Internal Dynamics and Energetics During Enzyme Catalysis

Arvind Ramanathan

CMU-CB-10-102
May 04, 2010

Lane Center for Computational Biology
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Christopher J. Langmead, Co-chair
Pratul K. Agarwal, Co-chair
Ivet Bahar
Chakra S. Chennubhotla

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Keywords: Enzyme catalysis, quasi-anharmonic analysis, dynamic tensor analysis, data-mining, Rossmann Fold

To thatha*, who encouraged the pursuit of excellence.*

Abstract

Proteins have evolved to perform their targeted biochemical function precisely and efficiently. Growing evidence from experiments and computational approaches suggests an intimate synergy between an enzyme's structure, intrinsic dynamics and biochemical function. In this thesis, we investigate the role of intrinsic dynamics in enzyme catalysis by developing novel theoretical and computational techniques and using extensive atomistic level molecular dynamics simulations.

First, we show that there is significant similarity in collective conformational fluctuations during an enzyme's reaction-cycle. In particular, for the enzyme cyclophilin A, a peptidyl-prolyl isomerase, we show that there is substantial overlap (65%) in the dynamics before, during and after the catalytic step. Second, we show that dynamics associated with the catalytic step is evolutionarily conserved in multiple enzymes catalyzing the same biochemical reaction, even when they do not share a common fold. Finally, we show that there is a remarkable similarity in the fluctuations coupled to the catalytic step for several members of a super-family of enzymes sharing a common mechanistic substep in the reaction mechanism. The Rossmann fold family of enzymes investigated reveals the presence of three specific regions shared by all family members that exhibit collective fluctuations coupled to the catalytic step. These regions show the presence of a network formed by hydrogen bonds and hydrophobic interactions extending from the flexible surface regions all the way to the active site.

Our results indicate that intrinsic dynamics coupled to the catalytic step of enzymes may have imposed selective pressure over the course of evolution to promote biochemical function. These observations may have far-reaching implications in understanding how enzymes have evolved and may potentially serve as guiding principles for designing novel enzymes in industrial and therapeutic applications.

Acknowledgments

jaadyam dhiyO harati sinchati vaachi satyam
maanOnnatim dishati paapamapaakarOthi —
chetah prasaadayathi dikshu tanOthi keerthim
satsangathiH kathaya kim na karOthi pumsaam ——

So goes an (g)old saying (*subhaashita*) in Sanskrit. Translated, it means: "it removes lethargy from intellect, invests truth in one's speech, enhances one's greatness, casts off sins, clarifies doubts and spreads fame far and wide; tell me what the company of the high-souled persons does not provide a man?" My Ph.D. journey can be summarized by this saying. I am fortunate to have worked along with some of the brightest minds in the world.

First, I would like to acknowledge my advisers: Chris Langmead at CMU and Pratul K. Agarwal at ORNL. Chris has been always encouraging and supportive of my research interests since I joined his group. He has proved to be one of the easiest persons to work with and I have typically enjoyed my discussions with him. I would like to thank him for being there through out this exciting journey.

Pratul, in spite of being over at Oak Ridge, has constantly helped me through good times and bad and has always been enthusiastic in pursuing my dreams. Pratul's constant words of support and encouragement have proven to be more that of a discerning friend than that of an adviser. Without him, it would have been nearly impossible to appreciate the vast and exciting field of enzyme catalysis.

I would like to thank my thesis committee members: Drs. Ivet Bahar and Chakra Chennubhotla. Dr. Bahar's comments have been insightful and instructive in helping me shape my thoughts. Chakra has been a source of inspiration (and perspiration). It has been a pleasure to work with him on several ideas. It is indeed rare to find a good friend and colleague in the same person and Chakra has filled both shoes with ease. I thank him for teaching me to "respect my data more" and to understand the nuances of linear algebra.

I am grateful to my advisors from my Master's program: Prof. I. V. Ramakrishnan at Stony Brook University and Dr. Swaminathan Subramanyam at Brookhaven National Laboratory. Both of them have truly inspired me to pursue a Ph.D. in Computational Biology and I am glad that they insisted I apply to the Ph.D. program at Carnegie Mellon. Without them, this work would not have seen fruition.

My colleagues at Oak Ridge: Dr. Ganesh Kamath deserves a special thanks for patiently reading this document (in his pretext of "I am learning something new"), providing me results from his forthcoming papers and computing a number of different distance profiles from various simulations. He has been my "go-to-discussion-buddy" whenever I need something. I would also like to thank Dr. Jose Borroguero, a close collaborator and friend, who has patiently read several drafts and suggested valuable comments. A number of scientists at Oak Ridge, some of whom had lunch with me, some of whom entertained me and some of whom I met down the corridor - I thank them for their interactions and support.

Colleagues at CMU: Hetunandan Kamisetty, who introduced me to probabilistic graphical modeling of protein structures; Sumit Jha, who introduced me to probabalisitic model checking; Narges Sharif Razavian and Subhodeep Moitra, the latest lab rats who have playing around with some of my data – all of them deserve special thanks. I also thank Xin Gao for his assistance and insights into some of the papers that did not make it into the thesis.

Colleagues in my program: Aabid Shariff, Byoungkoo Lee, Ahmet Bakan, Luis Pedro Coelho, Jacob Joseph, Sabah Kadri, Grace Huang, Justin Hogg, Guy Zinman, Yevgeniya Monisova, Andrej Savol, John Arul Prakash, Cordelia Ziraldo, Shannon Quinn, Anindita Dutta, Lidio Meireles, Ross Curtis, Joshua Kangas, Armaghan Naik - I sincerely thank each of you for being there, hanging out and having fun! Andrej Savol has been an a special collaborator and friend. From sharing his photos, Bach collection and energy numbers, I could not have done this without your help - thank you!

Without friends, Pittsburgh or CMU would not have been memorable. Varsha, RK, Sangeetha, Ashwin, Nitin, Devika, Advay, Siddharth, Arvind, Hetu, Ashwini, John - each of you have enriched my experience here and I could not have made it here without you guys. I thank you all for putting up with me and tolerating my inevitability in getting to bed by 11 PM! Elvira Garcia Highley and Chris Highley have been patient listeners to my rants during the lows and Elvira has patiently read through many drafts of my thesis proposal, papers and the thesis itself. I thank both of them. A special thanks goes out to Aabid and Afreen. Aabid - for putting up with me as an apartment mate, teaching me to live life and sharing a friendship that has been more than wonderful. And Afreen for making all those delectable delights, hilarious medical school stories and entertaining all

of us to no end!

Laurine Patricia Haddock (Patsey) has been more than an able administrator at the Department of Chemistry; she has been a mom to me and talked me through several difficult moments of my Ph.D. I thank her sincerely for that. Ena Miceli, at the School of Biological Sciences, has been outright wonderful - from tracking paychecks to introducing me to Italian food in and around Pittsburgh - thank you very much for your support. Thom Gulish at the Lane Center for Computational Biology has gone out of his way to help me: dropping me off at home on his scooter (or car), introducing me to Phish and allowing me to hang out with his awesome kids (Lilly, Joseph and Lucy). I thank him for being a good friend, colleague and an administrator.

Finally, and most importantly, I would like to thank my family. *Amma*, for instilling a sense of discipline, work ethic and responsibility; for being there at highs and lows; for all her sacrifices that would enable me to achieve my dreams - this would not have been possible without you. *Appa*, in spite of his difficult schedules made sure he visited me at least once a year. He has always given me 100% and were it not for his support, I would have not done my Ph.D. *Paatti*, for telling me all those stories when I was young and inspiring me to do what I wanted to do - I cannot thank you enough for that! My uncles and aunts have doted since I was a kid and I am thankful for their affection and attention. Ravi Chitappa deserves a very special thanks: he has been constantly in the background supporting me through out school. Kannan, Geetha and Karan have made me feel at home; I thank them for that. My cousins, Anand and Priya, Lakshmi and Rajesh, have hosted me on more than one occasion and I sincerely thank them for their love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Enzymes are a naturally occurring class of proteins that achieve their designated biochemical function precisely and efficiently. Enzymes have intrigued biochemists/ biophysicists since they are highly efficient biocatalysts with the ability to accelerate their target reaction (anywhere between 10^3 and 10^{17} times) [203]. Their functioning involves novel mechanisms of control including allosteric and cooperative effects. Thus, a detailed understanding at the molecular level of how enzymes function is critical for engineering new and efficient enzymes (for industrial purposes) as well as designing novel therapeutic agents that can treat a number of protein related diseases.

The role of enzyme's three-dimensional shape (or *fold*) has been well understood in the context of its function; the lock-and-key [89], induced-fit [161] and transition state stabilization [275] theories have suggested that enzyme function largely depends on the direct structural interactions between the enzyme and its substrate. However, several experiments and theoretical/ computational studies also indicate that residues far-away from the active site of the enzyme contribute critically to the function of the enzyme [152, 115]. Particularly, some of these distal residues are conserved throughout evolution and their role in the context of enzyme fold and biochemical function is not fully understood.

Most proteins are linear polymers of amino acids; the constituent amino acids allow the protein to fold into specific three-dimensional conformations via specific network of interactions formed either by hydrogen bonding or by hydrophobic interactions. These networks, while stabilizing the overall architecture of a protein, also allow it to undergo dynamic fluctuations under physiological conditions. These dynamic fluctuations enable the protein to sample a complex energy landscape consisting of multiple minima that are kinetically accessible. A recently emerging paradigm in understanding the structure-function relationship in proteins is that these intrinsic dynamic fluctuations are largely responsible

for achieving the specific biochemical function that the protein was designed for [24, 124]. Particularly, in recent years, coupled networks of protein motions have been found to exist in several enzymes conserved across different species [8, 9, 6]. Solvent motions have been shown to dominate protein motions [5, 84, 85, 91]. However, the exact nature of how localized structural changes and energy fluctuations influence large-scale conformational fluctuations within a protein remains a significant challenge in understanding the *structure-dynamics-function* paradigm.

In this thesis, we investigate the relationship between a protein's three-dimensional shape (or *fold*), conformational fluctuations that are intrinsic to the fold (or *dynamics*) and its *biochemical function*. For investigating biochemical function, we have chosen to study enzyme catalysis. In the context of enzyme catalysis, we address the following questions:

- How can one identify and characterize intrinsic dynamics of a protein, incorporating spatial and temporal information?
- How is intrinsic dynamics in specific enzyme systems conserved through the course of evolution?

The first part of the thesis addresses the development of novel tools to identify and characterize the intrinsic flexibility of an enzyme in the context of its function. The second part of the thesis focuses on understanding how enzymes have evolved - specifically, if intrinsic flexibility of an enzyme implicated in catalysis is also conserved as part of the fold. The ability to answer these questions in enzyme catalysis will be applicable not only to the design of new enzymes or drugs that enhance or inhibit enzyme function, but also to further our understanding of how proteins, in general, have optimized their fold and flexibility to achieve their targeted function.

