial for communication between the β2-strand and the α2-helix has been found.

The optimal signalling pathways described here agree remarkably well with those found in previous studies. The NMR study by Fuentes et al.[222] and the molecular dy-

namics study by Kong and Karplus[446] have together identified 3 unique intramolecular pathways in the human PDZ2 domain. As detailed above, these pathways highly overlap with those detected by the dynamical network analysis method discussed in this chapter. In addition, "Pathway B" agrees well with the results of the statistical analysis based on a multiple sequence alignment of PDZ domains by Lockless et al.[403]

It is important to note that the two distal regions found to be linked to the binding site through "Pathway A" and "Pathway B" (i.e. α1-helix and the C-terminal end of β6-strand) are located on the surface to which PDZ1 domain has been found to bind in the NMR study by van Den Berk et al.[232] Since the binding of PDZ1 to PDZ2 causes long-range allosteric effects altering the binding specificity of the PDZ2 domain, it is tempting to speculate that "Pathway A" and "Pathway B" play a role in relaying the allosteric signals between the PDZ1/PDZ2 interaction site and the peptide binding site.

The present study has demonstrated the usefulness of the concept of dynamical networks in studying intramolecular signalling mechanisms. The strength of the dynamical network method is that it greatly simplifies the data produced by the MD simulation but still captures key information about the adjacency and cooperativity of residues. Network analysis has emerged as a popular and powerful tool in systems biology. Networks serve as highly simplified representations of the biological systems that incorporate relevant information about the interactions of components. Once a biological system is "translated" to the language of network science, a wide range of measures and algorithms developed in the field of graph theory can be applied to analyse its key properties. Therefore, it is proposed here that structural bioinformatics could also benefit more from the network analysis approach, provided that transforming structural data to network representations is relatively straightforward.

Chapter 7

Concluding remarks and future directions

Proteins are complex dynamical systems that exist in an ensemble of conformations determined by their free-energy landscapes (see Section 1.2.1). Although the conformational dynamics of many proteins are intimately linked to their biochemical functions (e.g. flexibility was found to be key to ligand binding, catalysis and allosteric signalling), little is known about how their intrinsic motions are encoded in their primary sequences and tertiary structures (see Section 1.4). We are just beginning to understand the relationship between protein sequence/structure and dynamics space and to reveal how exactly functional properties depend on conformational flexibility. Studying these questions requires comparative analysis of protein dynamics in a systematic manner.

We see today an explosion in the amount of data of protein motions generated by experimental approaches (e.g. NMR spectroscopy, fluorescence and single-molecule FRET) and computational methods (e.g. atomistic/coarse-grained MD simulations and Elastic Network Models).[16]. The massive increase of data allows comprehensive, large-scale investigation of protein dynamics and comparative approaches are to play important role in this field. However, while many successful algorithms and similarity measures are available for comparing protein sequences and structures (reviewed by Mount 2004[447] and Koehl 2001[116]), the optimal way of quantifying the similarity of protein dynamics is still an open question.

Most of the work covered in this thesis represents an attempt to create a systematic framework for the comparative analysis of protein dynamics. The approach introduced here has been tested on several members of the PDZ domain family. Besides shedding light on the strength and weaknesses of the methodology, this work has lead to several inter-

esting conclusions about the relationship between conformational flexibility and binding specificity of PDZ domains.

The main conclusions of this research are discussed in this last chapter which is divided into three sections: conclusions of methodological results, conclusions about PDZ domains and general findings about the relationship between sequence, structure and dynamics. Finally, possible directions of future research are discussed in the last section.

7.1 Conclusions of methodological results

7.1.1 The five challenges of comparative analysis

Although several different strategies have been proposed so far for the comparative analysis of protein motions, every approach had to address the following five challenges.

Challenge 1: Representation of protein dynamics

The input of comparative analysis usually comprises two or more conformational ensembles (e.g. provided by MD simulations or NMR experiments) that represent the conformational subspaces visited by the proteins. In particular, continuous MD simulations provide high-dimensional trajectories in the conformational space. The direct comparison of these 3N-dimensional trajectories (where N is the number of atoms of the protein) is a problem difficult to address. One therefore needs a simplified (lower-dimensional) representation of the MD data.

Dimensionality reduction (Section 2.2) is a widely used approach in the field of data mining often applied when comparing high-dimensional datasets. As discussed in Section 4.2, many comparative MD studies were based on classical dimensionality reduction methods such as Principal Component Analysis (PCA) or Multidimensional Scaling Analysis (MDS) to create lower-dimensional representations of the high-dimensional data. For example, a selected subset of principal components corresponding to the largest eigenvalues provide a reduced input dataset for comparative analysis while still capturing essential information about functionally relevant modes of motions of the proteins. Residue

fluctuations are often characterized using RMSF profiles which also serve as simplified description of complex motions (Section 2.1.2). Whatever method is used for this purpose, the objective is to extract characteristic features from the large datasets describing protein dynamics at atomic resolution. Similarly, the dynamics of Elastic Network Models (ENMs) (which are *per se* simplified representations of protein structures) are usually characterized using Normal Mode Analysis (NMA) (Section 2.3,2) to extract a set of normal mode vectors capturing the essential modes of motions.

Challenge 2: Residue matching problem

In order to compare the dynamics of different but homologous protein structures, every method must have its own solution for the problem of finding the best mapping between the two non-identical sets of residues, referred to as the 'residue matching problem' (see Section 2.5.4). If the proteins to be compared have significantly similar structures, using their prior sequence or structural alignment can be a straightforward solution. However, this may not be the case; we should be able to compare the dynamics of structurally dissimilar proteins as well. In addition, even when comparing the dynamics of structurally similar proteins, one may not want to constrain the search for dynamically similar regions using a prior structural alignment.

Instead of relying on a prior alignment, an alternative option is creating an alignment that is optimized to match the dynamically most similar regions of the proteins. In other words, in this case the alignment is not the input but the output of the algorithm. As a result, the dynamics-based alignment score provides a measure of dynamic similarity reflecting how well the motions of the proteins agree, much like the DALI alignment score serves as a measure of structural similarity.

Challenge 3: Method of comparison

Once the complex, high-dimensional data of protein dynamics is transformed into a simplified, low-dimensional representation, the extracted features are to be compared across the different proteins (or different simulations of the same protein). Various algorithms

and similarity measures have been proposed for comparative analysis, depending on how the dynamics of proteins are represented. For example, when MD or ENM data are described with a set of principal components or normal modes, the subspace overlap measure is often used to estimate the similarities of conformational subspaces spanned by these principal component or normal mode vectors. Similarly, the covariance matrix overlap is commonly used for comparing two covariance matrices. Another example is the ensemble averaged RMSD (eRMSD) measure calculated between two conformational ensembles[448]. (An overview of different comparative methods introduced in the literature are given in Section 2.2 and Section 4.2) The objective of these methods is to provide a global (protein-level) similarity score and/or to perform local (residue-level) comparison.

Challenge 4: Significance analysis

Once the extracted features of dynamics are compared between different proteins or simulations, the statistical significance of similarity must also be estimated. This is especially important, because one can expect a certain extent of dynamic similarity between random pairs of proteins. The statistical significance (e.g. p-value) of dynamic similarity therefore needs to be calculated relative to the background similarity score distribution of random proteins. While this analysis is necessary to make reliable conclusions from the similarity scores, several previous studies did not use a rigorous significance analysis framework.

In addition, one should also take into account the sampling problems of MD simulations which result in detectable difference between the dynamics of the very same protein calculated for different simulations. For example, while overlaying RMSF profiles is a widely used approach for comparing dynamics, the significance of RMSF-difference between different proteins as compared to the expected RMSF-difference between different MD simulations of the same protein has hardly been discussed.

Challenge 5: Posterior analysis of comparative results

Finally, the results of comparative analysis may form the basis of further investigations aiming to study how the functional similarity of proteins relates to their dynamic simi-

larity, to explain the differences of their dynamics based on their primary sequences and tertiary structures and to understand the underlying mechanisms of protein evolution. Addressing these questions would require the integration of the results of comparative protein dynamics study with comparative sequence and structure analysis, in addition to functional, mutational and molecular phylogenetic data.

7.1.2 The methodology introduced in the thesis

The approach used in this work for analysing and comparing protein dynamics was mostly inspired by the structural alignment algorithm DALI. There are several analogies between the method presented in this thesis and the DALI algorithm. First of all, as discussed in Section 2.5.2, DALI represents protein structures with their distance matrices summarizing inter-atomic distances. As mentioned in 2.5.2, it can be shown that the distance matrix contains enough information for reconstructing the original three-dimensional structure. DALI addresses the pairwise structural alignment problem by aiming to solve the matrix alignment of the two input distance matrices. Similarly, the basic idea of the work summarized here was to introduce applicable matrix representations of the conformational dynamics of proteins. Since the structural flexibility of a protein is best captured by its conformational ensemble, instead of static inter-atomic distances used by DALI, inter-atomic distance distributions observed in the whole ensemble are to be characterized. In order to construct a simple matrix representation of conformational dynamics, several summary statistics of the inter-atomic distance distributions can be used including the mean, the standard deviation and variance, the range and interquartile range etc. The result is an NxN matrix (where N is the number of atoms included in the analysis) that can be compared between different proteins, as distance matrix is compared in DALI. Three characteristic matrices, the DFM, F and X matrix have been defined and used in Chapters 4 and 5 based on the standard deviation, variance and range summary statistics, respectively. These matrices serve as simplified representations of the high-dimensional MD trajectories, thereby providing solution for *Challenge 1.* discussed above.

In order to compare conformational ensembles of the *same* protein, one only has to

overlay the above-defined matrices calculated for the different ensembles. Since the proteins described by these ensembles are identical, there is no need for an alignment. This is the case, for example, when apo and holo simulations of a protein are compared.

One the other hand, when comparing *different* proteins, one has to address to 'residue matching problem' (discussed above and in Section 2.5.4). Both strategies proposed to solve this question (*Challenge 2*) have been explored in this thesis. Prior pairwise alignments define LxL submatrices of the dynamics-based matrices of both proteins (where L is the number of aligned residues). The two LxL submatrices capturing the relative mobility of aligned residues can be directly compared. Alternatively, the matrix alignment algorithm introduced can be used to identify similar pairs of submatrices in the two input matrices (*Challenge 3*).

To address the NP-hard problem of pairwise matrix alignment, a heuristic strategy has been developed based on the MCMC (Markov chain Monte Carlo) approach discussed in details in Section 2.4. The method uses the Simulated Annealing (SA) metaheuristic: i.e. a Markov chain of decreasing "temperature" parameter is generated that is designed to converge to an optimal matrix alignment ideally corresponding to the global maximum of the submatrix similarity score. Since the algorithm performs stochastic optimization, each run may give different outputs and therefore a number of random restarts were created and the best alignment (i.e. with the largest similarity score) found in the different Markov chains were selected. Note that in this strategy, the submatrix similarity score is used as the objective function to be optimized and the maximal score encountered is given as the final alignment score of the two proteins. The matrix alignment algorithm introduced here was applied to the comparison of DFM matrices of PDZ domains (Chapter 4). To assess the statistical significance of similarity (*Challenge 4*), alignment scores were compared to a random background score distribution calculated for a set of evolutionarily and functionally unrelated proteins.

An apparent weak point of the matrix alignment algorithm is that it does not guarantee finding the best possible alignment between the two input matrices as the Markov chain may be trapped by local minima in the search space. This common problem of global

optimization methods is often called the quasi-ergodicity of the search space referring to the fact that although the Markov chain can escape from the local minimum with non-zero probability, this is such a small probability that it actually never happens during the simulation. Unlike in the case of pairwise sequence alignments where the dynamic programming approach (Needleman-Wunsch algorithm) guarantees to give the exact solution (i.e. the global maximum of alignment score), in case of matrix alignment this approach is not feasible due to the NP-hardness of the problem. However, performing random restarts greatly increases our confidence about the results given by the MCMC algorithm.

Another issue to consider is the convergence of DFM patterns during the MD simulations. In particular, when comparing the dynamic fingerprint matrices of different simulations, one should be aware of the uncertainty of individual DFMs due to the incomplete conformational sampling. Therefore the convergence of DFM patterns was tested by comparing five independent 20 ns simulations of the same protein, PSD-95 PDZ3 domain. Importantly, the matrix alignment method has shown that the five DFMs had highly significant similar scores and were much more similar than those of different PDZ domains. These results suggested that the DFMs calculated based on 20 ns MD simulations were converged enough to be used for comparative analysis. However, DFM convergence time may vary across different proteins as it probably depends on the topography of the energy landscape. Therefore more work is necessary to clarify the uncertainty and convergence of dynamic fingerprint matrices in general. Note that since the DFM-based "average fluctuation profile" highly correlates with the standard RMSF plot (Section 4.4.1), the assessment of DFM convergence is intimately linked to the question of convergence of RMSF profiles.