1.1 Protein Dynamics

Under physiologically relevant conditions, proteins are not static structures but constantly undergo conformational fluctuations. These fluctuations are a result of a constant interplay with the surrounding environment (solvent, ions, other proteins, etc.) [84, 85, 91]. These fluctuations involve several different spatial and temporal scales. As illustrated in Fig. 1.1, the fastest time-scale fluctuations involve individual bond-vibrations which are of the order of femtoseconds (10^{-15}s). At pico-second (10^{-12}s) and nano-second time-scales (10^{-9}s), one can observe motions along individual side-chains or a small groups of side-chains involving *localized* regions in the protein. Beyond nano-seonds and micro-second (10^{-6}s)

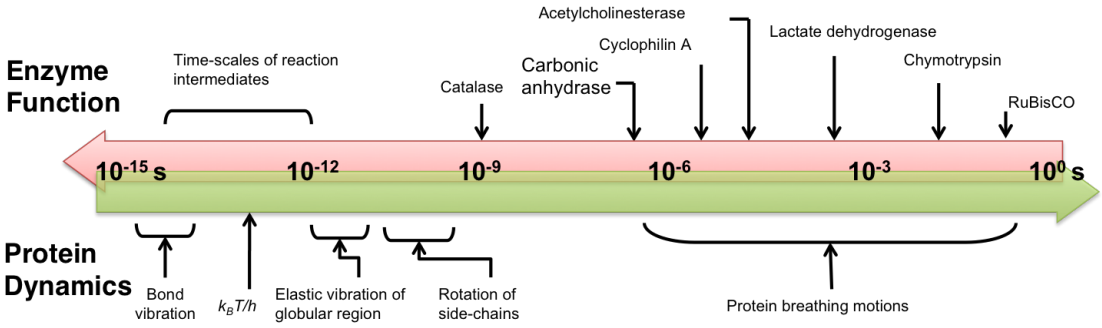

Figure 1.1: **Multi-scale nature of enzyme function and internal protein motions**. An illustration of how enzyme function and flexibility at multiple time-scales and spatial resolution is shown here. Some commonly known enzymes with known turnover numbers are shown along the red arrow. Protein motions ranging from bond vibrations to breathing motions overlap with enzyme function, leading to the question if enzyme flexibility and function are inter-related.

time-scales, motions within a protein can span several regions including whole secondary structures such as α-helices, β-sheets and flexible regions. These motions involving several regions of the protein are commonly referred to as *collective conformational fluctuations*, or breathing motions, and have attracted considerable attention in the literature due to their likely role in protein function [52].

Enzyme catalysis is a biochemical process involving chemical changes (e.g., breakage or formation of a bond) at the active site of the protein. However, as we observe in Fig. 1.1, the range of time-scales involved in substrate turn-over step of catalyzed reactions and the internal protein motions are very similar. This similarity in time-scales leads to the natural question: *are internal protein motions at various temporal and spatial scales and enzyme turn-over steps interconnected?* The answer to this question requires us to first build an integrated view of internal motions at various levels of spatial and temporal resolution in relation to the catalytic step of an enzyme.

Only recently, such an integrated view of protein structure, dynamics and function is emerging. A plethora of experimental techniques continue to enhance our understanding of the interconnection between protein motions and function. Neutron scattering has been used to monitor fast thermal motions (picosecond to 100 ps) [39, 42, 293]; nuclear magnetic resonance (NMR) experiments have provided information in the intermediate range (nanoseconds and longer) [125, 49, 269]; spin-echo neutron scattering on the microsecond to millisecond range [38]; hydrogen-deuterium exchange has been used to measure

slower conformational changes occurring in proteins (milliseconds) [294]; and crystallography [121] has revealed conformational changes occurring during the course of a protein's functional cycle. Moreover, recent Mssbauer effect and neutron scattering experiments have indicated that bulk solvent and hydration-shell fluctuations control protein motions and function [71, 84, 85]. More recently, ambient temperature X-ray crystallography and electron density sampling have revealed in exquisite detail the nature of conformational sub-states that may eventually affect protein function [90].

Experimental investigations of protein motions have faced an inherent problem; as mentioned above, the internal motions of proteins occur at a wide range of time scales, but the range of observed motions strongly depends on the relatively narrow time scale resolution window of the experimental method used [6, 124]. Further, apart from X-ray crystallography and NMR, it is relatively hard to relate back the observations from other experimental techniques in the context of individual protein motions. Thus, interpreting protein dynamics in the context of individual atomic motions is challenging.

Theoretical and computational studies are bridging this vital gap by linking motions at multiple time-scales to protein function such as ligand binding, enzyme catalysis and biological signaling. Bio-molecules, and proteins in particular have been challenging in order to characterize their dynamics at various time-scales. Substantial progress has been made in understanding protein motions since the first simulations of bio-molecules were carried out in 1977 [193]. Here, we briefly review some of the significant developments in the theoretical and computational literature with regards to the study of protein dynamics. Two approaches have been commonly used to investigate protein dynamics: (a) Normal Mode Analysis (NMA) based methods and (b) Molecular dynamics (MD)/ Monte-Carlo (MC) based sampling techniques.

1.1.1 Normal Mode Analysis (NMA): Analysis of Protein Motions using the Harmonic Approximation

Simple analytical models such as normal mode analysis (NMA) have contributed significantly to our understanding of collective conformational fluctuations involved in ligand/ substrate binding. Pioneered by Brooks and Karplus [51] and followed up by several others [103, 201, 179], NMA based approaches offer an intuitive and simplified representation of a protein's conformational landscape. An implicit assumption of NMA based approaches is that the effective potential of interaction between two atoms can be expressed as a sum of harmonic potentials between all pairs of atoms. The potential energy V for a given conformation, $\vec{r}^0$, where $\vec{r}$ represents the Cartesian coordinates of the protein, can

be expanded as a Taylor series in $\vec{r}$, as follows:

$$V(\vec{r}) = V(\vec{r^0}) + \sum_i \Big[\frac{\partial V}{\partial r_i}\Big]^0 (r_i - r_i^0) + \frac{1}{2}\sum_{ij} \Big[\frac{\partial^2 V}{\partial r_i \partial r_j}\Big]^0 (r_i - r_i^0)(r_j - r_j^0) + \dots \quad (1.1)$$

where $V(\vec{r^0})$ is some constant (at equilibrium), and the first derivative (second term in the summation) is zero at equilibrium. Hence, the potential can be set to the sum of pairwise potentials, given by:

$$\begin{aligned} V(\vec{r}) &= \frac{1}{2}\sum_{ij} \Big[\frac{\partial^2 V}{\partial r_i \partial r_j}\Big]^0 (r_i - r_i^0)(r_j - r_j^0) \\ &= \frac{1}{2}\sum_{ij} (r_i - r_i^0) H_{ij} (r_j - r_j^0) \\ &= \frac{1}{2}\Delta\vec{r}^T \mathbf{H} \Delta\vec{r} \end{aligned} \quad (1.2)$$

The Hessian, **H**, is positive semi-definite if it is built from a conformation at equilibrium, and hence can be diagonalized to obtain a set of eigenvalues (λ) and eigvenvectors (**U**). The Hessian has a total of 6 zero eigenvalues representing the 6 trivial rotational and translational motions (as a rigid body). While the eigenvalues, λ_i, represent the relative frequency of motions (small values of λ_i represent slow time-scale motions) where as U_i represents the directions of these motions.

Depending on how the Hessian **H** is built, several flavors of NMA techniques exist. The straightforward approach is to use an all-atom representation of the protein and build a Hessian, assuming that the conformation is at some stable energy minimum. In this thesis, all-atom NMA has been used. Several programs are available to do NMA. In this thesis, NMA has been performed using nmode [204]. The details of specific settings in NMA are explained in the methods section of the relevant chapters.

Another powerful approach is to coarse-gain the representation of the protein. This forms a class of NMA techniques commonly referred to as the elastic network model (ENM). Inspired by the works of Tirion [259] and extended to describe motions in proteins by Bahar and co-workers [111, 23] as well as others [76, 254], this coarse-grained approach relies on the ability to simplify the representation of the protein's potential energy using a network of masses connected with springs. This approach, is commonly called as the Guassian Network model (GNM) [111] where each amino-acid residue in the protein is 'abstracted' into a node in a graph and interactions between residues are replaced by springs (or edges) with a single uniform spring constant. Thus, GNM's potential V_{GNM}

would be given as:

$$V_{GNM} = \frac{\gamma}{2}\Big[\sum_{ij}^{N} \Gamma_{ij}\Delta\vec{r}_{ij}^{2}\Big], \tag{1.3}$$

where $\vec{r}_{ij}$, is the distance vector between residues i and j, and Γ_{ij} is a simple encoding of the connectivity of the network, given by the Graph Laplacian (also called the Kirchoff matrix):

$$\Gamma_{ij} = \begin{cases} -1 & \text{, if } i \neq j \text{ and } r_{ij} \leq r_c \\ 0 & \text{, if } i \neq j \text{ and } r_{ij} > r_c \\ -\sum_{j,j\neq i}\Gamma_{ij} & \text{, if } i = j. \end{cases} \tag{1.4}$$

Here r_c is referred to as the cut-off distance. For a protein with N residues, Γ will be a $N \times N$ matrix, representing the connectivity between residues. The eigenvectors $\vec{u}_{GNM}$ and eigenvalues Λ_{GNM} of Γ, represent the collective modes of the network, representing the native state dynamics of the protein. Note that the fluctuations from GNM are isotropic in all three directions and Gaussian. The statistical mechanics and the description of the methods are provided in detail in [28]. A subtle, yet important modification [25] to Eq. 1.3 results in the more popular Anisotropic Network Model (ANM), which uses a $3N \times 3N$ Hessian matrix to describe the potential ($3N$ to describe anisotropic Gaussian fluctuations in x, y and z directions for N residues in the protein). ANM's potential allows one to track changes in the direction of the inter-residue vector ($\Delta\vec{r}_{ij} = \vec{r}_i - \vec{r}_j$), in addition to the changes in just inter-residue distances as tracked by GNM.

Even with such simplifying assumptions, ENM based approaches have provided significant insights into the nature of dynamics and how it correlates with function [255]. Beginning with the study of human-immuno deficiency (HIV)- type 1 reverse transcriptase (RT) [22, 26], tremendous progress has been made in elucidating the nature of large-scale conformational fluctuations in the context of function. ANM (and ENM in general) is highly scalable, as illustrated by its applications to GRoEL-GRoES [60], ribosome complex [271], viruses [289] and recently, to the entire nuclear pore complex [181]. Further these approaches lend themselves easily to investigate hierarchical motions [62, 59] and to study signal processing mechanisms in proteins [61].

1.1.2 Molecular Dynamics (MD): Representing Anharmonic Motions in Proteins

In spite of their intuitive appeal, NMA based approaches do have some limitations [185]. The harmonic assumption is valid in and around the vicinity of a local minimum in the

complex energy landscape [119]. Further, as the protein is displaced from the local minimum, the harmonic assumption fails and may lead to violations in the observed fluctuations (and conformations) [247, 288].

Conformational fluctuations in a protein are governed by a rugged potential energy surface spanned by the protein [92]. Thus, one can imagine the energy landscape to be formed of "hillls" and "valleys" populated by discrete conformations of the protein. Within each valley, one would find the population to share significant similarity in terms of their conformations as well as internal energies. Thus, a set of protein conformations that share similar structure and energetic properties is usually referred to as a *conformational substate*. A protein may not indefinitely stay in one sub-state, but constantly keeps switching between different sub-states, in response to an external perturbation, such as ligand/ substrate binding or solvent motions and thermal fluctuations.

The rugged potential energy surface causes the protein to "jump" between various conformational sub-states giving rise to motions that are *anharmonic*. In the context of enzyme catalysis (as well as other functional processes such as ligand/ substrate binding), anharmonic motions play a critical role [94]. Particularly, it has been shown using NMR, in the case of dihydrofolate reductase (DHFR) [47] that the enzyme cycles through a set of discrete conformational sub-states separated by higher-energy sub-states during the course of its catalytic cycle. Similar observations have also been made for enzyme systems such as cyclophilin A (CypA) [77, 49], ribonuclease A (RNase A) [65, 163] and many other enzyme systems [125]. Thus, in order to understand the nature of protein flexibility and its impact on catalysis, it is necessary to characterize the dynamic energy landscape of the enzyme with respect to both anharmonic motions and the various conformational substates accessible via these motions [90].

Molecular dynamics (MD) and Monte-Carlo (MC) simulations [96], in the past thirty years, have played a significant role in modeling the anharmonic conformational landscape spanned by a protein. MD simulations also provide a rich description of the protein's conformational landscape at various levels of spatial resolution (including all-atom, back-bone, side-chain or just C^{α} atoms). They can also take into account solvent and environment conditions that mimic physiologically relevant conditions. By solving the Newton's laws of motion at every time-step, MD simulations can potentially sample the vast conformational landscape of a protein and account for its rugged potential energy surface, which can then be analyzed to gain insights into protein motions.

In MD, a atomic interactions in a protein (or any molecule) is described via a molecular mechanics (MM) force field [172]. Force fields usually consist of a predefined set of atom-type definitions/ descriptions based on specific chemical environment (see Fig. 1.2. For example, the behavior of a carbon in methane and carbon in an aromatic compound like

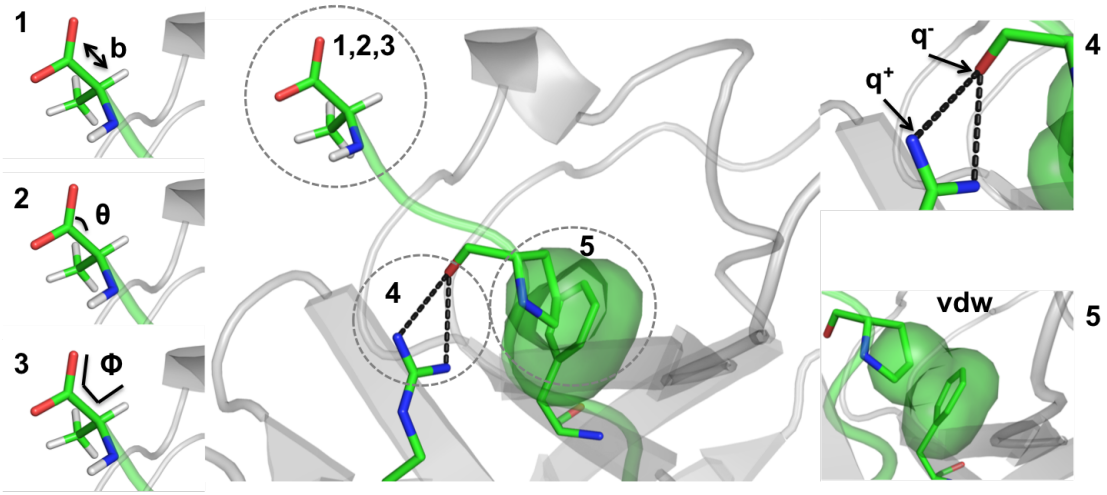

Figure 1.2: **Illustration of Molecular Modeling**. A protein (shown in light gray cartoon representation) with its constituent atoms are connected via bonds. The first three terms in a typical molecular mechanics force field are shown on the left. b represents bond lengths, θ represents bond angles, ϕ refers to the torsion angle. The right hand side of the image represents the last two interaction terms in Eq. 1.5: electrostatic and van der Waals interactions. Note this is only a schematic representation.

benzene is different, and a force-field helps describe this change succinctly. Force fields are derived from rigorous first principle techniques such as ab-initio quantum mechanical calculations and are then fit to experimentally known thermodynamic data to describe the atom of interest and thus called empirical force fields . The potential energy of a molecule $V_{MM}(\vec{r})$, based on a given conformation $\vec{r}$, under a typical force field is a summation of two groups of interaction terms: (a) bonded (V_b) and (b) non-bonded (V_{nb}). As described in Eq. 1.5,

$$V_{MM}(\vec{r}) = \overbrace{\sum_{i,j} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} K_\phi[1-cos(n\phi)]}^{V_b} + \underbrace{\sum_{non-bonded}\left[\frac{A_{ik}}{r_{ik}^{12}} + \frac{C_{ik}}{r_{ik}^{6}}\right] + \sum_{non-bonded}\frac{q_i q_j}{D r_{ik}}}_{V_{nb}} \quad (1.5)$$

The first three terms are the bonded interactions in the protein (see Fig.1.2). $V_{bond-stretch} = \sum_{i,j} K_b(b-b_0)^2$ describes the harmonic potential between atomic pairs connected by a covalent bond. The energy of a bond is approximated as a function of a bond's displacement from its ideal bond length, b_0, and the force-constant K_b determines the relative strength of the bond. The second term, $V_{bond-angle} = \sum_{angles} K_\theta(\theta-\theta_0)^2$, describes the alteration of bond angles from an idealized geometry θ_0. $V_{bond-stretch}$ and $V_{bond-angle}$ describe deviations from the ideal geometry and hence describe penalties for violating these constraints. The third term, $V_{dihedrals}$, models the steric barriers between 4 atoms, usually separated by 3 covalent bonds. The motion associated with this term is a rotation along the dihedral angle around the middle bond. This term is also harmonic, but conveniently coded as a periodic cosine function. Note, the n in the equation represents the periodicity of the dihedral (for example, in a C-C sp^3 hybridization, $n = 3$).

The last two terms are collectively referred to as non-bonded interaction energy (Fig. 1.2). The fourth term in Eq. 1.5, is referred to as the Lennard-Jones potential, V_{LJ} and represents the van der Waals interactions. It can be either attractive or repulsive; repulsion mainly occurring because of strong electron-electron interaction between two atoms and attraction occurring due to dispersion effects in the electron clouds. It is common to use a 6-12 potential to describe effects arising out of van der Waals interactions. The terms A and C are dependent on the atom type and can be estimated via gas-phase scattering experiments. r_{ik} represents the distance between the atoms. Of all the sources of interactions, van der Waals interactions perhaps represent the most important in terms of stabilizing the overall conformations of biological macro-molecules [203, 172]. The fifth term in Eq. 1.5

represents the Coulombic interaction, which arises due to charged interactions in the protein; q_i represents the charge of the atom/ particle and D represents the effective dielectric function of the medium in which the atoms are immersed.

These empirical potentials do suffer from inaccuracies [145] and hence, force field development is an active area of investigation. All empirical force fields describe electrostatic interactions as point-charges, which may not be appropriate for certain pH fluctuations. Recent interest in the area of developing polarizable force fields has allowed some progress to be made in overcoming the limitations of such inaccuracies [110].

The potential energy function defined in Eq. 1.5 is differentiable with respect to atomic coordinates, and hence one can deduce the force acting on an atom (both direction and magnitude) and this can be used to integrate the Newton's laws of motion. Several integration schemes for MD exist; most popular methods include (a) Verlet algorithm, (b) Leap-frog method, (c) velocity Verlet and (d) the Beeman technique [172, 96]. Each method comes with advantages and disadvantages and the implementations are entirely controlled by the choice of MD simulation software being used.

The output from an MD simulation (or a collection of simulations) is usually a set of conformations, referred to as an *ensemble*. Depending on the time resolution of MD and how often individual conformations were stored, these ensembles can be quite large. For example, simulating a protein for 10 nanoseconds (ns) storing conformations every picosecond can result in a total of 10,000 conformations. The time-scale resolution, similar to the spatial-resolution is up to the end-user. In each of the simulations used in this thesis, we mention the time-scale resolution used explicitly in the methods section. This thesis primarily makes use of the AMBER (Assisted Model Building with Energy Refinement) [56] software suite for molecular modeling purposes. It uses the parm98 [218] force-field, which has been verified to be better suited for describing protein dynamics [9].

MD based approaches have formed the basis of investigating anharmonic fluctuations in proteins. MD is routinely used in several target application contexts [151]. MD offers the flexibility of being easily integrated with a number of finer-grained approaches (such as quantum dynamics) or with coarse-grained approaches (such as ANM/ ENM [137]) and thus, provides a unique framework to gain insights into the mechanistic underpinnings of a variety of proteins. MD is also unique since it allows the ability to mimic real cellular conditions within a simulation [194] and hence, provides a unique computational microscope into the workings of a protein [173].

However, atomistic MD simulations do suffer from drawbacks. It is quite well known that MD simulations cannot often sample motions at biologically relevant time-scales [64, 87]. This is primarily due to the all-atom resolution at which MD functions. Several im-

provements in the sampling problem of MD have now enabled us to overcome some of the serious limitations. In the context of software development, the concerted efforts of several groups including AMBER [56], GROMACS [127], NAMD [220] and Desmond [50] have provided us with the ability to scale MD systems to hundreds of thousands (if not millions) atoms routinely. Further more, customized hardware development using field programmable gate arrays (FPGAs) [11] and application-specific integrated circuits (such as Anton [242]) have accelerated the speeds at which MD can be carried out. There is also growing interest in developing MD packages on graphics processing units (GPUs), as evidenced by several studies [82, 249]. Apart from these, from the algorithmic side, there has been tremendous progress in developing improved sampling techniques [175], coarse-grained MD simulations [139] as well as continuum models to increase the accessible times for MD [159].

1.1.3 Analysis of Collective Conformational Fluctuations from MD

Once a simulation has been completed, it is then necessary to analyze the ensemble[1] to gather insights on the actual behavior of the protein. Structural metrics such as root-mean squared deviations (RMSD[2] [143]) are commonly used to assess the stability and quality of simulations. However, to gather mechanistic insights about internal protein dynamics and collective conformational fluctuations, it becomes necessary to use advanced statistical techniques. Collective conformational fluctuations have traditionally been analyzed using principal component analysis (PCA) [142]. Various techniques [118, 155, 25] are available of which three techniques are most popular: (a) quasi-harmonic analysis (QHA) [150] (b) essential dynamics (ED) [15] and (c) dihedral PCA (dPCA) [14].