The prior alignment-based comparative strategy was tested in Chapter 4 and 5 using Needleman-Wunsch pairwise sequence alignments to guide the comparison of dynamics. In case the DFM, F or X matrices are compared based on prior sequence or structural alignments, the overall submatrix similarity score measures how much the motions of the two aligned residue sets agree in the two proteins. Furthermore, in addition to the overall similarity score that quantifies the global dynamic similarity of proteins, overlaying the aligned submatrices enable us to study which particular regions have most similar and

most different dynamics in the two structures. For example, analogously to the *difference distance matrix* commonly used for visualizing the location and extent of structural differences between two conformations (Section 5.3.1), the *difference fluctuation matrix* can help us to detect local dynamic differences between two structural ensembles (Section 3.4.1).

7.2 Conclusions about PDZ domains

PDZ domains have diverse peptide binding specificity characteristics which can be explained to a large extent by differences of their binding site residues that form specific interactions with the peptide ligands (see Section 1.5.4). However, as suggested by earlier studies and confirmed by the results presented in this thesis, the conformational dynamics of PDZ domains also appear to play important role in their binding specificity. As PDZ domains have highly conserved global folds and binding sites, the preliminary expectations was that they would also show very similar inherent dynamics. Surprisingly, the opposite was found: the comparative approach introduced here has revealed important differences between the global and local (i.e. binding site) fluctuations of the studied PDZ domains.

First of all, while the global motions of some PDZ domains were identified to be significantly similar, other pairs of PDZ domains were found to be dynamically different (discussed in Chapter 4). In particular, the global fluctuations of metazoan PDZ domains were shown to be dynamically more conserved than those of non-metazoan PDZ domains. The slight structural differences between the metazoan and circularly permuted non-metazoan PDZ fold seems to be reflected in the dissimilarity of global dynamics. A more focused analysis of binding pocket residues (discussed in Chapter 5) has also revealed distinguishing dynamic properties of five PDZ binding sites. In addition, the identified differences of binding pocket flexibility in these PDZ domains were found to correlate with functional properties (i.e. peptide binding promiscuity). In addition, optimal (dynamics-mediated) intramolecular pathways were highlighted in the mouse PTP-BL PDZ2 domain (Chapter 6). As shown by other studies, these signalling pathways are crucial for the allosteric communication between the peptide binding site and distal sites of the domain.

Thus the DFM methodology was proven to be efficient in highlighting the detailed differences in dynamics. However, several important questions arise based on these results. First of all, what sequence or structural features account for the observed dynamic differences of PDZ domains? How do single point mutations alter their conformational dynamics? Is it possible to identify paths of consecutive mutations or individual transition points in sequence space (discussed in Section 1.4.3 and 1.4.4) that connect these distinct states of dynamics? Can one describe mutations that are able to rigidify the binding pocket or make it more flexible thereby altering peptide binding specificity?

The detailed analysis of the flexibility of five PDZ binding sites has helped to explain some of their important functional characteristics. For example, the prominent flexibility of the binding sites of Dvl2 PDZ and InaD PDZ1 domains is likely to be linked to their ability to bind multiple ligands in significantly different binding modes. By contrast, the relative rigidity of the Erbin PDZ binding pocket is suggested to be related to the reduced specificity range of the domain. On the other hand, the peptide binding site of the GRIP1 PDZ7 domain was found to remain closed in the course of the 200 ns MD simulation, explaining the observation that this binding pocket is unable to interact with peptides. While both the conformational selection and induced fit mechanisms seem to be involved in the ligand recognition process of PDZ domains, the results indicate that out of the five PDZ domains studied, it is the PTP-BL PDZ2 domain in which induced fit was found to have the largest effect.

Taking these results together, it seems likely that the dynamics of the $\alpha 2/\beta 2$ binding pocket is under evolutionary selection pressure. If so, binding site residues are not only optimized for determining binding specificity via specific interactions with peptide ligands, but also for contributing to optimal binding pocket flexibility necessary for interacting with a range of partners. Taking into account the flexibility of the PDZ binding sites is particularly important for structure-based drug design as was recognized in flexible docking studies of PDZ domains.[248,249] Designing small molecule compounds that bind into the $\alpha 2/\beta 2$ binding groove with high affinity and selectivity could be guided by the knowledge of the characteristic motions of the PDZ binding sites of interest.

To summarize, despite having significantly similar three-dimensional structures, the peptide binding sites of PDZ domains were found to be remarkably versatile and capable of performing very different conformational dynamics. Such versatility might be crucial for PDZ domains to adopt diverse binding specificity properties. Although the fold is highly conserved, the striking sequence divergence of the PDZ domain family is likely to explain the observed divergence in dynamics space.

7.3 General conclusions about sequence, structure and dynamics

In addition to the specific results about PDZ domains, this work has lead to some interesting general conclusions about the relationship between protein sequence, structure and dynamics. As discussed in details in Section 1.4, the mapping between protein sequence space, structure space and function space is unexpectedly complex which makes sequence and structure-based function prediction a rather difficult problem. However, as proposed in Section 1.4.4, the dynamic similarity of proteins could also be used as a measure to predict protein functions since it may correlate with functional similarity even when sequence and structure are not reliable predictors. The results summarized in this thesis suggest that characterizing the distances of proteins in the dynamics space could provide the bridge between sequence/structure and biochemical function.

As illustrated by the results of Chapter 4, there is detectable correlation between the global structural similarity and the global dynamic similarity of proteins, as measured by DALI Z-scores and dynamics-based alignment scores, respectively. However, conservation of the tertiary structure does not necessarily indicate that the dynamics is also conserved as exemplified by the results about PDZ domains in Chapter 4 and 5. These results suggest that, such like the amino acid sequence, conformational dynamics can diverge faster than structure. Furthermore, this conclusion is supported by other findings of a number of studies discussed in Section 1.4.4. The possible higher rate of divergence in protein dynamics space than in structure space could explain fast evolutionary transitions between distinct biochemical functions of highly conserved protein structures.

What is suggested here is not that protein dynamics generally diverges faster than protein structure: this is obviously not the case as a series of studies have demonstrated that many proteins of conserved structures also have conserved motions. However, discontinuities (transition points) are likely to exist in the mapping from sequence/structure to dynamics: i.e. proteins for which small changes of sequence or structure results in large differences in dynamics (see Section 1.4.4). These transition points at which protein dynamics diverges faster than sequence or structure are suggested to facilitate the emergence of new functional properties in evolution. Importantly, the existence of transition points in the sequence→dynamics interface imposes limitations on coarse-grained ENM models which represent proteins structures without taking into account their sequences (Section 2.3).

7.4 Future directions of research

It is the essential nature of scientific research that with every answered question several new questions arise. Indeed, a few ideas and hypotheses that were not explored thoroughly in this thesis are proposed here as sensible starting points for further research work. Some of the possible directions of future investigations are discussed in this last section.

In the first place, several improvements could be made to the dynamics-based alignment method to increase the speed and accuracy of the algorithm. For example, instead of searching the alignment space by adding or removing single amino acid pairs at every step (see Section 4.3.5), the alignment could be assembled from small blocks of sub-alignments as for example performed by the DALI algorithm (see Section 2.5.2). Another possible improvement of the method would be to enable the construction of alignments in which the two proteins are not necessarily in the same sequence order, thus allowing the correct match of inverted or circularly permuted sequence regions. In its current form, the algorithm only allows the identification of protein alignments that keep the original sequence order of the input proteins.

Furthermore, the statistical framework of the algorithm could also be improved by

performing a more comprehensive assessment of the significance of dynamic similarity scores. To this end, the random background score distribution might be derived based on an extended collection of evolutionarily and functionally unrelated proteins. In addition, the critical simulation length necessary to obtain converged DFM patterns may also be studied in more details by comparing a larger set of repeated simulations of different lengths and this convergence analysis could be carried out for multiple different proteins.

Another idea proposed here is to develop a more refined algorithm that would aim to create *local* dynamics-based alignments of proteins. Although the alignments generated by the method introduced in this thesis are "local" in the sense that the alignment score does not depend on unaligned regions, they are also "global" in that the evaluation of statistical significance of the similarity score depends on the total lengths of sequences. Therefore, highly scoring local alignments are not detected due to their low significance. However, one may be interested in finding local sequence or structural regions that have similar dynamics in two proteins despite their dissimilar global motions. Serving as a complementary approach to local sequence and structural alignment tools (discussed in Section 2.5.3), a local dynamics-based alignment method could help to explain and predict conservation of biochemical functions based on local similarity of proteins dynamics.

Although the comparative approach described in the thesis have been used to study PDZ domains as a test case, the method is proposed to be generally applicable for studying the conservation of protein motions. Therefore one straightforward extension of the research presented here is to use the introduced methodology for comparing the dynamics of a more diverse set of proteins. As a number of studies have used the DALI similarity score to map the protein structure space (Section 1.4.1), the dynamic similarity score could be used to map the protein dynamics space. However, creating a large-scale map of the dynamics space seems to be difficult challenge given the large computational cost of MD simulations. Other methods that use ENM representations of proteins as the input of comparative analysis are naturally much more scalable, but less accurate.

One interesting application of the dynamics-based alignment method would be identifying pairs of proteins that have significantly different sequences and three-dimensional

structures, yet share similar conformational dynamics (as shown by the schematic illustration of Figure 1.16F). However, it is not clear if the concept of dynamic similarity can be defined in a sensible way in the absence of structural similarity. Nevertheless, comparing protein pairs (if any) that have low structural similarity score but high dynamic similarity score could be an interesting question.

Another possible application of the methodology described in the thesis is the comparison of the fluctuations of proteins in their apo and holo states. Various studies have found that upon ligand binding proteins can undergo functionally important changes regarding their dynamics (e.g. rigidification of the binding pocket or dynamics-mediated allosteric communication between the binding site and distal sites). In order to characterize these dynamic differences between the apo and holo states, one can directly compare the overlaid DFM patterns.

Clearly, the comparison of two-dimensional DFMs provides more detailed information about the difference of mobility of residues than simply comparing the one-dimensional RMSF profiles. However, as shown in Section 4.4.1, the dynamic fingerprint matrix can also be converted into a one-dimensional profile characterizing residue fluctuations that is referred to as the average fluctuation profile (AFP). The average fluctuation value of a residue is defined as the mean of each row of the DFM. As shown in Section 4.4.1, in case of a 20 ns simulation of PSD-95 PDZ3, the AFP was almost perfectly correlated with the RMSF profile. The minor difference between the two profiles is probably due to the fact that RMSF depends on arbitrary parameters of the superposition process while AFP is independent of superposition. Therefore, it is hypothesized here that AFP may serve as a more reliable measure of residue fluctuations than RMSF and the correlation between the two profiles decreases with the overall flexibility of the protein. Since PDZ domains have relatively rigid structures, the correlation between AFP and RMSF is over 90%. However, in case of highly flexible proteins that undergo considerable large-scale motions (and consequently, rigid-body superposition fails to fit their dissimilar conformations), the difference between AFP and RMSF is likely to be more prominent. This hypothesis could be tested easily, studying the possible benefits of AFP analysis. On the other hand, these

results also suggest that the standard superposition-based analysis of protein dynamics might be replaced with fully distance-based analysis that is not biased by superposition (see Section 4.5). To this end, the first step could be to investigate the possibility of introducing a superposition-independent form of the coordinate-covariance matrix.

Finally, several interesting questions arise based on the results about PDZ domains presented in the thesis. First of all, does the observed correlation between binding site flexibility and binding promiscuity generally holds in the family of PDZ domains? It would be straightforward to extend the comparative MD study discussed in Chapter 5, by comparing a larger set of PDZ domains based on longer (or repeated) MD simulations. Such an extended study would aim to collect more data about the relationship of their flexibility and peptide binding specificity to gain a more comprehensive picture of the mapping of PDZ domains from dynamics space to function space. On the other hand, an even more important question is to understand how the dynamics of PDZ domains are encoded in their sequences and structures. The integration of comparative dynamics analysis with the results of comparative sequence and structural studies and mutational data could give valuable insight into the mapping between the sequence, structure and dynamics space of PDZ domains. Some of the most important questions are the followings: How can a series of mutations change the flexibility of the binding pocket or the whole domain? Is it possible to identify transition points at which mutations can switch between distinct dynamic characteristics? What structural features (e.g. hydrophobic interactions, hydrogen bonding, salt bridges or loop rigidity) account for the diverse dynamics of different PDZ domains? Is it possible to infer the evolutionary relationships between PDZ domains from the comparison of their dynamics only? (*Challenge 5.*)

In addition, a specific hypothesis is formulated in Chapter 5, regarding the interaction of InaD PDZ1 with the NinaC peptide. As discussed in Section 5.4.3, it is expected that the binding mode of InaD PDZ1 with NinaC is significantly different from that of the structurally characterized interaction with the NorpA peptide. Currently no structural data are available to confirm this idea, and the hypothesis could be tested experimentally by solving the structure of InaD PDZ1 domain in complex with the NinaC peptide.

Bibliography

1. Ares, M. (2004). Interdisciplinary research and the undergraduate biology student. *Nature Structural and Molecular Biology*, **11**(12):1170–2.