QHA [150, 149] was initially proposed to estimate the configurational entropy of macromolecules. QHA approximates the Boltzmann partition function arising from the anharmonic conformational landscape with multi-variate Gaussian distributions. It captures the large-scale conformational fluctuations from an ensemble of protein conformations by diagonalizing the mass-weighted covariance matrix known as the atomic fluctua-

[1] Ensembles are also commonly referred to as MD trajectories. In this thesis, we will use both 'ensemble' and 'trajectory' interchangeably.

[2] RMSD is a measure of the average distance between two superimposed structures. Given two conformations represented by $\vec{r}_i$ and $\vec{r}_j$, RMSD is defined as:

$$RMSD = \sqrt{\frac{1}{n}\sum_i \|\vec{r}_i - \vec{r}_j\|^2},$$

where n is the total number of atoms on which alignment is carried out.

tion matrix (F_{ij}). For a system with N atoms, F_{ij} is defined as follows:

$$F_{ij} = \langle m_i^{1/2}(\vec{r}_i(t) - \langle \vec{r}_i(t) \rangle) m_j^{1/2}(\vec{r}_j(t) - \langle \vec{r}_j(t) \rangle) \rangle \tag{1.6}$$

where i and j represent the $3N$ degrees of freedom in Cartesian space; m is the mass of an atom and the quantity within $\langle \dot{} \rangle$ denotes an average over the ensemble. The inverse square roots of the eigenvalues determined by diagonalizing F_{ij} represent the frequencies associated with protein eigenmodes. The eigenvectors represent the displacement vectors of the individual atoms along particular eigenmode components. The eigenvalues are typically arranged in ascending order; the first six frequencies are zero, corresponding to the global translation/ rotation motions. The lowest frequencies (from 7 onwards) correspond to large-scale cooperative motions in the protein and the higher frequencies represent localized motions. For a system with N atoms, there are $3N - 6$ internal modes; however, typically only a limited number of slow modes (lowest frequencies) are computed since they are considered to be functionally relevant [295]. QHA allows identification of protein motions at a variety of time scales, as the atomic fluctuation matrix can be computed from protein conformations sampled during a single MD simulation (short time scale) [57], a collection of MD trajectories (collectively representing long time scale) [9] or a set of conformations obtained using various sampling techniques (which could represent protein present in different stages during its functional cycle).

ED [15, 1] was introduced to develop an approach to separate the configurational space into two subspaces: (a) essential subspace consisting of only a few degrees of freedom that capture anharmonic motions and contributes to the maximal variance in positional fluctuations and (b) a constrained subspace which has Gaussian-like behavior and is localized to specific regions of the protein. It differs from QHA in that it uses only the atomic displacements instead of the mass-weighting these displacements. Thus, given an ensemble of conformations ($\vec{r}(t)$) from a MD simulation, the covariance matrix for ED is defined as follows:

$$C_{ij} = \langle (\vec{r}_i(t) - \langle \vec{r}_i(t) \rangle)(\vec{r}_j(t) - \langle \vec{r}_j(t) \rangle) \rangle \tag{1.7}$$

where i and j still represent the $3N$ degrees of freedom in the Cartesian space and the quantity within $\langle \dot{} \rangle$ denotes an average over the ensemble. The eigenvalues of C determine the amplitude of the fluctuations; hence, a higher amplitude (and larger eigenvalues) represents more fluctuations and lower eigenvalues determine constrained motions. The eigenvectors represent the directionality of motion along the ED modes. ED has been used to describe large-scale protein motions [263], identify hinge sites [264], study protein folding [70] and describe motions from ensembles of experimentally determined structures [170].

Both QHA and ED require that the ensemble of conformations be aligned to some reference structure before any of the analysis is performed. This approach of aligning can be particularly cumbersome, if a well defined structure is not available. For example, in protein folding studies, where the protein adopts multiple conformations before reaching its native state conformation, the description of a good reference structure can be very difficult to find. To overcome this limitation, it was proposed that one could perform PCA based analysis in the dihedral space [14]. This approach, commonly referred to as dihedral PCA (dPCA) uses the internal dihedral angles ($\phi - \psi$) from the peptide backbone as the variables on which PCA is performed. The idea here is to convert each $\phi - \psi$ into an Euclidean representation, where each ϕ_i and ψ_i for residue i is converted into a 4 tuple defined as follows:

$$\begin{aligned} x_{i-3} &= cos(\phi_i); \\ x_{i-2} &= sin(\phi_i); \\ x_{i-1} &= cos(\psi_i); \\ x_i &= sin(\psi_i). \end{aligned} \tag{1.8}$$

This transformation results in a $4M \times 4M$ covariance matrix as defined in Eq. 1.7, where M is the total number of $\phi - \psi$ angles in the protein. This approach has also been widely used to study the importance of collective motions in protein folding as well as describe free-energy landscapes of several proteins [186, 199].

1.2 Evolutionary Linkage between Enzyme Fold, Flexibility and Catalysis

In the first part of the thesis, we examine how we can characterize the conformational landscape of an enzyme and its intrinsic flexibility in the context of catalysis. In the second part of the thesis, we examine if internal motions in enzymes are evolutionarily conserved as part of their catalytic step. Enzymes have evoked considerable interest in biochemistry for their relevance in industrial processes as well as design of novel therapeutic agents that can overcome protein-related diseases. Hence, in order to design better enzymes as well as novel drugs, a good understanding of how enzymes have evolved with respect to their flexibility and catalysis is crucial. In this section, we examine the latest developments in our understanding of how enzymes have evolved.

Reaction	Enzyme class	E.C.	Example
Oxidation/ Reduction	oxidoreductases	1	dihydrofolate reductase
Group transfer	transferases	2	MAP2-kinase
Hydrolysis of bond	hydrolases	3	Serine protease
Formation/removal of double bonds	lyases	4	Carbonic anhydrase
Isomerization of functional groups	isomerases	5	Cyclophilin A
Single bond formation	ligases	6	Aminoacyl tRNA synthetase

Table 1.1: **Summary of enzyme catalyzed reactions**. Each unique reaction mechanism receives a E.C. number, which can be then used to identify the enzyme.

1.2.1 Sequence Analysis of Mechanistically Diverse Enzymes

Six basic types of enzyme catalysis are known, as summarized in Table 1.1. Each enzyme is classified by the Enzyme Commission (E.C.), based on the specific chemical reaction that the enzyme catalyzes. We primarily focused on three enzyme classes: (a) oxidoreductases and (b) isomerases and (c) hydrolases. The specific enzymes and the enzyme mechanisms will be covered in subsequent chapters (see Chapters 5 and 6). These three enzyme classes constitute nearly 60% of naturally observed enzymes [29] and are most biologically relevant for developing therapeutic agents against a variety of diseases.

In spite of having a restricted number of protein folds available, enzymes have developed features over the course of evolution that make them especially well suited to catalyze their target chemistry as well as a diverse set of biochemical reactions [12]. This remarkable diversity of naturally occurring enzymes begs the question whether each protein fold has evolved independently [202]. Sequence analysis of enzyme super-families have identified structural and biochemical features [102] that have been conserved as part of their enzymatic mechanism. Further, Meng and co-workers have also proposed the idea of active-site templates in several protein super-families [196]. Babbitt and coworkers have reported that mechanistically diverse enzymes that share a common sub-step have very similar protein folds [19]. Based on the exploration of a family of proteins consisting of a sub-set of two dinucleotide binding domains flavin-proteins, Ojha and co-workers have identified features of an enzyme's fold that can allow for diversification of biochemical function [209]. It was also observed that the conserved structural features are optimal for stereo-specic hydride transfer that is stabilized by specic interactions with amino acids from several motifs distributed among both di-nucleotide binding domains [209]. Similar observations have been reported in the enolase, amidohydrolase/phospotriesterase and other families [102, 101].

1.2.2 Dynamical contributions to Enzyme Catalysis

Structural interactions and fold similarity provide valuable insights into how active site geometry and various structural motifs/ templates may be utilized in evolution to "diversify" chemical reactions. However, it does not provide any rationale towards why mutations to distal residues in the enzyme can play a significant role in affecting catalytic function. This requires an understanding of the various features that impact the progress of the reaction. Two competing theories have emerged in terms of the features that control the catalytic step of the enzyme: (a) dynamical properties or motions of the enzyme/ substrate at various time-scales that impact the catalytic step of the enzyme (or its reaction coordinate) [114, 47, 49, 77, 78, 163, 90] and (b) electrostatic pre-organization, where by the protein active site provides the necessary charged environment to lower the barrier required to move from the reactant (R) state to the transition (T) state [275]. While there is substantial debate in the literature as to which of these factors contribute the most to the catalytic step [158], we believe that it is important to understand the dynamical contributions (motions in the enzyme and substrate) that lead to catalysis.

Dynamical contributions in the context of enzyme substrate binding have been extensively investigated. In the context of enzymes, ANM (and its variants) have provided insights into how ligands may bind at target locations in the protein. Yang and Bahar, were able to isolate the common mechano-chemical characteristics in a large data-base of enzymes and found that catalytically important residues were co-localized with global hinge centers in these enzymes [287]. Further, Bakan and Bahar were able to show that the intrinsic fluctuations of an enzyme plays a dominant role in structural changes upon ligand binding [30]. These studies were limited by the coarse-grained representation (at C^{α} level) and modeling the protein's conformational landscape using a purely harmonic model. Moreover, enzyme catalysis is a biochemical process involving the breakage and formation of a chemical bond, which cannot be captured by ANM (or any harmonic model) since the processes are essentially away from equilibrium [185].

ENM has also been useful in comparing/ contrasting the dynamics of ligand-binding in enzymes from several proteins. For example, Keskin and co-workers were able to study the common features of ligand-binding dynamics in several proteins that shared structural homology [153]. Similarly, Carnevale and co-workers compared large-scale conformational fluctuations associated with ligand-binding in the entire protease super-family and were able to show significant similarity in dynamics along the ligand-binding sites in proteases [55]. More recently, a comprehensive analysis of the binding dynamics of the RAS GTPase family of enzymes revealed an exquisite similarity of motions in the conserved core of the family and specific motions associated with respect to functional specializa-

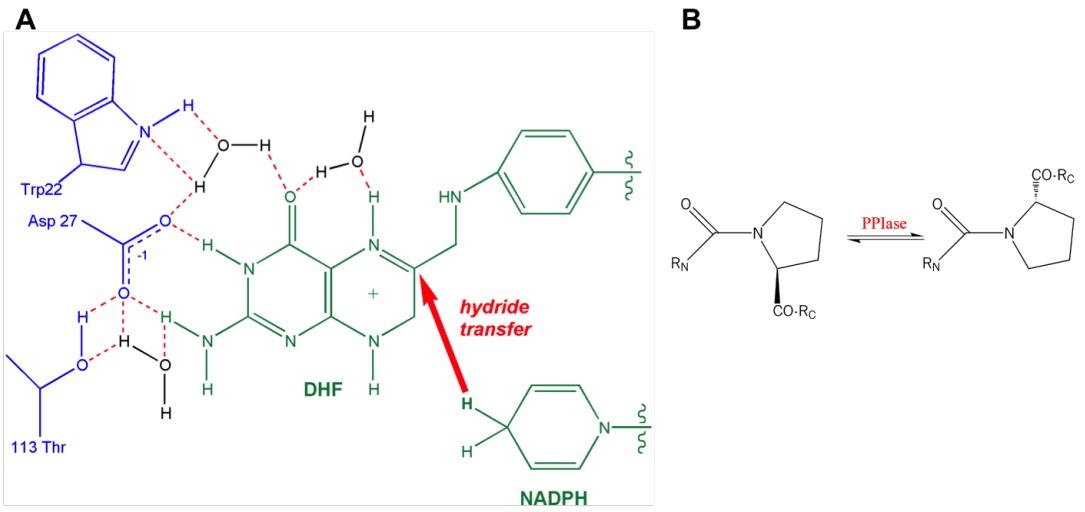

Figure 1.3: **Reaction Mechanisms studied in this dissertation**. (A) shows the reaction mechanism of hydride transfer in the enzyme DHFR. (B) shows the reaction mechanism of isomerization of prolyl-peptidyl substrate in the enzyme CypA.

tion [224].

Enzyme catalysis was traditionally studied using quantum mechanical (QM) or quantum mechanics/molecular mechanics (QM/MM) modeling. However, the ability to implement rigorous QM/MM approaches for various enzyme systems can prove to be extremely challenging. Hence, a more novel approach called the Empirical Valence Bond (EVB) [276] method was proposed. This approach allows one to map the reaction as a series of potential surfaces that drive the system from the reactant to the product state. Using MD simulations to sample along the potential surfaces and then using standard thermodynamical approaches allows one to generate a free-energy profile for the reaction being studied. EVB has been successfully applied to a number of enzyme systems and is summarized in [274, 210, 158].

Agarwal and co-workers were able to identify a "correlated network of coupled motions" coupled to catalysis in the enzyme dihydrofolate reductase (DHFR) [7, 8]. DHFR is an enzyme that catalyzes the reduction of 7,8-dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate. As illustrated in Fig. 1.3A, the hydride from the co-factor nicotinamide adenine dinucleotide phosphate (NADPH) is transferred to the C6 of the pterin ring in the substrate (DHF). The collective motions in DHFR, termed as *reaction coupled flexibility*, involve fluctuations that extend from the surface all the way to the active site of the enzyme. These

motions occur at multiple time-scales from μs and beyond (spanning the entire protein) to fast time-scale fluctuations at nanoseconds and picoseconds (in and around the active site). Extensive experimental and computational/ theoretical work [40, 47, 41], following this have now firmly established that coupled motions are indeed an essential feature required for DHFR to catalyze its reaction.

Similar networks of correlated motions were also observed in the enzyme cyclophilin A (CypA) [9, 4], which was later verified independently via experimental techniques including NMR [78, 90]. CypA is an enzyme that catalyzes the isomerization of prolyl-peptidyl bonds in a variety of substrates ranging from small peptides to large proteins. The isomerization reaction mechanism is illustrated in Fig. 1.3B. These studies and a number of others point to the importance of the enzyme's *intrinsic dynamics* and how these motions enable the enzyme to overcome the energy barrier(s) separating the reactant and product states. More recently, it was also proposed that solvent motions (especially close to the surface of an enzyme) can also play a substantial role in modulating an enzyme's catalytic step [6].

1.3 Specific Aims

A summary of techniques used to examine protein dynamics at multiple time-scales was described in Sec. 1.1. Note that all the current techniques described make an implicit assumption that protein motions can be described as being harmonic or approximate the anharmonic landscape spanned by a protein using multi-variate Gaussian distributions. This would naturally lead to the question whether one could do better in describing the protein's conformational landscape (and energetics). The specific questions investigated in the context of describing the internal protein motions at various time-scales (and spatial resolution) are as follows:

1. How do techniques such as NMA and PCA based on harmonic approximation of the landscape compare to experimentally determined large-scale (slow time-scale) fluctuations?

2. Given that NMA and PCA based approaches face limitations, how can one characterize the anharmonic nature of collective fluctuations?

3. Furthermore, with the increasing ability to sample protein motions at biologically relevant time-scales, are there possible means to understand collective conformational fluctuations as simulations are progressing?

We hypothesize that the internal dynamics in an enzyme at multiple time-scales involving a hierarchy of local and global conformational fluctuations, are in fact crucial for enzyme catalysis and may have induced selective pressure over the course of evolution. A critical test for this dynamical model of enzyme catalysis to be successful is to explore the connection between the chemical step during enzyme catalysis and protein exibility in the context of natural evolution of enzymes. In this part of the thesis, we lay the ground work for analyzing the evolutionary linkage between enzyme flexibility and catalysis in a systematic way. In order to study the relationship between an enzyme's flexibility and catalytic function we examine two important questions which are described below:

1. Is there significant similarity in the intrinsic dynamics of enzymes before (substrate-free), during (along the reaction pathway) and after the catalytic step?
2. Given the same bio-chemical reaction, even if two enzymes do not share a common fold/ shape, is the flexibility associated with the catalytic step similar?
3. If enzymes from a super-family share a common mechanistic step during catalysis, is the flexibility associated with the chemical step conserved across the super-family?

The answers to these questions will have significant impact in understanding protein dynamics (at various spatial and temporal scales) and also in enhancing our understanding of how enzymes function. Our studies have resulted in the publication of three published research articles and two articles that are currently being resubmitted for publication. These are listed here:

1. Arvind Ramanathan, Pratul K. Agarwal, *Computational Identification of Slow Conformational Fluctuations in Proteins*, J. Phys. Chem. B., 2009, 113 (52), pp 16669–16680.
2. Arvind Ramanathan, Andrej J. Savol, Christopher J. Langmead, Pratul K. Agarwal, and Chakra S. Chennubhotla, *Higher-Order Correlations in Internal Protein Motions and Energetics*, Phys. Rev. Lett., 2010, (in submission).
3. Arvind Ramanathan, Pratul K. Agarwal, Maria G. Kurnikova, Christopher J. Langmead, An online approach to mine collective behavior from molecular dynamics simulations, J. Comp. Bio. (2010), 17:3 (in press).
4. Arvind Ramanathan, Pratul K. Agarwal, *Evolutionary Linkage between Enzyme Fold, Flexibility and Catalysis*, .
5. Arvind Ramanathan, Pratul K. Agarwal, *Enzyme Superfamily shows Conservation of Flexibility linked to Catalysis*, in revision.

1.4 Outline of the Thesis

In chapter 2 of this thesis, we examine how well NMA and PCA based techniques (QHA) provide insights into long time-scale collective conformational fluctuations and whether conformational sub-states identified may be relevant in protein function. We show that while PCA can be useful to describe experimentally observed root-mean squared fluctuations (RMSF) and collective motions, it may not provide relevant information regarding the organization of conformational sub-states in the landscape. Particularly, we observe that the sub-states identified using PCA do not share enegetic/ conformational similarities, hence making it difficult to interpret these states in the context of biological function. This observation makes it difficult to characterize protein flexibility in the context of enzyme catalysis.

Given that PCA based techniques lack interpretability in terms of identifying conformational sub-states, the natural question is how do we learn about the complex anharmonic conformational landscape of a protein? This question is investigated in chapter 3 of the thesis. We perform a detailed analysis of the positional and internal energy distributions of proteins and characterize the essential statistical "features" of the conformational landscape. Particularly, we show that the anharmonic behavior in protein fluctuations gives rise to unique features which can then be exploited to build a representation of the conformational landscape. This model, which we term as quasi-anharmonic analysis (QAA), can identify and characterize unique conformational sub-states that share significant similarity in both energy/ conformational space. Further, we also show that the large-scale conformational fluctuations as described by QAA can be easily interpreted in terms of a protein's function.

The recent improvement in hardware/ software has made all-atom MD simulations to access biologically relevant time-scales (μs and beyond). The data from such large-scale MD simulations can easily reach several terabytes. This unprecedented access to data have also brought along difficulty in storing, analyzing and retrieving biologically relevant information to the end-users. Thus, there is a need for analyzing such large simulations on-the-fly or *online*. Although it is possible to use a few coarse-grained structural metrics to track MD simulations, these properties doe not necessarily provide any information regarding collective behavior, which is of interest to the end-users. Hence, in chapter 4 of this thesis, we introduce a novel online approach to mine MD simulations as they are progressing. By building a novel multi-dimensional representation of MD data using *tensors*, we show that it is possible to reason about collective motions and how these behaviors evolve typically in a long MD simulation. Unlike tracking RMSD, our approach allows us to cluster the protein into collectively moving regions based on similarity in

distance fluctuations as well as identify time-points during which the collective behaviors change significantly.

In chapter 5, we ask the question, if the large-scale collective conformational fluctuations in an enzyme is similar throughout its functional cycle. Particularly, we study three scenarios for the enzyme cyclophilin A (CypA), a ubiquitous enzyme catalyzing the cis/trans isomerization of peptidyl-prolyl bonds in several peptides and proteins. We ask if the slow conformational dynamics, intrinsic to the enzyme are similar before, during and after the catalytic sub-step. The study shows the remarkable similarity of intrinsic motions along the reaction pathway as well as in its substrate-free state implying that perhaps, enzymes have evolved precisely to move in directions that enable function.

In chapter 6 of this thesis, we focus on answering if dynamics coupled to the catalytic step of an enzyme (in spite of having different folds) show any similarity. By identifying the common features in the reaction coupled flexibility across multiple species in the enzymes that catalyze the aforementioned reaction mechanisms, we show that reaction coupled flexibility is conserved across the enzyme fold. We also show that the distal (and flexible) residues located on the surface of the enzyme are connected to the relatively rigid active site of the protein via non-covalent linkages (hydrogen bonds and hydrophobic interactions) that are conserved irrespective of sequence/ structural homology.

In chapter 7, we present evidence to show that enzymes from a super-family sharing a common mechanistic step during catalysis also show similar flexibility during the chemical step of the reaction. We examine the super-family of oxido-reductases, which utilize dinucleotide cofactors: nicotinamide adenine dinucleotide (NAD+), its phosphorylated and reduced forms (NADP+, NADH and NADPH). These cofactors play a central role in cellular metabolism and energy production, as hydride-accepting and hydride donating coenzymes, in many essential biochemical processes including glycolysis and the citric acid cycle. This enzyme super-family is referred to as the dinucleotide binding Rossmann fold proteins (DBRPs) due to the presence of the Rossmann fold, β-α-β motif that binds a nucleotide. The investigated enzymes have very low sequence identity with diverse secondary structural topologies. The reaction coupled exibility shows existence of reaction coupled motions that are remarkably similar. Detailed analyses indicate that in addition to structural constraints, motions that play a promoting role in catalysis are also a conserved feature of the entire super-family.

In chapter 8, we conclude with a perspective of how each of the techniques introduced in the first part of the thesis are useful. The outcomes of the second half of the study may have far reaching implications for our understanding of how proteins function as well as what evolutionary constraints may influence enzyme catalysis. A survey of possible applications are also discussed in this chapter.