2. Csete, M. E. and Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science*, **295**(5560):1664–9.

3. Wooley, J. C. and Lin, H. S., eds. (2005) *Catalyzing Inquiry at the Interface of Computing and Biology* (National Academies Press (US)).

4. Newman, M. E. J. (2011). Complex Systems: A Survey. *American Journal of Physics*, **79**(8):800.

5. Kitano, H. (2001) Systems Biology: Toward System-level Understanding of Biological Systems, in *Foundations of Systems Biology*.

6. Kitano, H. (2002). Systems biology: a brief overview. *Science*, **295**(5560):1662–4.

7. Kohl, P., Crampin, E. J., Quinn, T. A., and Noble, D. (2010). Systems biology: an approach. *Clinical Pharmacology and Therapeutics*, **88**(1):25–33.

8. Cuccato, G., Della Gatta, G., and di Bernardo, D. (2009). Systems and Synthetic biology: tackling genetic networks and complex diseases. *Heredity*, **102**(6):527–32.

9. Reichhardt, T. (1999). It's sink or swim as a tidal wave of data approaches. *Nature*, **399**(6736):517–20.

10. Butcher, E. C., Berg, E. L., and Kunkel, E. J. (2004). Systems biology in drug discovery. *Nature Biotechnology*, **22**(10):1253–9.

11. Kettman, J. R., Coleclough, C., Frey, J. R., and Lefkovits, I. (2002). Clonal proteomics: one gene - family of proteins. *Proteomics*, **2**(6):624–31.

12. Kurakin, A. (2007). Self-organization versus Watchmaker: ambiguity of molecular recognition and design charts of cellular circuitry. *Journal of Molecular Recognition*, **20**(4):205–14.

13. Mechelke, M. and Habeck, M. (2010). Robust probabilistic superposition and comparison of protein structures. *BMC bioinformatics*, **11**(1):363.

14. Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quarterly Reviews of Biophysics*, **36**(3):307–40.

15. Teilum, K., Olsen, J. G., and Kragelund, B. B. (2009). Functional aspects of protein flexibility. *Cellular and Molecular Life Sciences*, **66**(14):2231–47.

16. Henzler-Wildman, K. and Kern, D. (2007). Dynamic personalities of proteins. *Nature*, **450**(7172):964–72.

17. Zhang, J., Sapienza, P. J., Ke, H., Chang, A., Hengel, S. R., Wang, H., Phillips, G. N., and Lee, A. L. (2010). Crystallographic and nuclear magnetic resonance evaluation of the impact of peptide binding to the second PDZ domain of protein tyrosine phosphatase 1E. *Biochemistry*, **49**(43):9280–91.

18. Garnier, S., Gautrais, J., and Theraulaz, G. (2007). The biological principles of swarm intelligence. *Swarm Intelligence*, **1**(1):3–31.

19. Bajec, I. L. and Heppner, F. H. (2009). Organized flight in birds. *Animal Behaviour*, **78**(4):777–789.

20. Deisboeck, T. S. and Couzin, I. D. (2009). Collective behavior in cancer cell populations. *BioEssays*, **31**(2):190–7.

21. Sannita, W. G. (2008). Neuronal functional diversity and collective behaviors. *Journal of Biological Physics*, **34**(3-4):267–78.

22. Bianca, C. and Pennisi, M. (2012). The triplex vaccine effects in mammary carcinoma: A nonlinear model in tune with SimTriplex. *Nonlinear Analysis: Real World Applications*, **13**(4):1913–1940.

23. Carnevale, V., Raugei, S., Micheletti, C., and Carloni, P. (2006). Convergent dynamics in the protease enzymatic superfamily. *Journal of the American Chemical Society*, **128**(30):9766–72.

24. Grant, B. J., Gorfe, A. A., and McCammon, J. A. (2010). Large conformational changes in proteins: signaling and other functions. *Current Opinion in Structural Biology*, **20**(2):142–7.

25. Janovjak, H., Sapra, K. T., Kedrov, A., and Müller, D. J. (2008). From valleys to ridges: exploring the dynamic energy landscape of single membrane proteins. *Chemphyschem: a European journal of chemical physics and physical chemistry*, **9**(7):954–66.

26. Leeson, D. T. and Wiersma, D. A. (1995). Looking into the energy landscape of myoglobin. *Nature Structural Biology*, **2**(10):848–51.

27. Frauenfelder, H., Fenimore, P. W., and Young, R. D. (2007). Protein dynamics and function: insights from the energy landscape and solvent slaving. *IUBMB Life*, **59**(8-9):506–12.

28. Kitao, A. and Go, N. (1999). Investigating protein dynamics in collective coordinate space. *Current Opinion in Structural Biology*, **9**(2):164–9.

29. Gerstein, M. (1998). A database of macromolecular motions. *Nucleic Acids Research*, **26**(18):4280–4290.

30. Zhao, Y., Stoffler, D., and Sanner, M. (2006). Hierarchical and multi-resolution representation of protein flexibility. *Bioinformatics*, **22**(22):2768–74.

31. Kuznetsov, I. B. (2009). Simplified computational methods for the analysis of protein flexibility. *Current Protein and Peptide Science*, **10**(6):607–13.

32. Vaccaro, L., Koronakis, V., and Sansom, M. S. P. (2006). Flexibility in a drug transport accessory protein: molecular dynamics simulations of MexA. *Biophysical Journal*, **91**(2):558–64.

33. Barrett, C. P., Hall, B. A., and Noble, M. E. M. (2004). Dynamite: a simple way to gain insight into protein motions. *Acta Crystallographica. Section D, Biological Crystallography*, **60**(Pt 12 Pt 1):2280–7.

34. Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**(4):433–459.

35. Cui, Q. and Bahar, I. (2006) *Normal Mode Analysis: Theory And Applications to Biological And Chemical Systems* (CRC Press).

36. Amadei, A., Linssen, A. B., and Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins*, **17**(4):412–25.

37. Amadei, A., Linssen, A. B., de Groot, B. L., van Aalten, D. M., and Berendsen, H. J. (1996). An efficient method for sampling the essential subspace of proteins. *Journal of Biomolecular Structure and Dynamics*, **13**(4):615–25.

38. Teodoro, M. L., Phillips, G. N., and Kavraki, L. E. (2002) A dimensionality reduction approach to modeling protein flexibility, in *Proceedings of the sixth annual international conference on Computational biology - RECOMB '02*, pages 299–308 (ACM Press, New York, New York, USA).

39. Teodoro, M. L., Phillips, G. N., and Kavraki, L. E. (2003). Understanding protein flexibility through dimensionality reduction. *Journal of Computational Biology*, **10**(3-4):617–34.

40. Berendsen, H. J. and Hayward, S. (2000). Collective protein dynamics in relation to function. *Current Opinion in Structural Biology*, **10**(2):165–9.

41. Miller, D. W. and Agard, D. A. (1999). Enzyme specificity under dynamic control: a normal mode analysis of alpha-lytic protease. *Journal of Molecular Biology*, **286**(1):267–78.

42. Chau, P. L., van Aalten, D. M., Bywater, R. P., and Findlay, J. B. (1999). Functional concerted motions in the bovine serum retinol-binding protein. *Journal of Computer-aided Molecular Design*, **13**(1):11–20.

43. Peters, G. H. and Bywater, R. P. (2002). Essential motions in a fungal lipase with bound substrate, covalently attached inhibitor and product. *Journal of Molecular Recognition*, **15**(6):393–404.

44. Chilliemi, G., Fiorani, P., Benedetti, P., and Desideri, A. (2003). Protein concerted motions in the DNA-human topoisomerase I complex. *Nucleic Acids Research*, **31**(5):1525–1535.

45. Hammes, G. G., Benkovic, S. J., and Hammes-Schiffer, S. (2011). Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. *Biochemistry*, **50**(48):10422–30.

46. Derreumaux, P. and Schlick, T. (1998). The loop opening/closing motion of the enzyme triosephosphate isomerase. *Biophysical Journal*, **74**(1):72–81.

47. Williams, J. C. and McDermott, A. E. (1995). Dynamics of the flexible loop of triosephosphate isomerase: the loop motion is not ligand gated. *Biochemistry*, **34**(26):8309–19.

48. Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavlou, G., Eisenmesser, E. Z., and Kern, D. (2004). Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nature Structural and Molecular Biology*, **11**(10):945–9.

49. Ostermann, A., Waschipky, R., Parak, F. G., and Nienhaus, G. U. (2000). Ligand binding and conformational motions in myoglobin. *Nature*, **404**(6774):205–8.

50. Lange, O. F., Lakomek, N.-A., Farès, C., Schröder, G. F., Walter, K. F. A., Becker, S., Meiler, J., Grubmüller, H., Griesinger, C., and de Groot, B. L. (2008). Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**(5882):1471–5.

51. James, L. C., Roversi, P., and Tawfik, D. S. (2003). Antibody multispecificity mediated by conformational diversity. *Science*, **299**(5611):1362–7.

52. Baker, K. A., Tzitzilonis, C., Kwiatkowski, W., Choe, S., and Riek, R. (2007). Conformational dynamics of the KcsA potassium channel governs gating properties. *Nature structural & molecular biology*, **14**(11):1089–95.

53. Haliloglu, T. and Ben-Tal, N. (2008). Cooperative transition between open and closed conformations in potassium channels. *PLoS Computational Biology*, **4**(8):e1000164.

54. Yang, H., Yu, Y., Li, W.-G., Yu, F., Cao, H., Xu, T.-L., and Jiang, H. (2009). Inherent dynamics of the acid-sensing ion channel 1 correlates with the gating mechanism. *PLoS Biology*, **7**(7):e1000151.

55. Liu, J., Zhang, J., Yang, Y., Huang, H., Shen, W., Hu, Q., Wang, X., Wu, J., and Shi, Y. (2008). Conformational change upon ligand binding and dynamics of the PDZ domain from leukemia-associated Rho guanine nucleotide exchange factor. *Protein Science*, **17**(6):1003–14.

56. Biehl, R., Hoffmann, B., Monkenbusch, M., Falus, P., Préost, S., Merkel, R., and Richter, D. (2008). Direct observation of correlated interdomain motion in alcohol dehydrogenase. *Physical Review Letters*, **101**(13):138102.

57. Bu, Z., Biehl, R., Monkenbusch, M., Richter, D., and Callaway, D. J. E. (2005). Coupled protein domain motion in Taq polymerase revealed by neutron spin-echo spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(49):17646–51.

58. Smock, R. G. and Gierasch, L. M. (2009). Sending signals dynamically. *Science*, **324**(5924):198–203.

59. Jarymowycz, V. A. and Stone, M. J. (2008). Remote changes in the dynamics of the phosphotyrosine-binding domain of insulin receptor substrate-1 induced by phosphopeptide binding. *Biochemistry*, **47**(50):13371–82.

60. Popovych, N., Sun, S., Ebright, R. H., and Kalodimos, C. G. (2006). Dynamically driven protein allostery. *Nature Structural and Molecular Biology*, **13**(9):831–8.

61. Dodson, G. G., Lane, D. P., and Verma, C. S. (2008). Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO reports*, **9**(2):144–50.

62. Debye, P. (1913). Interferenz von Röntgenstrahlen und Wärmebewegung. *Annalen der Physik*, **348**(1):49–92.

63. Waller, I. (1923). Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen. *Zeitschrift für Physik A: Hadrons and Nuclei*, **17**(1):398–408.

64. Dobrianov, I., Caylor, C., Lemay, S., Finkelstein, K., and Thorne, R. (1999). X-ray diffraction studies of protein crystal disorder. *Journal of Crystal Growth*, **196**(2-4):511–523.

65. Fisette, O., Lagüe, P., Gagné, S., and Morin, S. (2012). Synergistic applications of MD and NMR for the study of biological systems. *Journal of Biomedicine and Biotechnology*, **2012**:254208.

66. Teng, Q. (2005) *Structural Biology: Practical NMR Applications* (Birkhäuser).

67. Tzakos, A. G., Grace, C. R. R., Lukavsky, P. J., and Riek, R. (2006). NMR techniques for very large proteins and rnas in solution. *Annual Review of Biophysics and Biomolecular Structure*, **35**:319–42.

68. Zhao, X. (2012). Protein structure determination by solid-state NMR. *Topics in Current Chemistry*, **326**:187–213.

69. Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C., and Ha, T. (2008). Advances in single-molecule fluorescence methods for molecular biology. *Annual Review of Biochemistry*, **77**:51–76.

70. Roy, R., Hohng, S., and Ha, T. (2008). A practical guide to single-molecule FRET. *Nature Methods*, **5**(6):507–16.

71. Diez, M., Zimmermann, B., Börsch, M., König, M., Schweinberger, E., Steigmiller, S., Reuter, R., Felekyan, S., Kudryavtsev, V., Seidel, C. A. M., and Gräber, P. (2004). Proton-powered subunit rotation in single membrane-bound F0F1-ATP synthase. *Nature Structural and Molecular Biology*, **11**(2):135–41.

72. Pirchi, M., Ziv, G., Riven, I., Cohen, S. S., Zohar, N., Barak, Y., and Haran, G. (2011). Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nature Communications*, **2**:493.

73. Karplus, M. and Kuriyan, J. (2005). Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(19):6679–85.

74. Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science*, **330**(6002):341–6.

75. Adcock, S. A. and McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical Reviews*, **106**(5):1589–615.

76. Bruccoleri, R. E. and Karplus, M. (1990). Conformational sampling using high-temperature molecular dynamics. *Biopolymers*, **29**(14):1847–62.

77. Sugita, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, **314**(1-2):141–151.

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

79. Streett, W., Tildesley, D., and Saville, G. (1978). Multiple time-step methods in molecular dynamics. *Molecular Physics*, **35**(3):639–648.

80. Phillips, S. C., Essex, J. W., and Edge, C. M. (2000). Digitally filtered molecular dynamics: The frequency specific control of molecular dynamics simulations. *Journal of Chemical Physics*, **112**(6):2586–2597.

81. Tozzini, V. (2005). Coarse-grained models for proteins. *Current Opinion in Structural Biology*, **15**(2):144–50.

82. Sansom, M. S. P., Scott, K. A., and Bond, P. J. (2008). Coarse-grained simulation: a high-throughput computational approach to membrane proteins. *Biochemical Society transactions*, **36**(Pt 1):27–32.

83. Ayton, G. S., Noid, W. G., and Voth, G. A. (2007). Multiscale modeling of biomolecular systems: in serial and in parallel. *Current Opinion in Structural Biology*, **17**(2):192–8.

84. de Groot, B. L., van Aalten, D. M., Scheek, R. M., Amadei, A., Vriend, G., and Berendsen, H. J. (1997). Prediction of protein conformational freedom from distance constraints. *Proteins*, **29**(2):240–51.

85. Tama, F. and Sanejouand, Y.-H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Engineering Design and Selection*, **14**(1):1–6.

86. Chennubhotla, C., Rader, A. J., Yang, L.-W., and Bahar, I. (2005). Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Physical Bology*, **2**(4):S173–80.

87. Doruker, P., Atilgan, A. R., and Bahar, I. (2000). Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins*, **40**(3):512–24.

88. Cheng, S. and Niv, M. Y. (2010). Molecular dynamics simulations and elastic network analysis of protein kinase B (Akt/PKB) inactivation. *Journal of Chemical Information and Modeling*, **50**(9):1602–10.

89. Spiwok, V., Lipovová, P., Skálová, T., Dušková, J., Dohnálek, J., Hašek, J., Russell, N., and Králová, B. (2007). Cold-active enzymes studied by comparative molecular dynamics simulation. *Journal of Molecular Modeling*, **13**(4):13.

90. Barreca, M. L., Lee, K. W., Chimirri, A., and Briggs, J. M. (2003). Molecular dynamics studies of the wild-type and double mutant HIV-1 integrase complexed with the 5CITEP inhibitor: mechanism for inhibition and drug resistance. *Biophysical Journal*, **84**(3):1450–63.

91. Perryman, A. L., Forli, S., Morris, G. M., Burt, C., Cheng, Y., Palmer, M. J., Whitby, K., McCammon, J. A., Phillips, C., and Olson, A. J. (2010). A dynamic model of HIV integrase inhibition and drug resistance. *Journal of Molecular Biology*, **397**(2):600–15.

92. Lücke, C., Fushman, D., Ludwig, C., Hamilton, J. A., Sacchettini, J. C., and Rüterjans, H. (1999). A comparative study of the backbone dynamics of two closely related lipid binding proteins: bovine heart fatty acid binding protein and porcine ileal lipid binding protein. *Molecular and Cellular Biochemistry*, **192**(1-2):109–21.

93. Gutiérrez-González, L. H., Ludwig, C., Hohoff, C., Rademacher, M., Hanhoff, T., Rüterjans, H., Spener, F., and Lücke, C. (2002). Solution structure and backbone dynamics of human epidermal-type fatty acid-binding protein (E-FABP). *The Biochemical Journal*, **364**(Pt 3):725–37.

94. Tai, K., Baaden, M., Murdock, S., Wu, B., Ng, M. H., Johnston, S., Boardman, R., Fangohr, H., Cox, K., Essex, J. W., and Sansom, M. S. P. (2007). Three hydrolases and a transferase: comparative analysis of active-site dynamics via the BioSimGrid database. *Journal of Molecular Graphics & Modelling*, **25**(6):896–902.

95. Tai, K., Murdock, S., Wu, B., Ng, M. H., Johnston, S., Fangohr, H., Cox, S. J., Jeffreys, P., Essex, J. W., and Sansom, M. S. P. (2004). BioSimGrid: towards a worldwide repository for biomolecular simulations. *Organic and Biomolecular Chemistry*, **2**(22):3219–21.

96. Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(27):11079–84.

97. Povolotskaya, I. S. and Kondrashov, F. A. (2010). Sequence space and the ongoing expansion of the protein universe. *Nature*, **465**(7300):922–6.

98. Armstrong, K. A. and Tidor, B. (2008). Computationally mapping sequence space to understand evolutionary protein engineering. *Biotechnology progress*, **24**(1):62–73.

99. Hou, J., Jun, S.-R., Zhang, C., and Kim, S.-H. (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(10):3651–6.

100. Redfern, O. C., Dessailly, B., and Orengo, C. A. (2008). Exploring the structure and function paradigm. *Current Opinion in Structural Biology*, **18**(3):394–402.

101. Koonin, E. V. and Galperin, M. Y. (2003) *Sequence - Evolution - Function* (Kluwer Academic).

102. Needleman, S. and WUNSCH, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteinsâŸĘ. *Journal of Molecular Biology*, **48**(3):443–453.

103. Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1):195–7.

104. Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(22):10915–9.

105. Dayhoff, M. O. and Schwartz, R. M. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, **5**(3):345–351.

106. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3):403–10.

107. Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17):3389–3402.

108. Lipman, D. and Pearson, W. (1985). Rapid and sensitive protein similarity searches. *Science*, **227**(4693):1435–1441.

109. Kriventseva, E. V., Fleischmann, W., Zdobnov, E. M., and Apweiler, R. (2001). CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Research*, **29**:33 – 36.

110. Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**(7):1575–84.

111. Krause, A., Stoye, J., and Vingron, M. (2000). The SYSTERS protein sequence cluster set. *Nucleic Acids Research*, **28**(1):270–2.

112. Keefe, A. D. and Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, **410**(6829):715–8.

113. Krause, A., Stoye, J., and Vingron, M. (2005). Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics*, **6**(1):15.

114. Frenkel, Z. M. and Trifonov, E. N. (2007). Evolutionary networks in the formatted protein sequence space. *Journal of Computational Biology*, **14**(8):1044–57.

115. Xia, Y. and Levitt, M. (2004). Simulating protein evolution in sequence and structure space. *Current Opinion in Structural Biology*, **14**(2):202–7.

116. Koehl, P. (2001). Protein structure similarities. *Current Opinion in Structural Biology*, **11**(3):348–353.

117. Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, **233**:123–138.

118. Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering Design and Selection*, **11**(9):739–747.

119. Gibrat, J. F., Madej, T., and Bryant, S. H. (1996). Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, **6**(3):377–85.

120. Taylor, W. R., Flores, T. P., and Orengo, C. A. (1994). Multiple protein structure alignment. *Protein Science*, **3**(10):1858–70.

121. Stark, A. and Russell, R. B. (2003). Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic acids research*, **31**(13):3341–4.

122. Konc, J. and Janezic, D. (2010). ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, **26**(9):1160–8.

123. Holm, L. and Sander, C. (1996). Mapping the protein universe. *Science*, **273**(5275):595–603.

124. Osadchy, M. and Kolodny, R. (2011). Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(30):12301–6.

125. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, **247**(4):536–40.

126. Orengo, C. A., Pearl, F. M. G., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L., and Thornton, J. M. (1999). The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Research*, **27**(1):275–279.

127. Holm, L. and Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research*, **22**(17):3600–9.

128. Skolnick, J., Arakaki, A. K., Lee, S. Y., and Brylinski, M. (2009). The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(37):15690–5.

129. Sadreyev, R. I., Kim, B.-H., and Grishin, N. V. (2009). Discrete-continuous duality of protein structure space. *Current Opinion in Structural biology*, **19**(3):321–8.

130. Pascual-García, A., Abia, D., Ortiz, A. R., and Bastolla, U. (2009). Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Computational Biology*, **5**(3):e1000331.

131. Raes, J., Harrington, E. D., Singh, A. H., and Bork, P. (2007). Protein function space: viewing the limits or limited by our view? *Current Opinion in Structural Biology*, **17**(3):362–9.

132. Erdin, S., Lisewski, A. M., and Lichtarge, O. (2011). Protein function prediction: towards integration of similarity metrics. *Current Opinion in Structural Biology*, **21**(2):180–8.

133. Sangar, V., Blankenberg, D. J., Altman, N., and Lesk, A. M. (2007). Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinformatics*, **8**(1):294.

134. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**(1):25–9.

135. Webb, E. C. (1992). Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. *Enzyme Nomenclature*.

136. Chagoyen, M., Carazo, J. M., and Pascual-Montano, A. (2008). Assessment of protein set coherence using functional annotations. *BMC Bioinformatics*, **9**(1):444.

137. Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E. N., Falcão, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9**:S4.

138. Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**(7):976–8.

139. Alvarez, M. A. and Yan, C. (2011). A graph-based semantic similarity measure for the gene ontology. *Journal of Bioinformatics and Computational Biology*, **9**(6):681–95.

140. Higdon, R., Louie, B., and Kolker, E. (2010) Modeling sequence and function similarity between proteins for protein functional annotation, in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing - HPDC '10*, page 499 (ACM Press, New York, New York, USA).

141. Louie, B., Bergen, S., Higdon, R., and Kolker, E. (2010). Quantifying protein function specificity in the gene ontology. *Standards in Genomic Sciences*, **2**(2):238–44.

142. Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of Molecular Biology*, **318**(2):595–608.

143. Zeldovich, K. B. and Shakhnovich, E. I. (2008). Understanding Protein Evolution: From Protein Physics to Darwinian Selection. *Annual Review of Physical Chemistry*, **59**:105–127.

144. Zhang, M. and Wang, W. (2003). Organization of signaling complexes by PDZ-domain scaffold proteins. *Accounts of Chemical Research*, **36**(7):530–8.

145. Valencia, A., Kjeldgaard, M., Pai, E. F., and Sander, C. (1991). GTPase domains of ras p21 oncogene protein and elongation factor Tu: analysis of three-dimensional structures, sequence families, and functional sites. *Proceedings of the National Academy of Sciences of the United States of America*, **88**(12):5443–7.

146. Shatsky, M., Nussinov, R., and Wolfson, H. J. (2006). Optimization of multiple-sequence alignment based on multiple-structure alignment. *Proteins*, **62**(1):209–17.

147. de Jong, R. M., Tiesinga, J. J. W., Rozeboom, H. J., Kalk, K. H., Tang, L., Janssen, D. B., and Dijkstra, B. W. (2003). Structure and mechanism of a bacterial haloalcohol dehalogenase: a new variation of the short-chain dehydrogenase/reductase fold without an NAD(P)H binding site. *The EMBO Journal*, **22**(19):4933–44.

148. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, **12**(2):85–94.

149. Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews. Molecular cell Biology*, **8**(12):995–1005.

150. Rigden, D. J. (2009) *Protein Structure to Function With Bioinformatics*.

151. Bornberg-Bauer, E. (1997). How are model protein structures distributed in sequence space? *Biophysical Journal*, **73**(5):2393–403.

152. Bornberg-Bauer, E. (1999). Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. *Proceedings of the National Academy of Sciences*, **96**(19):10689–10694.

153. Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, **24**(6):1501–1509.

154. Lau, K. F. and Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, **22**(10):3986–3997.

155. Helling, R., Li, H., Mélin, R., Miller, J., Wingreen, N., Zeng, C., and Tang, C. (2001). The designability of protein structures. *Journal of Molecular Graphics & Modelling*, **19**(1):157–67.

156. Wingreen, N. S., Li, H., and Tang, C. (2004). Designability and thermal stability of protein structures. *Polymer*, **45**(2):699–705.

157. He, Y., Chen, Y., Alexander, P., Bryan, P. N., and Orban, J. (2008). NMR structures of two designed proteins with high sequence identity but different fold and function. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(38):14412–7.

158. Bonneau, R. and Baker, D. (2001). Ab initio protein structure prediction: progress and prospects. *Annual Review of Biophysics and Bomolecular Structure*, **30**:173–89.

159. Goldsmith-Fischman, S. and Honig, B. (2003). Structural genomics: computational methods for structure analysis. *Protein Science*, **12**(9):1813–21.

160. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**(6381):86–9.

161. Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(29):11963–8.

162. Shah, I. and Hunter, L. (1997). Predicting enzyme function from sequence: a systematic appraisal. *Proc International Conference on Intelligent Systems for Molecular Biology*, **5**:276–83.

163. Pawlowski, K., Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000). Sensitive sequence comparison as protein function predictor. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 42–53.