Chapter 2

Analyzing Slow Protein Fluctuations using Quasi-harmonic Analysis

Slow conformational fluctuations occurring at the microsecond to millisecond time scale have recently attracted considerable interest in connection to the mechanism of enzyme catalysis. In this chapter, we present our studies on identification and characterization of microsecond flexibility of ubiquitin, based on quasi-harmonic analysis (QHA) and normal-mode analysis (NMA). First, a strategy to sample motions at long time-scales is presented, which effectively combines information from multiple starting structures and accounts for the natural structural diversity ubiquitin may exhibit. Next, we show that the identified slow (large-scale) fluctuations via this strategy are in good agreement with the fluctuations observed from available experimental ensembles. It is also demonstrated that the motions along the slowest modes have functional implications for ubiquitin binding. All-atom NMA, on the other hand, is found to strongly depend on the reference conformations. Analysis of the internal energy of conformations along the top-most (slow) modes from QHA reveal that the conformations do not share much homogeneity in the energy/ conformation space. A perspective on the advantages and limitations of using the harmonic approximation and its ability to characterize long time-scale fluctuations in proteins is presented in the end. [1]

[1]Parts of this chapter is adapted from: Arvind Ramanathan, Pratul K. Agarwal, *Computational Identification of Slow Conformational Fluctuations in Proteins*, J. Phys. Chem. B., 2009, 113 (52), pp 16669–16680.

2.1 Introduction

Slow, collective conformational fluctuations, also referred to as protein breathing motions [52] have attracted considerable interest due to their possible role in enabling a protein to perform its designated function such as enzyme catalysis [6, 52, 125, 113, 8, 114, 53, 77, 40, 5, 78]. As explained in the previous chapter (Sec. 1.1), experimental and computational techniques are playing a significant role in elucidating the nature of intrinsic protein dynamics and function [6, 114, 100].

Particularly, NMA [51, 25, 179, 282] and its various extensions [286] (see Sec. 1.1.1), have provided insights into the conformational fluctuations associated with individual protein structures, as well as flexibility intrinsically associated with the overall shape of proteins and its linkage to protein function. In NMA, a description of the internal motions of a protein conformation assumed to be in a local energy minimum (or its vicinity) is computed by diagonalization of the Hessian matrix [197]. Even though NMA provides information about fast and slow protein motions, these are representative of conformational fluctuations observed in the vicinity of the reference structure (Fig. 2.1) [185]. Moreover, it is widely discussed that the slow conformational changes in the proteins show a large degree of anharmonicity [212, 72, 71, 261]. The harmonic approximation limits the use of NMA to small amplitude motions in the potential energy surface associated with a local minimum [34] (corresponding to the gray region in Fig.2.1). Therefore, NMA is not well suited to study conformational changes associated with biochemical processes such as enzyme catalysis, which covers distant areas of the conformational energy landscape. Methods such as time-averaged normal coordinate analysis (TANCA) [205] have been developed to overcome some of the inherent limitations of NMA. TANCA provides more reproducible modes by diagonalizing the time-averaged Hessian matrix. The fast motions that differ considerably between the reference structures are removed in TANCA. However, TANCA is computationally very expensive, as it requires calculation of the Hessian for each structure of the molecular dynamics (MD) trajectory.

Quasi-harmonic analysis (QHA) [150] and related principal component analysis techniques [15] (see Sec. 1.1.2), based on the eigenvalue decomposition of an ensemble of protein conformations, have provided a useful method for identifying motions particularly at long time scales (see Fig. 2.1). QHA allows identification of protein motions at a variety of time scales, as the atomic fluctuation matrix can be computed from protein conformations sampled during a single MD simulation (short time scale) [57], a collection of MD trajectories (collectively representing long time scale) [9] or a set of conformations obtained using various sampling techniques (which could represent protein present in different stages during its functional cycle).

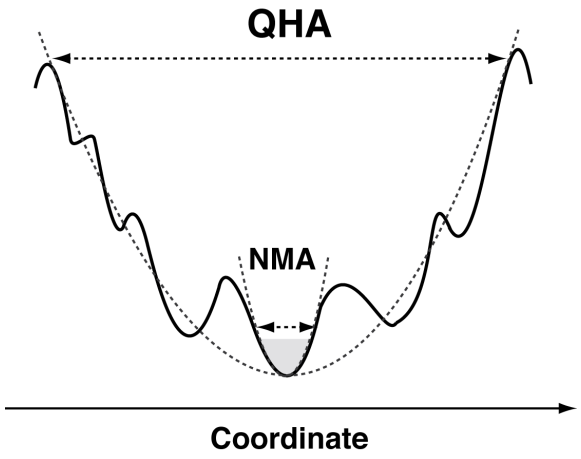

Figure 2.1: **Conformational fluctuations identified by QHA and NMA based on alternate ways to approximate the potential energy surface.** NMA identifies motions in the vicinity of the energy minimum, while QHA can identify motions that span distant areas of the energy surface.

QHA is emerging as an important method for identification of slow conformational fluctuations within proteins. However, the degree of certainty with which the motions identified by QHA at long time scales (microsecond and longer) can reproduce the motions observed through experimental methods such as NMR/X-ray crystallography [99, 170] remains an important concern. The amount of conformational sampling that is required to observe the slow motions corresponding to experimentally available data is also of concern [1, 68, 64]. Further, QHA does not provide any information about the time scales of motions – only the relative direction and the magnitude of motions are available from the analysis. The eigenvalues provide relative information about the frequencies associated with the eigenmodes; however, in a number of cases, the exact time scale of the computed motions cannot be assessed definitively. For example, if the motions are computed by combining conformations sampled by several MD simulations based on different starting structures, information about the time scale of the identified motions is missing [31, 169]. A related concern is that of identifying and characterizing conformational sub-states. QHA (and related PCA techniques) have been regularly used in identifying sub-states [192, 296, 297]. However, it is not clear if these sub-states share any conformational or energetic homogeneity. This has important bearing on understanding the organization of the conformational landscape of a protein and therefore needs to be clarified.

In this chapter, we present a systematic characterization of slow protein conformational fluctuations identified using QHA and NMA. Further, a methodology for identifying conformational fluctuations associated with an enzyme reaction is also discussed. Microsecond motions in ubiquitin, a small globular protein, are identified, characterized, and compared to the experimentally available microsecond ensemble of ubiquitin structures. The selection of ubiquitin is based on the availability of a number of conformations through NMR studies as well as a large number of X-ray crystal structures. Recent studies by Lange and co-workers have provided insights into the structural heterogeneity of ubiquitin, characterized by NMR at the microsecond time scale [170]. Our results indicate that the computationally obtained slow motions can reliably reproduce the conformational fluctuations observed from experimental techniques. The use of multiple MD simulations based on different crystal structures accounting for the structural diversity in the protein allows the range of slow motions to be explored quickly. Further, the regions identified to have significant conformational flexibility identified by QHA correspond to the regions that show structural deviations in the different structures from X-ray as well as NMR.

However, while interpreting the conformational landscape using any of the PCA techniques, one must be more cautious. Our results indicate that the in the conformational landscape spanned by the top three basis vectors from PCA, the conformational sub-states identified do not share either structural or energetic homogeneity. Further, even though these basis vectors correspond to the slowest modes of motion, our analysis reveals that it is difficult to identify and characterize conformational sub-states relevant to function. This highlights the limitations of using a harmonic approximation to describe an inherently anharmonic landscape. Hence, we motivate the need to characterize anharmonic fluctuations in a protein.

2.2 Simulating Long time-scale Fluctuations from Shorter MD Simulations

This section introduces a strategy to simulate long time-scale from MD simulations from a collection of shorter MD runs. This approach was proposed as early as 1998 [57], which showed that time-scale accessible to MD simulations from a single 1 ns run was shorter than the time-scale accessible to a collection of 10 individual MD runs that lasted 100 ps. Further, Shirts and Pande [243] were able to show that using a large number of smaller MD runs could approximate long time-scale fluctuations derived from a single long MD run. Within the protein folding community, this approach has gathered significant support, especially with the Folding@Home project [213]. Although parallel independent simula-

tions approach to simulate long time-scale motions are valid, one needs to exercise caution while interpreting kinetic parameters from such simulations, as pointed out by Fersht [88].

2.2.1 Ubiquitin: A work-horse for Protein Dynamics

Ubiquitin is found in all eukaryotes, known to have an important role in labeling proteins for degradation [221]. Its structure is evolutionarily conserved, consisting of five β-strands arranged as an antiparallel sheet interspersed with two α-helices located close to the N- and C-termini of the protein [74, 278, 268]. It belongs to the well-known β-grasp fold-family [138]. Ubiquitin is known to bind diverse targets, and therefore, its flexibility associated with binding other proteins is of interest. As illustrated in Fig. 2.2, the protein consists of 5 anti-parallel β-strands ($\beta_1 - \beta_5$) interspersed between two α-helices labeled α_1 and α_2. The loops $\beta_1 - \beta_2$ and $\beta_3 - \beta_4$ and the surface extending from these loops form the primary binding site for ubiquitin. Although there are secondary (and tertiary) binding sites in ubiquitin, the flexibility associated with the two loops ($\beta_1 - \beta_2$ and $\beta_3 - \beta_4$) is of prime interest since it can bind to a variety of substrates [75].

Human ubiquitin (76 residues in a single chain) was selected for computationally identifying and characterizing the microsecond conformational flexibility due to the structural diversity available from different experimental techniques as well as microsecond flexibility as recently observed with NMR [170]. The set of conformations in this NMR study included all of the structural heterogeneity observed from 46 X-ray crystallographic structures in which ubiquitin was complexed with diverse proteins. Moreover, it was observed that a linear combination of a small set of principal components was able to explain the pincer-like motion of ubiquitin residues involved in forming the binding interactions; and conformational selection, rather than the induced-fit mechanism, was sufficient to explain all of the structural heterogeneity in ubiquitin binding dynamics.