164. Joshi, T. and Xu, D. (2007). Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics*, **8**(1):222.

165. Tian, W. and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, **333**(4):863–82.

166. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., and Ofran, Y. (2003). Automatic prediction of protein function. *Cellular and Molecular Life Sciences*, **60**(12):2637–50.

167. Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology*, **297**(1):233–49.

168. Kim, S.-H., Shin, D. H., Choi, I.-G., Schulze-Gahmen, U., Chen, S., and Kim, R. (2003). Structure-based functional inference in structural genomics. *Journal of Structural and Functional Genomics*, **4**(2-3):129–35.

169. Hegyi, H. and Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology*, **288**(1):147–64.

170. Ferrè, F., Ausiello, G., Zanzoni, A., and Helmer-Citterich, M. (2005). Functional annotation by identification of local surface similarities: a novel tool for structural genomics. *BMC Bioinformatics*, **6**(1):194.

171. Hvidsten, T. R., Laegreid, A., Kryshtafovych, A., Andersson, G., Fidelis, K., and Komorowski, J. (2009). A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PloS one*, **4**(7):e6266.

172. Hwang, K. Y., Chung, J. H., Kim, S. H., Han, Y. S., and Cho, Y. (1999). Structure-based identification of a novel NTPase from Methanococcus jannaschii. *Nature Structural Biology*, **6**(7):691–6.

173. Tokuriki, N. and Tawfik, D. S. (2009). Protein dynamism and evolvability. *Science*, **324**(5924):203–7.

174. Meier, S., Jensen, P. R., David, C. N., Chapman, J., Holstein, T. W., Grzesiek, S., and Ozbek, S. (2007). Continuous molecular evolution of protein-domain structures by single amino acid changes. *Current Biology*, **17**(2):173–8.

175. Kosloff, M. and Kolodny, R. (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins*, **71**(2):891–902.

176. Murzin, A. G. (2008). Biochemistry. Metamorphic proteins. *Science*, **320**(5884):1725–6.

177. Gerlt, J. A. and Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual Review of Biochemistry*, **70**:209–46.

178. Anantharaman, V., Aravind, L., and Koonin, E. V. (2003). Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Current Opinion in Chemical Biology*, **7**(1):12–20.

179. Gherardini, P. F. and Helmer-Citterich, M. (2008). Structure-based function prediction: approaches and applications. *Briefings in Functional Genomics & Proteomics*, **7**(4):291–302.

180. Wierenga, R. K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS letters*, **492**(3):193–8.

181. Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *Journal of Molecular Biology*, **321**(5):741–65.

182. Hasson, M. S. (1998). Evolution of an enzyme active site: The structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. *Proceedings of the National Academy of Sciences*, **95**(18):10396–10401.

183. Zen, A., Carnevale, V., Lesk, A. M., and Micheletti, C. (2008). Correspondences between low-energy modes in enzymes: dynamics-based alignment of enzymatic functional families. *Protein Science*, **17**(5):918–29.

184. Zen, A., de Chiara, C., Pastore, A., and Micheletti, C. (2009). Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains. *Bioinformatics*, **25**(15):1876–83.

185. Pang, A., Arinaminpathy, Y., Sansom, M. S. P., and Biggin, P. C. (2005). Comparative molecular dynamics–similar folds and similar motions? *Proteins*, **61**(4):809–22.

186. Hensen, U., Meyer, T., Haas, J., Rex, R., Vriend, G., and Grubmüller, H. (2012). Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PloS one*, **7**(5):e33931.

187. Keskin, O., Jernigan, R. L., and Bahar, I. (2000). Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophysical Journal*, **78**(4):2093–106.

188. Bahar, I., Erman, B., Haliloglu, T., and Jernigan, R. L. (1997). Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry*, **36**(44):13512–23.

189. Haliloglu, T., Bahar, I., and Erman, B. (1997). Gaussian Dynamics of Folded Proteins. *Physical Review Letters*, **79**(16):3090–3093.

190. Maguid, S., Fernández-Alberti, S., Parisi, G., and Echave, J. (2006). Evolutionary conservation of protein backbone flexibility. *Journal of Molecular Evolution*, **63**(4):448–57.

191. Maguid, S., Fernandez-Alberti, S., and Echave, J. (2008). Evolutionary conservation of protein vibrational dynamics. *Gene*, **422**(1-2):7–13.

192. Liu, Y. and Bahar, I. (2012). Sequence Evolution Correlates with Structural Dynamics. *Molecular Bology and Evolution*.

193. Whitley, M. J., Zhang, J., and Lee, A. L. (2008). Hydrophobic core mutations in CI2 globally perturb fast side-chain dynamics similarly without regard to position. *Biochemistry*, **47**(33):8566–76.

194. Verma, D., Jacobs, D. J., and Livesay, D. R. (2012). Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged. *PLoS Computational Biology*, **8**(3):e1002409.

195. Mittermaier, A. and Kay, L. E. (2004). The response of internal dynamics to hydrophobic core mutations in the SH3 domain from the Fyn tyrosine kinase. *Protein Science*, **13**(4):1088–99.

196. Jelen, F., Oleksy, A., Smietana, K., and Otlewski, J. (2003). PDZ domains - Common players in the cell signaling. *Acta Biochimica Polonica*, **50**(4):985–1017.

197. Harris, B. Z. and Lim, W. A. (2001). Mechanism and role of PDZ domains in signaling complex assembly. *Journal of Cell Science*, **114**(Pt 18):3219–31.

198. Jemth, P. and Gianni, S. (2007). PDZ domains: folding and binding. *Biochemistry*, **46**(30):8701–8.

199. Hung, A. Y. and Sheng, M. (2002). PDZ domains: structural modules for protein complex assembly. *The Journal of Biological Chemistry*, **277**(8):5699–702.

200. Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J.-H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D., and Sidhu, S. S. (2008). A specificity map for the PDZ domain family. *PLoS Biology*, **6**(9):e239.

201. Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A., and MacBeath, G. (2008). Predicting PDZ domain-peptide interactions from primary sequences. *Nature Biotechnology*, **26**(9):1041–5.

202. Ponting, C. P. (1997). Evidence for PDZ domains in bacteria, yeast, and plants. *Protein Science*, **6**(2):464–8.

203. Sakarya, O., Conaco, C., Egecioglu, O., Solla, S. A., Oakley, T. H., and Kosik, K. S. (2010). Evolutionary expansion and specialization of the PDZ domains. *Molecular Biology and Evolution*, **27**(5):1058–69.

204. Cesareni, G., Gimona, M., Sudol, M., and Yaffe, M. (2006) *Modular Protein Domains* (John Wiley & Sons).

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

206. Bray, D. (1998). Signaling complexes: biophysical constraints on intracellular communication. *Annual review of biophysics and biomolecular structure*, **27**:59–75.

207. Feng, W. and Zhang, M. (2009). Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. *Nature Reviews. Neuroscience*, **10**(2):87–99.

208. Sheng, M. and Sala, C. (2001). PDZ domains and the organization of supramolecular complexes. *Annual Review of Neuroscience*, **24**:1–29.

209. Brenman, J. E., Chao, D. S., Gee, S. H., McGee, A. W., Craven, S. E., Santillano, D. R., Wu, Z., Huang, F., Xia, H., Peters, M. F., Froehner, S. C., and Bredt, D. S. (1996). Interaction of nitric oxide synthase with the postsynaptic density protein PSD-95 and alpha1-syntrophin mediated by PDZ domains. *Cell*, **84**(5):757–67.

210. Hillier, B. J. (1999). Unexpected Modes of PDZ Domain Scaffolding Revealed by Structure of nNOS-Syntrophin Complex. *Science*, **284**(5415):812–815.

211. Lee, H.-J. and Zheng, J. J. (2010). PDZ domains and their binding partners: structure, specificity, and modification. *Cell Communication and Signaling*, **8**:8.

212. Subbaiah, V. K., Kranjec, C., Thomas, M., and Banks, L. (2011). PDZ domains: the building blocks regulating tumorigenesis. *The Biochemical Journal*, **439**(2):195–205.

213. Scannevin, R. H. and Huganir, R. L. (2000). Postsynaptic organization and regulation of excitatory synapses. *Nature Reviews. Neuroscience*, **1**(2):133–41.

214. Fanning, A. S. and Anderson, J. M. (1999). PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane. *The Journal of Clinical Investigation*, **103**(6):767–72.

215. Dong, H., Zhang, P., Song, I., Petralia, R. S., Liao, D., and Huganir, R. L. (1999). Characterization of the Glutamate Receptor-Interacting Proteins GRIP1 and GRIP2. *The Journal of Neuroscience*, **19**(16):6930–6941.

216. Srivastava, S., Osten, P., Vilim, F. S., Khatri, L., Inman, G., States, B., Daly, C., DeSouza, S., Abagyan, R., Valtschanoff, J. G., Weinberg, R. J., and Ziff, E. B. (1998). Novel anchorage of GluR2/3 to the postsynaptic density by the AMPA receptor-binding protein ABP. *Neuron*, **21**(3):581–91.

217. Xu, X. Z., Choudhury, A., Li, X., and Montell, C. (1998). Coordination of an array of signaling proteins through homo- and heteromeric interactions between PDZ domains and target proteins. *The Journal of Cell Biology*, **142**(2):545–55.

218. Nourry, C., Grant, S. G. N., and Borg, J.-P. (2003). PDZ domain proteins: plug and play! *Science's STKE*, **2003**(179):RE7.

219. Kaufmann, K., Shen, N., Mizoue, L., and Meiler, J. (2011). A physical model for PDZ-domain/peptide interactions. *Journal of Molecular Modeling*, **17**(2):315–24.

220. Christopherson, K. S. (1999). PSD-95 Assembles a Ternary Complex with the N-Methyl-D-aspartic Acid Receptor and a Bivalent Neuronal NO Synthase PDZ Domain. *Journal of Biological Chemistry*, **274**(39):27467–27473.

221. Petit, C. M., Zhang, J., Sapienza, P. J., Fuentes, E. J., and Lee, A. L. (2009). Hidden dynamic allostery in a PDZ domain. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(43):18249–54.

222. Fuentes, E. J., Der, C. J., and Lee, A. L. (2004). Ligand-dependent Dynamics and Intramolecular Signaling in a PDZ Domain. *Journal of Molecular Biology*, **335**(4):1105–1115.

223. Kong, Y. and Karplus, M. (2007). The signaling pathway of rhodopsin. *Structure*, **15**(5):611–23.

224. Im, Y. J., Lee, J. H., Park, S. H., Park, S. J., Rho, S.-H., Kang, G. B., Kim, E., and Eom, S. H. (2003). Crystal structure of the Shank PDZ-ligand complex reveals a class I PDZ interaction and a novel PDZ-PDZ dimerization. *The Journal of Biological Chemistry*, **278**(48):48099–104.

225. Utepbergenov, D. I., Fanning, A. S., and Anderson, J. M. (2006). Dimerization of the scaffolding protein ZO-1 through the second PDZ domain. *The Journal of Biological Chemistry*, **281**(34):24671–7.

226. Fanning, A. S., Lye, M. F., Anderson, J. M., and Lavie, A. (2007). Domain swapping within PDZ2 is responsible for dimerization of ZO proteins. *The Journal of Biological Chemistry*, **282**(52):37710–6.

227. Gallardo, R., Ivarsson, Y., Schymkowitz, J., Rousseau, F., and Zimmermann, P. (2010). Structural diversity of PDZ-lipid interactions. *Chembiochem : a European journal of chemical biology*, **11**(4):456–67.

228. Zhang, Y., Yeh, S., Appleton, B. A., Held, H. A., Kausalya, P. J., Phua, D. C. Y., Wong, W. L., Lasky, L. A., Wiesmann, C., Hunziker, W., and Sidhu, S. S. (2006). Convergent and divergent ligand specificity among PDZ domains of the LAP and zonula occludens (ZO) families. *The Journal of Biological Chemistry*, **281**(31):22299–311.

229. Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E., Zaslavskaia, L. A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. *Science*, **317**(5836):364–9.

230. Birrane, G., Chung, J., and Ladias, J. A. A. (2003). Novel mode of ligand recognition by the Erbin PDZ domain. *The Journal of Biological Chemistry*, **278**(3):1399–402.

231. Wong, H.-C., Bourdelas, A., Krauss, A., Lee, H.-J., Shao, Y., Wu, D., Mlodzik, M., Shi, D.-L., and Zheng, J. (2003). Direct binding of the PDZ domain of Dishevelled to a conserved internal sequence in the C-terminal region of Frizzled. *Molecular Cell*, **12**(5):1251–60.

232. van den Berk, L. C. J., Landi, E., Walma, T., Vuister, G. W., Dente, L., and Hendriks, W. J. A. J. (2007). An allosteric intramolecular PDZ-PDZ interaction modulates PTP-BL PDZ2 binding specificity. *Biochemistry*, **46**(47):13629–37.

233. Zhang, J., Petit, C. M., King, D. S., and Lee, A. L. (2011). Phosphorylation of a PDZ domain extension modulates binding affinity and interdomain interactions in postsynaptic density-95 (PSD-95) protein, a membrane-associated guanylate kinase (MAGUK). *The Journal of Biological Chemistry*, **286**(48):41776–85.

234. Verpy, E., Leibovici, M., Zwaenepoel, I., Liu, X. Z., Gal, A., Salem, N., Mansour, A., Blanchard, S., Kobayashi, I., Keats, B. J., Slim, R., and Petit, C. (2000). A defect in harmonin, a PDZ domain-containing protein expressed in the inner ear sensory hair cells, underlies Usher syndrome type 1C. *Nature Genetics*, **26**(1):51–5.

235. Boerkoel, C. F., Takashima, H., Stankiewicz, P., Garcia, C. A., Leber, S. M., Rhee-Morris, L., and Lupski, J. R. (2001). Periaxin mutations cause recessive Dejerine-Sottas neuropathy. *American Journal of Human Genetics*, **68**(2):325–33.

236. Cheng, J., Moyer, B. D., Milewski, M., Loffing, J., Ikeda, M., Mickle, J. E., Cutting, G. R., Li, M., Stanton, B. A., and Guggino, W. B. (2002). A Golgi-associated PDZ domain protein modulates cystic fibrosis transmembrane regulator plasma membrane expression. *The Journal of Biological Chemistry*, **277**(5):3520–9.

237. Dev, K. K. (2004). Making protein interactions druggable: targeting PDZ domains. *Nature Reviews. Drug discovery*, **3**(12):1047–56.

238. Javier, R. T. and Rice, A. P. (2011). Emerging theme: cellular PDZ proteins as common targets of pathogenic viruses. *Journal of virology*, **85**(22):11544–56.

239. Obenauer, J. C., Denson, J., Mehta, P. K., Su, X., Mukatira, S., Finkelstein, D. B., Xu, X., Wang, J., Ma, J., Fan, Y., Rakestraw, K. M., Webster, R. G., Hoffmann, E., Krauss, S., Zheng, J., Zhang, Z., and Naeve, C. W. (2006). Large-scale sequence analysis of avian influenza isolates. *Science*, **311**(5767):1576–80.

240. Hirbec, H., Francis, J. C., Lauri, S. E., Braithwaite, S. P., Coussen, F., Mulle, C., Dev, K. K., Coutinho, V., Meyer, G., Isaac, J. T. R., Collingridge, G. L., Henley, J. M., and Couthino, V. (2003). Rapid and differential regulation of AMPA and kainate receptors at hippocampal mossy fibre synapses by PICK1 and GRIP. *Neuron*, **37**(4):625–38.

241. Daw, M. I., Chittajallu, R., Bortolotto, Z. A., Dev, K. K., Duprat, F., Henley, J. M., Collingridge, G. L., and Isaac, J. T. (2000). PDZ proteins interacting with C-terminal GluR2/3 are involved in a PKC-dependent regulation of AMPA receptors at hippocampal synapses. *Neuron*, **28**(3):873–86.

242. Thorsen, T. S., Madsen, K. L., Rebola, N., Rathje, M., Anggono, V., Bach, A., Moreira, I. S., Stuhr-Hansen, N., Dyhring, T., Peters, D., Beuming, T., Huganir, R., Weinstein, H., Mulle, C., Strø mgaard, K., Rø nn, L. C. B., and Gether, U. (2010). Identification of a small-molecule inhibitor of the PICK1 PDZ domain that inhibits hippocampal LTP and LTD. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(1):413–8.

243. Shan, J., Shi, D.-L., Wang, J., and Zheng, J. (2005). Identification of a specific inhibitor of the dishevelled PDZ domain. *Biochemistry*, **44**(47):15495–15503.

244. Fujii, N., You, L., Xu, Z., Uematsu, K., Shan, J., He, B., Mikami, I., Edmondson, L. R., Neale, G., Zheng, J., Guy, R. K., and Jablons, D. M. (2007). An antagonist of dishevelled protein-protein interaction suppresses beta-catenin-dependent tumor cell growth. *Cancer Research*, **67**(2):573–9.

245. Shan, J. and Zheng, J. J. (2009). Optimizing Dvl PDZ domain inhibitor by exploring chemical space. *Journal of Computer-aided Molecular Design*, **23**(1):37–47.

246. Grandy, D., Shan, J., Zhang, X., Rao, S., Akunuru, S., Li, H., Zhang, Y., Alpatov, I., Zhang, X. A., Lang, R. A., Shi, D.-L., and Zheng, J. J. (2009). Discovery and characterization of a small molecule inhibitor of the PDZ domain of dishevelled. *The Journal of Biological Chemistry*, **284**(24):16256–63.

247. Shan, J., Zhang, X., Bao, J., Cassell, R., and Zheng, J. J. (2012). Synthesis of potent dishevelled PDZ domain inhibitors guided by virtual screening and NMR studies. *Chemical Biology and Drug Design*, **79**(4):376–83.

248. Niv, M. Y. and Weinstein, H. (2005). A Flexible Docking Procedure for the Exploration of Peptide Binding Selectivity to Known Structures and Homology Models of PDZ Domains. *Journal of the American Chemical Society*, **127**(40):14072–14079.

249. Gerek, Z. N. and Ozkan, S. B. (2010). A flexible docking scheme to explore the binding selectivity of PDZ domains. *Protein Science*, **19**(5):914–28.

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

251. Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, **9**(9):646–52.

252. Ponder, J. W. and Case, D. A. (2003). Force fields for protein simulations. *Advances in Protein Chemistry*, **66**:27–85.

253. Ryckaert, J.-P. and Bellemans, A. (1978). Molecular dynamics of liquid alkanes. *Faraday Discussions of the Chemical Society*, **66**:95.

254. Jorgensen, W. L. and Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, **110**(6):1657–1666.

255. Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D., and van Gunsteren, W. F. (2005). The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry*, **26**(16):1719–51.

256. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, **117**(19):5179–5197.

257. MacKerell, A. D., Bashford, D., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins âĂă. *The Journal of Physical Chemistry B*, **102**(18):3586–3616.

258. Verlet, L. (1967). Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, **159**(1):98–103.

259. Mesirov, J. P. and Schulten, K. (1996) *Mathematical Approaches to Biomolecular Structure and Dynamics* (Springer).

260. Andersen, H. C. (1983). Rattle: A âĂIJvelocityâĂİ version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, **52**(1):24–34.

261. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, **23**(3):327–341.

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

263. Berendsen, H., Postma, J., van Gunsteren, W., and Hermans, J. (1981). Interaction models for water in relation to protein hydration. *Intermolecular Forces*, pages 331 – 342.

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

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

266. Bernal, J. D. and Fowler, R. H. (1933). A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions. *The Journal of Chemical Physics*, **1**(8):515.

267. Stillinger, F. H. (1974). Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, **60**(4):1545.

268. Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, **112**(16):6127–6129.

269. Schaefer, M. and Karplus, M. (1996). A Comprehensive Analytical Treatment of Continuum Electrostatics. *The Journal of Physical Chemistry*, **100**(5):1578–1599.

270. Wesson, L. and Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, **1**(2):227–35.

271. Saito, M. (1994). Molecular dynamics simulations of proteins in solution: Artifacts caused by the cutoff approximation. *The Journal of Chemical Physics*, **101**(5):4055.

272. Schreiber, H. and Steinhauser, O. (1992). Molecular dynamics studies of solvated polypeptides: Why the cut-off scheme does not work. *Chemical Physics*, **168**(1):75–89.

273. Schreiber, H. and Steinhauser, O. (1992). Cutoff size does strongly influence molecular dynamics results on solvated polypeptides. *Biochemistry*, **31**(25):5856–5860.

274. Darden, T., Perera, L., Li, L., and Pedersen, L. (1999). New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*, **7**(3):R55–R60.

275. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, **81**(8):3684.

276. Steinhauser, M. O. and Hiermaier, S. (2009). A review of computational methods in materials science: examples from shock-wave and polymer physics. *International Journal of Molecular Sciences*, **10**(12):5135–216.

277. Leach, A. R. (2001) *Molecular Modelling: Principles and Applications* (Pearson Education).

278. Nocedal, J. (1980). Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation*, **35**(151):773 – 782.

279. Borg, I. and Groenen, P. J. F. (2005) *Modern Multidimensional Scaling: Theory And Applications*.

280. Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500):2323–6.

281. Scholkopf, B., Smola, A., and Müller, K.-R. (1999) Kernel principal component analysis, in *Advances in Kernel Methods - Support Vector Learning*, pages 327 – 352.

282. Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500):2319–23.

283. Hess, B. (2002). Convergence of sampling in protein simulations. *Physical Review E*, **65**(3).

284. Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, **77**(9):1905–1908.

285. Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D*, **47**(1):69–100.

286. Markov, A. A. (1954). The theory of algorithms. *Trudy Matematicheskogo Instituta imeni V. A. Steklova*, **42**:3–375.

287. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**(6):1087.

288. Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1):97–109.

289. Kinoshita, K., Murakami, Y., and Nakamura, H. (2007). eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Research*, **35**(Web Server issue):W398–402.

290. Angaran, S., Bock, M. E., Garutti, C., and Guerra, C. (2009). MolLoc: a web tool for the local structural alignment of molecular surfaces. *Nucleic Acids Research*, **37**(Web Server issue):W565–70.

291. Cascella, M. and Dal Peraro, M. (2009). Challenges and Perspectives in Biomolecular Simulations: From the Atomistic Picture to Multiscale Modeling. *CHIMIA International Journal for Chemistry*, **63**(1):14–18.

292. Zwier, M. C. and Chong, L. T. (2010). Reaching biological timescales with all-atom molecular dynamics simulations. *Current Opinion in Pharmacology*, **10**(6):745–52.

293. Salsbury, F. R. (2010). Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Current Opinion in Pharmacology*, **10**(6):738–44.

294. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, **4**(2):187–217.

295. Pearlman, D. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, **91**(1-3):1–41.

296. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, **26**(16):1781–802.

297. Gosling, J. and Mcgilton, H. (1996) *The Java Language Environment: A White Paper*.

298. Holland, R. C. G., Down, T. A., Pocock, M., Prlić, A., Huen, D., James, K., Foisy, S., Dräger, A., Yates, A., Heuer, M., and Schreiber, M. J. (2008). BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**(18):2096–7.

299. Novák, A., Miklós, I., Lyngsø, R., and Hein, J. (2008). StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*, **24**(20):2403–4.

300. Hanson, R. M. (2010). Jmol - a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography*, **43**(5):1250–1260.

301. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9):1189–91.

302. Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., and Wilson, A. (2011), Geneious v5.4, http://www.geneious.com/.

303. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*.

304. Hinsen, K. and Sadron, R. C. (2000). The molecular modeling toolkit: a new approach to molecular simulations. *Journal of Computational Chemistry*, **21**:79–85.

305. Romo, T. D. and Grossfield, A. (2009). LOOS: an extensible platform for the structural analysis of simulations. *Conf Proc IEEE Eng Med Biol Soc*, **2009**:2332–5.

306. Biasini, M., Mariani, V., Haas, J., Scheuber, S., Schenk, A. D., Schwede, T., and Philippsen, A. (2010). OpenStructure: a flexible software framework for computational structural biology. *Bioinformatics*, **26**(20):2626–8.

307. Eckstein, R., Loy, M., and Wood, D. (1998) *Java Swing* (O'Reilly Media).

308. Rumbaugh, J., Jacobson, R., and Booch, G. (1999) *The Unified Modelling Language Reference Manual*.

309. Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA:protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(16):6620–5.

310. Batagelj, V. and Mrvar, A. (1998). Pajek - Program for Large Network Analysis. *Connections*, **21**(2):1–11.

311. Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, **31**(1):7–15.

312. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, **32**(5):922–923.

313. Koehl, P. (2006) Protein Structure Classification, in *Reviews in Computational Chemistry, Volume 22* (edited by K B Lipkowitz, T R Cundari and V J Gillet), page 392 (John Wiley and Sons).

314. Wu, D. and Wu, Z. (2010). Superimposition of protein structures with dynamically weighted RMSD. *Journal of Molecular Modeling*, **16**(2):211–22.

315. Moore, M. S. and Blobel, G. (1994). A G protein involved in nucleocytoplasmic transport: the role of Ran. *Trends in Biochemical Sciences*, **19**(5):211–216.

316. Heinig, M. and Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, **32**(Web Server issue):W500–2.

317. Ng, M., Johnston, S., Wu, B., Murdock, S., Tai, K., Fangohr, H., Cox, S., Essex, J., Sansom, M., and Jeffreys, P. (2006). BioSimGrid: Grid-enabled biomolecular simulation data storage and analysis. *Future Generation Computer Systems*, **22**(6):657–664.

318. Dodson, G. and Verma, C. S. (2006). Protein flexibility: its role in structure and mechanism revealed by molecular simulations. *Cellular and Molecular Life Sciences*, **63**(2):207–19.

319. Hammes-Schiffer, S. and Benkovic, S. J. (2006). Relating protein motion to catalysis. *Annual Review of Biochemistry*, **75**:519–41.

320. Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, **13**(3):373–80.

321. Hall, B. A., Kaye, S. L., Pang, A., Perera, R., and Biggin, P. C. (2007). Characterization of protein conformational states by normal-mode frequencies. *Journal of the American Chemical Society*, **129**(37):11394–401.