The availability of a number of X-ray structure/NMR ensembles and structural deviations within these structures provides information on the conformational fluctuations when the protein samples the kinetically accessible parts of the energy landscape. The recently reported microsecond NMR refinement with orientational restraints (EROS) with 116 structures (PDB code: 2K39) [170] was used for comparisons with the computational results. The 46 X-ray crystal structures of ubiquitin available in the PDB database [44] were also used. These structures were aligned to the reference structure 1UBQ [268] before simulation and analysis. The N-terminal residue 1 and the C-terminal tail consisting of residues 71-76 were excluded from our analysis due to the large displacements in the free ends, and for X-ray ensemble analysis, only the heavy atoms were used. Computational analysis of ubiquitin has served various purposes: (a) validation of force-fields and

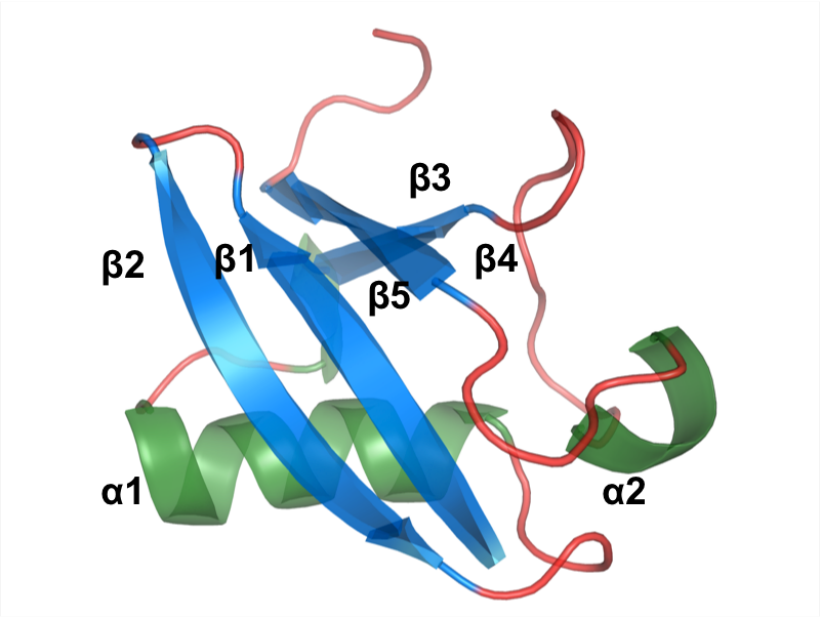

Figure 2.2: **Structure of Ubiquitin.** Ubiquitin is an ubiquitously expressed protein with five β-strands ($\beta 1 - \beta 5$) as well as two α helicaes ($\alpha 1$ and $\alpha 2$). The flexible loops are highlighted in dark red.

simulations [244, 189], (b) possible allosteric effects [182], (c) protein folding [190] and (d) develop novel MD algorithms to simulate longer time-scales [191]. Thus, one may consider ubiquitin to be a work-horse for evaluating protein dynamics at various time-scales.

Eight different starting X-ray crystal structures, which covered the structural diversity of ubiquitin, were used as starting points for MD simulations. These starting structures were obtained from the Protein Data Bank with the following access codes: 1UBQ, 1P3Q (chain U) [268], 1S1Q (chain B) [253], 1TBE (chain B) [66], 1YIW (chain A) [32], 2D3G (chain B) [129], 2FCQ (chain B) [33], and 2G45 (chain B) [228]. Each of the eight crystal structures was processed using the AMBER 8.0 simulation suite and the parm98 force-field [218]. A summary of the structural differences in the eight crystal structures are summarized in Table 2.1.

Structure	Chain	RMSD (Å)	Regions with structural changes
1UBQ	A	(reference)	-
1P3Q	U	0.417	$\beta_1 - \beta_2, \beta_3 - \beta_4$
1S1Q	B	0.501	$\beta_1 - \beta_2, \beta_3 - \beta_4, \beta_2 - \alpha_1$
1TBE	B	0.528	$\alpha_1 - \beta_3, \beta_1 - \beta_2, \beta_3 - \beta_4$
1YIW	A	0.574	$\alpha_1 - \beta_3, \beta_1 - \beta_2, \beta_3 - \beta_4$
2D3G	A	0.380	$\beta_1 - \beta_2, \beta_3 - \beta_4$
2FCQ	B	0.568	$\beta_1 - \beta_2, \beta_3 - \beta_4$
2G45	B	0.570	$\beta_1 - \beta_2, \beta_3 - \beta_4$

Table 2.1: **Summary of starting PDB structures used for detailed molecular dynamics simulations**. These crystal structures were chosen for showing diverse binding partners and also having sufficient structural diversity. RMSD was computed for the backbone of each of the structures to 1UBQ (base structure). The summary of structural changes describes areas highly flexible from the ensemble, as observed by structural overlaps against 1UBQ.

2.2.2 MD Simulations

Each of the eight crystal structures was processed using the AMBER 8.0 simulation suite and the parm98 force-field [218]. Note, we have previously verified the ability of the parm98 force-field for its ability to reliably reproduce conformational fluctuations in proteins [9]. Standard amino acid residues were used to build the protein structures. After determining the protonation state for each amino acid residue at pH 7.0, missing hydrogen atoms were added to the protein. The structures were placed in a rectangular box of SPC/E water [233], such that the distance between the protein and the side of the water box was 10 Å.

The prepared systems were equilibrated using the following protocol. First, the water molecules were minimized using the steepest descent method for 500 steps, followed by conjugate gradients minimization until the root mean square (rms) of the gradients was less than 0.25 kcal/mol. Å. In the next step, the protein atoms were minimized using a similar procedure to release close contacts in crystal structures. A small MD simulation of 25.0 ps with a gradual increase of the temperature in the system to about 300 K, followed by a 25.0 ps constant pressure simulation in which the water molecules were unrestrained to allow occupation of vacuous regions. Five additional steps of equilibration at constant volume were performed with each step consisting of an energy minimization (threshold 0.001 kcal/mol. Å) followed by a 5.0 ps MD run. In these steps, positional restraints

were applied to solute atoms. For the first of the five steps, the force constant was 100 kcal/mol. $Å^2$, followed by a gradual scaling of 0.5 for subsequent steps. The final equilibration was performed without any positional restraints. Another MD simulation with a temperature ramp over 25.0 ps to readjust the temperature to 300 K followed by a 25.0 ps constant pressure MD step to fill any remaining voids was performed to reach the equilibrated structure.

All production runs were performed using the NVE ensemble, with periodic boundary conditions. The particle-mesh Ewald (PME) [223] method was used for electrostatic interactions; a 10 Å cutoff for Lennard-Jones interactions and SHAKE [233] was used for restricting motions of all covalent bonds with hydrogen atoms. All simulations were performed at 300 K and 1 atm pressure. The data from simulations were stored every 1 ps. A total of 62,500 snapshots were collected, totaling 62.5 ns of time for each production run. Collectively for the eight systems, the total conformations sampled are referred to as the 0.5 μs MD ensemble. As an aside, it should be noted that the equilibration procedure and the production run protocols were followed through out the thesis.

Collective Conformational Fluctuations using QHA

QHA, as outlined in Sec. 1.1.3 was performed on the conformations sampled in the MD simulations, and a corresponding principal component analysis was performed with the 116 structures in the EROS ensemble and the 46 X-ray crystal structures. For all eight MD simulations, the structures were fit to the reference structure of 1UBQ to remove rotational/translational movements. QHA modes were computed for the 0.1, 0.2, 0.3, 0.4, and 0.5 μs MD ensembles; for each analysis, 1,250 structures (stored at regular intervals) from each of the eight MD simulations, therefore a total of 10,000 structures, were used for each ensemble. Note that the 0.1, 0.2, 0.3, 0.4, and 0.5 μs MD ensembles are based on the 12.5, 25.0, 37.5, 50, and 62.5 ns individual MD trajectories, respectively, from each of the eight MD simulations. For each of the ensembles, the atomic fluctuation matrix 1.6 was built for all atoms and diagonalized using the ptraj [56] program as part of the AMBER suite. The results from QHA were visualized using custom scripts written in Python [265] and PyMOL [73]

Normal Mode Analysis

Twelve structures from each simulation (at 5 ns, 10 ns, 15 ns, ..., 55 ns, 60 ns), therefore, a total of 96 structures from the eight simulations, were analyzed with NMA. Prior to NMA, the starting structure was minimized using both the Newton-Raphson and con-

jugate gradients minimization methods until an rms tolerance of 10^{-4} kcal/mol. $Å^{-1}$ was reached. Solvent was excluded from the calculations, and the nmode program [204] from the AMBER suite of programs was used to compute the normal modes for ubiquitin.

2.3 Slow Conformational Dynamics of Ubiquitin

Slow conformational fluctuations in ubiquitin were identified with QHA and dPCA from the conformations extracted from the MD trajectories. In the most straightforward way, these fluctuations could be identified by generating a microsecond trajectory followed by the QHA of the system snapshots collected during the simulation. However, currently, computing microsecond trajectories of even small proteins is computationally expensive and time-prohibitive, as it would require several months of simulation run-time even with the best computer hardware. As an alternate approach, for this study, eight independent MD simulations each based on a different starting X-ray crystal structure and 62.5 ns in duration were generated. The total set of conformations used for QHA corresponds to 0.5μs sampling of the potential energy surface (obtained by combining all eight MD simulations). This approach leads to an obvious question: does the analysis of this total ensemble provide information about ubiquitin flexibility at the microsecond or nanosecond time scale? We provide an answer to this important question below after the detailed characterization of the QHA modes and comparison with the experimental information.

2.3.1 Analysis of Fluctuations in Ubiquitin

Fig. 2.3 depicts the ubiquitin backbone flexibility as identified on the basis of the individual MD trajectories as well as the 0.5μs MD ensemble. An indication of the backbone flexibility of ubiquitin is provided by the root-mean-square fluctuations (rmsfs) as computed from the 0.5μs ensemble of MD structure. The rmsf includes the complete set of motions; however, the motivation of this study is to characterize the slow conformational flexibility. Therefore, slow conformational fluctuations were identified by computing QHA modes from a set of 10,000 conformations in the individual 62.5 ns MD trajectories as well as the entire 0.5μs MD ensemble. The modes computed from the 0.5μs MD ensemble are referred to as $\text{QHA}_{0.5\mu s}$ in the remaining text of this report. As depicted in Figure 3, QHA analysis of individual 62.5 ns MD trajectories can identify the mobile regions qualitatively; however, it is unable to correctly characterize the range of motions (displacements) due to limited coverage of the energy landscape. An aggregation of the slowest 10 $\text{QHA}_{0.5\mu s}$ modes (obtained by summing up the atomic displacements in the modes) indicates that

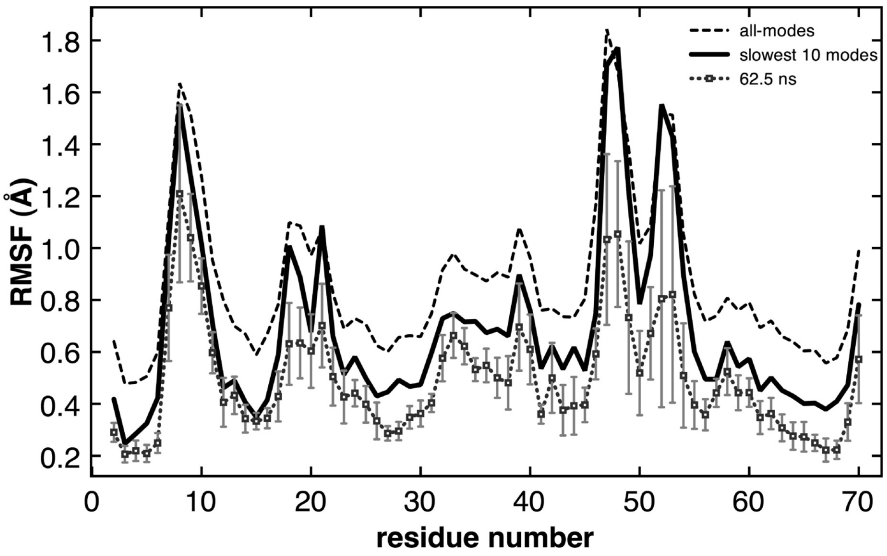

Figure 2.3: **Similarity in ubiquitin backbone flexibility as characterized by 0.5 μs and 62.5 ns conformation sampling.** The gray curve corresponds to an average of three individual MD trajectories (based on the 1P3Q, 1S1Q, and 1UBQ crystal structures) each 62.5 ns in duration, while the solid black curve corresponds to a total of 0.5 μs sampling from eight separate MD simulations. Flexibility as quantified by the slowest 10 QHA modes (solid black curve) is compared to the total rmsf of the 0.5 μs MD ensemble (dashed black curve), which corresponds to all modes.

they are sufficient to capture the majority of fluctuations in the most mobile region of ubiquitin when compared to all motions; in particular, these slowest modes show similar displacements in the most mobile regions of the ubiquitin backbone. Note the slowest modes are defined by the low frequencies (eigenvalues) corresponding to the modes. Quantitative estimates indicate that the slowest 10 modes represent 78.4% of flexibility reflected in the 0.5μs ensemble (the slowest 20 modes capture 87% and the slowest 50 modes capture 94% of flexibility). Therefore, the slowest 10 modes were selected for detailed analysis and characterization. The advantage of using the slowest 10 modes instead of the total rmsfs is that it can mitigate the effects of "noise" represented by fast fluctuations (at short time-scales) which are irrelevant for collective fluctuations involving global motions in the protein.