322. Dhulesia, A., Gsponer, J., and Vendruscolo, M. (2008). Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a PDZ domain protein. *Journal of the American Chemical Society*, **130**(28):8931–9.

323. Cox, K. and Sansom, M. S. P. (2009). One membrane protein, two structures and six environments: a comparative molecular dynamics simulation study of the bacterial outer membrane protein PagP. *Molecular Membrane Biology*, **26**(4):205–14.

324. Lama, D. and Sankararamakrishnan, R. (2008). Anti-apoptotic Bcl-XL protein in complex with BH3 peptides of pro-apoptotic Bak, Bad, and Bim proteins: comparative molecular dynamics simulations. *Proteins*, **73**(2):492–514.

325. Yaneva, R., Springer, S., and Zacharias, M. (2009). Flexibility of the MHC class II peptide binding cleft in the bound, partially filled, and empty states: a molecular dynamics simulation study. *Biopolymers*, **91**(1):14–27.

326. Zacharias, M. and Springer, S. (2004). Conformational flexibility of the MHC class I alpha1-alpha2 domain in peptide bound and free states: a molecular dynamics simulation study. *Biophysical Journal*, **87**(4):2203–14.

327. Papaleo, E., Riccardi, L., Villa, C., Fantucci, P., and De Gioia, L. (2006). Flexibility and enzymatic cold-adaptation: a comparative molecular dynamics investigation of the elastase family. *Biochimica et biophysica acta*, **1764**(8):1397–406.

328. Maguid, S., Fernandez-Alberti, S., Ferrelli, L., and Echave, J. (2005). Exploring the common dynamics of homologous proteins. Application to the globin family. *Biophysical Journal*, **89**(1):3–13.

329. Potestio, R., Aleksiev, T., Pontiggia, F., Cozzini, S., and Micheletti, C. (2010). ALADYN: a web server for aligning proteins by matching their large-scale motion. *Nucleic acids research*, **38**(Web Server issue):W41–5.

330. Holm, L. and Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**(6):566–7.

331. Aung, Z. and Tan, K.-L. (2006). MatAlign: precise protein structure comparison by matrix alignment. *Journal of Bioinformatics and Computational Biology*, **4**(6):1197–216.

332. Johnson, D. S., Aragon, C. R., McGeoch, L. A., and Schevon, C. (1989). Optimization by Simulated Annealing: An Experimental Evaluation; Part I, Graph Partitioning. *Operations Research*, **37**(6):865–892.

333. Beique, J.-C. and Andrade, R. (2003). PSD-95 regulates synaptic transmission and plasticity in rat cerebral cortex. *The Journal of Physiology*, **546**(Pt 3):859–67.

334. De Los Rios, P., Cecconi, F., Pretre, A., Dietler, G., Michielin, O., Piazza, F., and Juanico, B. (2005). Functional dynamics of PDZ binding domains: a normal-mode analysis. *Biophysical Journal*, **89**(1):14–21.

335. Schneider, T. R. (2000). Objective comparison of protein structures: error-scaled difference distance matrices. *Acta Crystallographica. Section D, Biological Crystallography*, **56**(Pt 6):714–21.

336. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**(6):276–7.

337. Liao, D. I., Qian, J., Chisholm, D. A., Jordan, D. B., and Diner, B. A. (2000). Crystal structures of the photosystem II D1 C-terminal processing protease. *Nature Structural Biology*, **7**(9):749–53.

338. Krojer, T., Garrido-Franco, M., Huber, R., Ehrmann, M., and Clausen, T. (2002). Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. *Nature*, **416**(6879):455–9.

339. Wilken, C., Kitzing, K., Kurzbauer, R., Ehrmann, M., and Clausen, T. (2004). Crystal structure of the DegS stress sensor: How a PDZ domain recognizes misfolded protein and activates a protease. *Cell*, **117**(4):483–94.

340. Done, S. H., Brannigan, J. A., Moody, P. C., and Hubbard, R. E. (1998). Ligand-induced conformational change in penicillin acylase. *Journal of Molecular Biology*, **284**(2):463–75.

341. Calderone, V., Folli, C., Marchesani, A., Berni, R., and Zanotti, G. (2002). Identification and structural analysis of a zebrafish apo and holo cellular retinol-binding protein. *Journal of Molecular Biology*, **321**(3):527–35.

342. Takeda, M., Ogino, S., Umemoto, R., Sakakura, M., Kajiwara, M., Sugahara, K. N., Hayasaka, H., Miyasaka, M., Terasawa, H., and Shimada, I. (2006). Ligand-induced structural changes of the CD44 hyaluronan-binding domain revealed by NMR. *The Journal of Biological Chemistry*, **281**(52):40089–95.

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

344. Csermely, P., Palotai, R., and Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in Biochemical Sciences*, **35**(10):539–46.

345. Hammes, G. G., Chang, Y.-C., and Oas, T. G. (2009). Conformational selection or induced fit: a flux description of reaction mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(33):13737–41.

346. Silva, D.-A., Bowman, G. R., Sosa-Peinado, A., and Huang, X. (2011). A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS Computational Biology*, **7**(5):e1002054.

347. Nobeli, I., Favia, A. D., and Thornton, J. M. (2009). Protein promiscuity and its implications for biotechnology. *Nature Biotechnology*, **27**(2):157–67.

348. Fernández, A., Tawfik, D. S., Berkhout, B., Sanders, R., Kloczkowski, A., Sen, T., and Jernigan, B. (2005). Protein promiscuity: drug resistance and native functions–HIV-1 case. *Journal of Bomolecular Structure and Dynamics*, **22**(6):615–24.

349. Muralidhara, B. K., Sun, L., Negi, S., and Halpert, J. R. (2008). Thermodynamic fidelity of the mammalian cytochrome P450 2B4 active site in binding substrates and inhibitors. *Journal of Molecular Biology*, **377**(1):232–45.

350. Skopalík, J., Anzenbacher, P., and Otyepka, M. (2008). Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *The Journal of Physical Chemistry. B*, **112**(27):8165–73.

351. Thorpe, I. F. and Brooks, C. L. (2007). Molecular evolution of affinity and flexibility in the immune system. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(21):8821–6.

352. Shepherd, T. R., Hard, R. L., Murray, A. M., Pei, D., and Fuentes, E. J. (2011). Distinct ligand specificity of the Tiam1 and Tiam2 PDZ domains. *Biochemistry*, **50**(8):1296–308.

353. Gee, S. H., Quenneville, S., Lombardo, C. R., and Chabot, J. (2000). Single-amino acid substitutions alter the specificity and affinity of PDZ domains for their ligands. *Biochemistry*, **39**(47):14638–46.

354. Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., and Ranganathan, R. (2005). Natural-like function in artificial WW domains. *Nature*, **437**(7058):579–83.

355. Panni, S., Dente, L., and Cesareni, G. (2002). In vitro evolution of recognition specificity mediated by SH3 domains reveals target recognition rules. *The Journal of Biological Chemistry*, **277**(24):21666–74.

356. Gerek, Z. N., Keskin, O., and Ozkan, S. B. (2009). Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior. *Proteins*, **77**(4):796–811.

357. Basdevant, N., Weinstein, H., and Ceruso, M. (2006). Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. *Journal of the American Chemical Society*, **128**(39):12766–77.

358. Logan, C. Y. and Nusse, R. (2004). The Wnt signaling pathway in development and disease. *Annual Review of Cell and Developmental Biology*, **20**:781–810.

359. Widelitz, R. (2005). Wnt signaling through canonical and non-canonical pathways: recent progress. *Growth Factors*, **23**(2):111–6.

360. Habas, R. and Dawid, I. B. (2005). Dishevelled and Wnt signaling: is the nucleus the final frontier? *Journal of Biology*, **4**(1):2.

361. Wallingford, J. B. and Habas, R. (2005). The developmental biology of Dishevelled: an enigmatic protein governing cell fate and cell polarity. *Development*, **132**(20):4421–36.

362. Śmietana, K., Mateja, A., Krezel, A., and Otlewski, J. (2011). PDZ domain from Dishevelled – a specificity study. *Acta Biochimica Polonica*, **58**(2):243–9.

363. Schultz, J. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences*, **95**(11):5857–5864.

364. Zhang, Y., Appleton, B. A., Wiesmann, C., Lau, T., Costa, M., Hannoush, R. N., and Sidhu, S. S. (2009). Inhibition of Wnt signaling by Dishevelled PDZ peptides. *Nature Chemical Biology*, **5**(4):217–9.

365. Lee, H.-J., Wang, N. X., Shi, D.-L., and Zheng, J. J. (2009). Sulindac inhibits canonical Wnt signaling by blocking the PDZ domain of the protein Dishevelled. *Angewandte Chemie*, **48**(35):6448–52.

366. Borg, J. P., Marchetto, S., Le Bivic, A., Ollendorff, V., Jaulin-Bastard, F., Saito, H., Fournier, E., Adélaïde, J., Margolis, B., and Birnbaum, D. (2000). ERBIN: a basolateral PDZ protein that interacts with the mammalian ERBB2/HER2 receptor. *Nature Cell Biology*, **2**(7):407–14.

367. Olayioye, M. A., Neve, R. M., Lane, H. A., and Hynes, N. E. (2000). The ErbB signaling network: receptor heterodimerization in development and cancer. *The EMBO Journal*, **19**(13):3159–67.

368. Hynes, N. E., Horsch, K., Olayioye, M. A., and Badache, A. (2001). The ErbB receptor tyrosine family as signal integrators. *Endocrine-related cancer*, **8**(3):151–9.

369. Bublil, E. M. and Yarden, Y. (2007). The EGF receptor family: spearheading a merger of signaling and therapeutics. *Current Opinion in Cell Biology*, **19**(2):124–34.

370. Yu, D. and Hung, M. C. (2000). Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. *Oncogene*, **19**(53):6115–21.

371. Bryant, P. J. and Huwe, A. (2000). LAP proteins: what's up with epithelia? *Nature Cell Biology*, **2**(8):E141–3.

372. Jaulin-Bastard, F., Saito, H., Le Bivic, A., Ollendorff, V., Marchetto, S., Birnbaum, D., and Borg, J. P. (2001). The ERBB2/HER2 receptor differentially interacts with ERBIN and PICK1 PSD-95/DLG/ZO-1 domain proteins. *The Journal of Biological Chemistry*, **276**(18):15256–63.

373. Huang, Y. Z., Wang, Q., Xiong, W. C., and Mei, L. (2001). Erbin is a protein concentrated at postsynaptic membranes that interacts with PSD-95. *The Journal of Biological Chemistry*, **276**(22):19318–26.

374. Lebeau, S., Masouyé, I., Berti, M., Augsburger, E., Saurat, J.-H., Borradori, L., and Fontao, L. (2005). Comparative analysis of the expression of ERBIN and Erb-B2 in normal human skin and cutaneous carcinomas. *The British Journal of Dermatology*, **152**(6):1248–55.

375. Liu, N., Zhang, J., Zhang, J., Liu, S., Liu, Y., and Zheng, D. (2008). Erbin-regulated sensitivity of MCF-7 breast cancer cells to TRAIL via ErbB2/AKT/NF-kappaB pathway. *Journal of Biochemistry*, **143**(6):793–801.

376. Laura, R. P., Witt, A. S., Held, H. A., Gerstner, R., Deshayes, K., Koehler, M. F. T., Kosik, K. S., Sidhu, S. S., and Lasky, L. A. (2002). The Erbin PDZ domain binds with high affinity and specificity to the carboxyl termini of delta-catenin and ARVCF. *The Journal of Biological Chemistry*, **277**(15):12906–14.

377. Jaulin-Bastard, F., Arsanto, J.-P., Le Bivic, A., Navarro, C., Vély, F., Saito, H., Marchetto, S., Hatzfeld, M., Santoni, M.-J., Birnbaum, D., and Borg, J.-P. (2002). Interaction between Erbin and a Catenin-related protein in epithelial cells. *The Journal of Biological Chemistry*, **277**(4):2869–75.

378. Izawa, I., Nishizawa, M., Tomono, Y., Ohtakara, K., Takahashi, T., and Inagaki, M. (2002). ERBIN associates with p0071, an armadillo protein, at cell-cell junctions of epithelial cells. *Genes to cells: devoted to molecular & cellular mechanisms*, **7**(5):475–85.

379. Arikkath, J., Israely, I., Tao, Y., Mei, L., Liu, X., and Reichardt, L. F. (2008). Erbin controls dendritic morphogenesis by regulating localization of delta-catenin. *The Journal of Neuroscience*, **28**(28):7047–56.

380. Dai, P., Xiong, W. C., and Mei, L. (2006). Erbin inhibits RAF activation by disrupting the sur-8-Ras-Raf complex. *The Journal of Biological Chemistry*, **281**(2):927–33.

381. Legouis, R., Jaulin-Bastard, F., Schott, S., Navarro, C., Borg, J.-P., and Labouesse, M. (2003). Basolateral targeting by leucine-rich repeat domains in epithelial cells. *EMBO reports*, **4**(11):1096–1100.

382. Kelker, M. S., Dancheck, B., Ju, T., Kessler, R. P., Hudak, J., Nairn, A. C., and Peti, W. (2007). Structural basis for spinophilin-neurabin receptor interaction. *Biochemistry*, **46**(9):2333–44.

383. Feng, W., Fan, J.-S., Jiang, M., Shi, Y.-W., and Zhang, M. (2002). PDZ7 of glutamate receptor interacting protein binds to its target via a novel hydrophobic surface area. *The Journal of Biological Chemistry*, **277**(43):41140–6.

384. Mao, L., Takamiya, K., Thomas, G., Lin, D.-T., and Huganir, R. L. (2010). GRIP1 and 2 regulate activity-dependent AMPA receptor recycling via exocyst complex interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(44):19038–43.

385. Mejias, R., Adamczyk, A., Anggono, V., Niranjan, T., Thomas, G. M., Sharma, K., Skinner, C., Schwartz, C. E., Stevenson, R. E., Fallin, M. D., Kaufmann, W., Pletnikov, M., Valle, D., Huganir, R. L., and Wang, T. (2011). Gain-of-function glutamate receptor interacting protein 1 variants alter GluA2 recycling and surface distribution in patients with autism. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(12):4920–5.

386. Ye, B., Liao, D., Zhang, X., Zhang, P., Dong, H., and Huganir, R. L. (2000). GRASP-1: a neuronal RasGEF associated with the AMPA receptor/GRIP complex. *Neuron*, **26**(3):603–17.

387. Azevedo, A. W. and Rieke, F. (2011). Experimental protocols alter phototransduction: the implications for retinal processing at visual threshold. *The Journal of Neuroscience*, **31**(10):3670–82.

388. Ranganathan, R., Harris, W., and Zuker, C. (1991). The molecular genetics of invertebrate phototransduction. *Trends in Neurosciences*, **14**(11):486–493.

389. Zuker, C. S. (1996). The biology of vision in Drosophila. *Proceedings of the National Academy of Sciences*, **93**(2):571–576.

390. Scott, K. and Zuker, C. (1997). Lights out: deactivation of the phototransduction cascade. *Trends in Biochemical Sciences*, **22**(9):350–4.

391. Scott, K. and Zuker, C. S. (1998). Assembly of the Drosophila phototransduction cascade into a signalling complex shapes elementary responses. *Nature*, **395**(6704):805–8.

392. Kimple, M. E., Siderovski, D. P., and Sondek, J. (2001). Functional relevance of the disulfide-linked complex of the N-terminal PDZ domain of InaD with NorpA. *The EMBO Journal*, **20**(16):4414–22.

393. Tsunoda, S., Sierralta, J., Sun, Y., Bodner, R., Suzuki, E., Becker, A., Socolich, M., and Zuker, C. S. (1997). A multivalent PDZ-domain protein assembles signalling complexes in a G-protein-coupled cascade. *Nature*, **388**(6639):243–9.

394. van Huizen, R., Miller, K., Chen, D. M., Li, Y., Lai, Z. C., Raab, R. W., Stark, W. S., Shortridge, R. D., and Li, M. (1998). Two distantly positioned PDZ domains mediate multivalent INAD-phospholipase C interactions essential for G protein-coupled signaling. *The EMBO Journal*, **17**(8):2285–97.

395. Z Xu, X., Choudhury, A., Li, X., and Montell, C. (1998). Coordination of an array of signaling proteins through homo- and heteromeric interactions between PDZ domains and target proteins. *Journal of Cell Biology*, **142**(2):545–55.

396. Wes, P. D., Xu, X. Z., Li, H. S., Chien, F., Doberstein, S. K., and Montell, C. (1999). Termination of phototransduction requires binding of the NINAC myosin III and the PDZ protein INAD. *Nature Neuroscience*, **2**(5):447–53.

397. Bähner, M., Sander, P., Paulsen, R., and Huber, A. (2000). The visual G protein of fly photoreceptors interacts with the PDZ domain assembled INAD signaling complex via direct binding of activated Galpha(q) to phospholipase cbeta. *The Journal of Biological Chemistry*, **275**(4):2901–4.

398. Hicks, J. L., Liu, X., and Williams, D. S. (1996). Role of the ninaC proteins in photoreceptor cell structure: ultrastructure of ninaC deletion mutants and binding to actin filaments. *Cell Motility and the Cytoskeleton*, **35**(4):367–79.

399. Saras, J., Claesson-Welsh, L., Heldin, C. H., and Gonez, L. J. (1994). Cloning and characterization of PTPL1, a protein tyrosine phosphatase with similarities to cytoskeletal-associated proteins. *The Journal of Biological Chemistry*, **269**(39):24082–9.

400. Lee, S. H., Shin, M. S., Park, W. S., Kim, S. Y., Kim, H. S., Lee, J. H., Han, S. Y., Lee, H. K., Park, J. Y., Oh, R. R., Jang, J. J., Lee, J. Y., and Yoo, N. J. (1999). Immunohistochemical localization of FAP-1, an inhibitor of Fas-mediated apoptosis, in normal and neoplastic human tissues. *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica*, **107**(12):1101–8.

401. Erdmann, K. S. (2003). The protein tyrosine phosphatase PTP-Basophil/Basophil-like. Interacting proteins and molecular functions. *European Journal of Biochemistry / FEBS*, **270**(24):4789–98.

402. Erdmann, K. S., Kuhlmann, J., Lessmann, V., Herrmann, L., Eulenburg, V., Müller, O., and Heumann, R. (2000). The Adenomatous Polyposis Coli-protein (APC) interacts with the protein tyrosine phosphatase PTP-BL via an alternatively spliced PDZ domain. *Oncogene*, **19**(34):3894–901.

403. Lockless, S. W. (1999). Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, **286**(5438):295–299.

404. Gianni, S., Walma, T., Arcovito, A., Calosci, N., Bellelli, A., Engström, A., Travaglini-Allocatelli, C., Brunori, M., Jemth, P., and Vuister, G. W. (2006). Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering. *Structure*, **14**(12):1801–9.

405. Walma, T., Aelen, J., Nabuurs, S. B., Oostendorp, M., van den Berk, L., Hendriks, W., and Vuister, G. W. (2004). A Closed Binding Pocket and Global Destabilization Modify the Binding Properties of an Alternatively Spliced Form of the Second PDZ Domain of PTP-BL. *Structure*, **12**(1):11–20.

406. Grembecka, J., Cierpicki, T., Devedjiev, Y., Derewenda, U., Kang, B. S., Bushweller, J. H., and Derewenda, Z. S. (2006). The binding of the PDZ tandem of syntenin to target proteins. *Biochemistry*, **45**(11):3674–83.

407. Kang, B. S., Cooper, D. R., Jelen, F., Devedjiev, Y., Derewenda, U., Dauter, Z., Otlewski, J., and Derewenda, Z. S. (2003). PDZ tandem of human syntenin: crystal structure and functional properties. *Structure*, **11**(4):459–68.

408. Chi, C. N., Bach, A., Engström, A., Wang, H., Stromgaard, K., Gianni, S., and Jemth, P. (2009). A sequential binding mechanism in a PDZ domain. *Biochemistry*, **48**(30):7089–97.

409. Tzeng, S.-R. and Kalodimos, C. G. (2009). Dynamic activation of an allosteric regulatory protein. *Nature*, **462**(7271):368–72.

410. Koshland, D. E. (1998). Conformational changes: how small is big enough? *Nature Medicine*, **4**(10):1112–4.

411. Changeux, J.-P. and Edelstein, S. J. (2005). Allosteric mechanisms of signal transduction. *Science*, **308**(5727):1424–8.

412. Bray, D. and Duke, T. (2004). Conformational spread: the propagation of allosteric states in large multiprotein complexes. *Annual Review of Biophysics and Biomolecular Structure*, **33**(1):53–73.

413. Cooper, A. and Dryden, D. T. (1984). Allostery without conformational change. A plausible model. *European Biophysics Journal*, **11**(2):103–9.

414. Wand, A. J. (2001). On the dynamic origins of allosteric activation. *Science*, **293**(5534):1395.

415. Freire, E. (2000). Can allosteric regulation be predicted from structure? *Proceedings of the National Academy of Sciences of the United States of America*, **97**(22):11680–2.

416. Pan, H., Lee, J. C., and Hilser, V. J. (2000). 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*, **97**(22):12020–5.

417. Stevens, S. Y., Sanker, S., Kent, C., and Zuiderweg, E. R. (2001). Delineation of the allosteric mechanism of a cytidylyltransferase exhibiting negative cooperativity. *Nature Structural Biology*, **8**(11):947–52.

418. Mäler, L., Blankenship, J., Rance, M., and Chazin, W. J. (2000). Site-site communication in the EF-hand Ca2+-binding protein calbindin D9k. *Nature Structural Biology*, **7**(3):245–50.

419. Lee, A. L., Kinnear, S. A., and Wand, A. J. (2000). Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nature Structural Biology*, **7**(1):72–7.

420. Kern, D. and Zuiderweg, E. R. P. (2003). The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, **13**(6):748–57.

421. Tzeng, S.-R. and Kalodimos, C. G. (2011). Protein dynamics and allostery: an NMR view. *Current Opinion in Structural Biology*, **21**(1):62–7.

422. Gunasekaran, K., Ma, B., and Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins*, **57**(3):433–43.

423. Swain, J. F. and Gierasch, L. M. (2006). The changing landscape of protein allostery. *Current Opinion in Structural Biology*, **16**(1):102–8.

424. Stacklies, W., Xia, F., and Gräter, F. (2009). Dynamic allostery in the methionine repressor revealed by force distribution analysis. *PLoS Computational Biology*, **5**(11):e1000574.

425. Böde, C., Kovács, I. A., Szalay, M. S., Palotai, R., Korcsmáros, T., and Csermely, P. (2007). Network analysis of protein dynamics. *FEBS letters*, **581**(15):2776–82.

426. Doncheva, N. T., Klein, K., Domingues, F. S., and Albrecht, M. (2011). Analyzing and visualizing residue networks of protein structures. *Trends in Biochemical Sciences*, **36**(4):179–82.

427. Vishveshwara, S., Ghosh, A., and Hansia, P. (2009). Intra and Inter-Molecular Communications Through Protein Structure Network. *Current Protein and Peptide Science*, **10**(2):15.

428. Jacobs, D. J., Rader, A. J., Kuhn, L. A., and Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins*, **44**(2):150–65.

429. Dokholyan, N. V., Li, L., Ding, F., and Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(13):8637–41.

430. Fang, Y., Ma, D., Li, M., Wen, Z., and Diao, Y. (2010). Investigation of the proteins folding rates and their properties of amino acid networks. *Chemometrics and Intelligent Laboratory Systems*, **101**(2):123–129.

431. Peng, S.-L. and Tsay, Y.-W. (2010). Measuring protein structural similarity by maximum common edge subgraphs. *Intelligent Computing Theories and Applications*, **6216**:100–107.

432. Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N., and Sarawagi, S. (2003). Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *Journal of Molecular Biology*, **326**(3):955–78.

433. Huan, J., Wang, W., Washington, A., Prins, J., Shah, R., and Tropsha, A. (2004). Accurate classification of protein structural families using coherent subgraph analysis. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 411–22.

434. del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Science*, **15**(9):2120–8.

435. Brinda, K. V. and Vishveshwara, S. (2005). A network representation of protein structures: implications for protein stability. *Biophysical Journal*, **89**(6):4159–70.

436. Alves, N. and Martinez, A. (2007). Inferring topological features of proteins from amino acid residue networks. *Physica A: Statistical Mechanics and its Applications*, **375**(1):336–344.

437. Atilgan, A. R., Akan, P., and Baysal, C. (2004). Small-world communication of residues and significance for protein dynamics. *Biophysical Journal*, **86**(1 Pt 1):85–91.

438. Bagler, G. and Sinha, S. (2005). Network properties of protein structures. *Physica A: Statistical Mechanics and its Applications*, **346**(1-2):27–33.

439. del Sol, A., Fujihashi, H., and O'Meara, P. (2005). Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, **21**(8):1311–5.

440. Greene, L. H. and Higman, V. A. (2003). Uncovering network systems within protein structures. *Journal of Molecular Biology*, **334**(4):781–91.

441. del Sol, A. and Carbonell, P. (2007). The modular organization of domain structures: insights into protein-protein binding. *PLoS Computational Biology*, **3**(12):e239.

442. Estrada, E. (2010). Universality in protein residue networks. *Biophysical Journal*, **98**(5):890–900.

443. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, **344**(4):1135–46.

444. Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**(1):269–271.

445. Freeman, L. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, **40**(1):35–41.

446. Kong, Y. and Karplus, M. (2009). Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. *Proteins*, **74**(1):145–54.

447. Mount, D. W. (2004) *Bioinformatics: Sequence and Genome Analysis*, vol. 21 (CSHL Press).

448. Brüschweiler, R. (2003). Efficient RMSD measures for the comparison of two molecular ensembles. Root-mean-square deviation. *Proteins*, **50**(1):26–34.