Ubiquitin shows considerable flexibility at the microsecond time scale (see Fig. 2.4). In addition to the $\text{QHA}_{0.5\mu s}$, ubiquitin flexibility was also characterized on the basis of the structures from the X-ray and NMR ensembles. Modes based on 46 X-ray structures

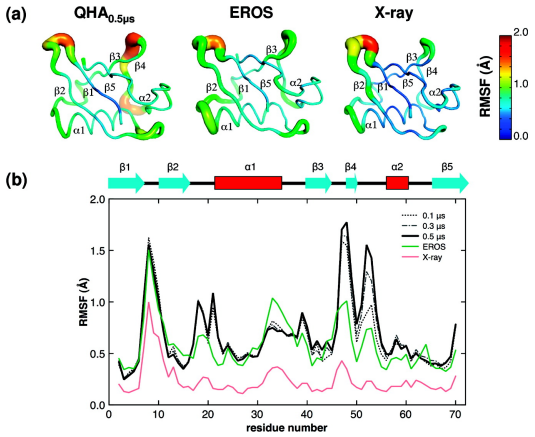

Figure 2.4: **Microsecond conformational fluctuations in ubiquitin as identified by QHA$_{0.5\mu s}$, NMR (EROS), and X-ray ensembles show high degree of similarity in flexible regions.** (a) Inverse frequency-weighted positional fluctuations for C^{α} atoms are shown in a tube-like representation; thicker tubes represent large-scale fluctuations, and thinner tubes represent lower fluctuations. The actual magnitude of the fluctuation is also color-coded, with red indicating regions with the highest conformational flexibility and blue representing regions within the protein that are relatively less flexible. (b) Comparison of the ubiquitin backbone fluctuation at the microsecond time scale. Inverse frequency-weighted positional fluctuations in ubiquitin. The positional fluctuations from MD ensembles are shown compared with the microsecond scale NMR and X-ray ensemble. The secondary structure of ubiquitin is overlaid on top of the plot for ease of visualization and identifying regions that show high flexibility.

and 116 structures in the NMR ensemble (EROS) were computed in a similar way to the QHA$_{0.5\mu s}$. Note that the EROS ensemble corresponds to the microsecond time scale as defined by the resolution of the NMR experiments [170]. Fig. 2.4a depicts slow conformational fluctuations of ubiquitin as an aggregate of the top 10 slowest modes from the QHA$_{0.5\mu s}$, EROS, and X-ray ensembles. The microsecond scale flexibility is observed in three major regions: $\beta_1 - \beta_2$ hairpin, $\alpha_1 - \beta_3$ loop, and $\beta_3 - \alpha_2$ loop. The computational results agree with the EROS results, except for the $\beta_3 - \alpha_2$ region, where the simulations indicate higher flexibility; also, computational simulations show additional flexibility in the $\beta_2 - \alpha_1$ region (see further discussion below). Fig. 2.4b depicts that computationally even at $0.1\mu s$ there is qualitative agreement within the experimentally observed flexibility, as the regions of high flexibility are similar. However, with additional sampling of the potential energy surface (0.2, 0.3, 0.4, and 0.5μs), the amplitude of the motions in the flexible

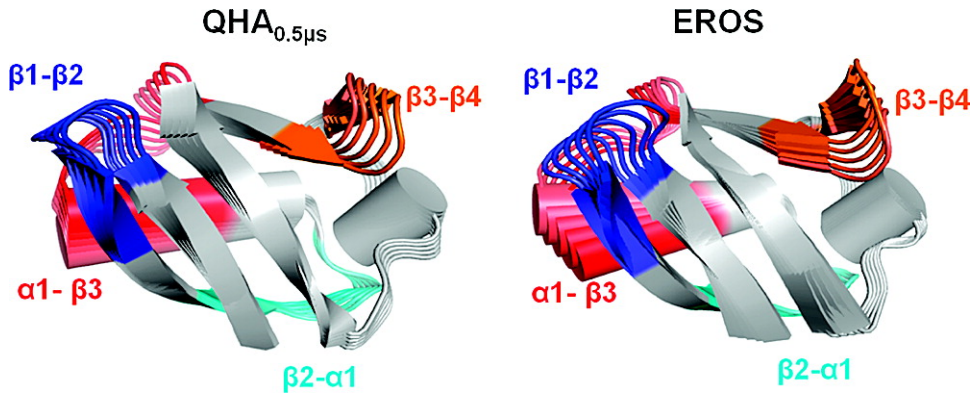

Figure 2.5: **Similarity in directions of the slowest mode of ubiquitin at the microsecond time scale based on the QHA$_{0.5\mu s}$ (computational) and EROS ensemble (experimental).** The modes are depicted in a movie-like fashion, with subsequent conformations shown in lighter colors. Four regions of the protein are highlighted with different colors and labeled; these regions indicate large displacements. These motions involve a pincer-like motion involving $\beta_1 - \beta_2$, the C-terminal part of the α_1 helix, and $\alpha_1 - \beta_3$ loops. The amplitudes of the motions are arbitrarily scaled for visualization; however, see text for a comparison of the coverage of the conformational landscape.

regions of the protein becomes enhanced and is able to qualitatively reproduce the experimentally observed backbone flexibility in the microsecond NMR ensemble. While the NMR ensemble agrees with computationally characterized flexibility, the X-ray ensemble shows subdued motions. This is potentially an artifact, as the amplitude of fluctuations observed are based on simulations of unbound ubiquitin in solution, compared to the ubiquitin from X-ray ensembles where it is present in complex with other proteins [45]. This has also been reported previously in the comparison of X-ray structure with the NMR ensemble [170]. Particularly, the residues Ile44 and His68 show close contacts with binding partners in the X-ray structures; in simulations where the binding partners are absent, these regions show increased flexibility.

2.3.2 Comparing Directions of Motions from Experimental and MD ensembles

Individual slow modes from the QHA$_{0.5\mu s}$ and EROS ensembles provide interesting insights into intrinsic large-scale motions of ubiquitin with a potentially functional role. Fig.

2.5 shows a movie-like representation of the slowest mode from the two ensembles; this mode shows an opening/closing motion associated with the binding site of ubiquitin [154]. As also previously reported, this mode involves the pincer-like motion of $\alpha_1 - \beta_3$, $\beta_1 - \beta_2$, and $\beta_3 - \beta_4$ loop regions. QHA$_{0.5\mu s}$ also reveals large motions in the $\beta_2 - \alpha_1$ regions and to a smaller extent in the $\beta_4 - \alpha_2$ loop region (see the Supporting Information for animated movies of the modes). These two loops show highly correlated movements with the opening/closing motions. The motion of residues Ile44, Lys63, and His68 that make close contacts with the binding proteins are observed to be restricted in this mode, which also agrees with the findings from the study of the EROS ensemble. A comparison of other slow modes indicates that the magnitude and the directions of motions in modes 2-10 from QHA$_{0.5\mu s}$ and EROS ensembles were also similar. As mentioned above, it was found that the slowest 10 modes contribute $>$78% of the ubiquitin flexibility; therefore, the similarity in the slowest 10 modes between the computational and experimentally calculated modes indicates overall similar microsecond flexibility of ubiquitin. Overall, similarity in flexible regions and the displacement amplitudes within modes from the MD simulations and experimental ensembles (EROS) indicates that QHA is able to reproduce the conformational fluctuations of ubiquitin.

The ability of QHA$_{0.5\mu s}$ to reproduce the experimentally observed ubiquitin flexibility at the microsecond time scale was characterized in further detail. In order to perform a quantitative comparison, the large-scale fluctuations computed using QHA on the total 0.5μs MD simulations and EROS ensemble were compared by calculation of the overlap for the slowest modes. The overlap between two subspaces spanned by the eigenvectors is defined by Hess's metric [126] as follows:

$$\gamma = \frac{1}{D_1} \sum_{i=1}^{D_1} \sum_{j=1}^{D_2} (\mathbf{v}_i^1 . \mathbf{v}_j^2)^2 \tag{2.1}$$

where D_1 and D_2 represent the number of eigenvectors considered from each of the ensembles; v_i^1 represents the ith eigenvector from each of the ensembles. γ indicates the degree of similarity of motions between the two ensembles, with a value of 1 indicating identical motions. On the basis of the observation that the slowest 10 modes are able to qualitatively reproduce the flexibility in the most mobile regions of the proteins, we computed the overlap for the slowest 10 modes. As indicated in Table 2.2, the QHA modes computed on the basis of 0.1, 0.2, 0.3, 0.4, and 0.5μs MD sampling show close to 80% overlap, indicating that the large-scale conformation fluctuations are reproducible across various sections of the MD sampling. This also provides an indication of the robustness of QHA methodology in identifying slow conformational fluctuations.

A corresponding calculation of overlaps of the modes from MD simulation with that

	0.1 μs	0.2 μs	0.3 μs	0.4 μs	0.5 μs
0.1 μs		0.872	0.795	0.774	0.755
0.2 μs			0.880	0.855	0.783
0.3 μs				0.980	0.880
0.4 μs					0.895
EROS	0.748	0.747	0.749	0.751	0.747
X-ray	0.405	0.396	0.362	0.364	0.349

Table 2.2: Similarity of Motions Spanned by QHA, EROS, and X-ray Ensembles for Ubiquitin. Each entry in the table is computed via eq 2, for the slowest 10 modes determined from each of the ensembles from the MD simulation.

of EROS also indicates about 75% agreement between the two ensembles (see Table 2.2). Note that the window of previous NMR investigations has been very broadly defined on the microsecond time scale. The X-ray ensemble shows a lower degree of similarity; as mentioned before, this is possibly due to the comparison of bound ubiquitin in X-ray complexes with apo-ubiquitin. The largest differences in flexibility between the computational and EROS ensembles are observed in the $\beta_2 - \alpha_1$ and $\beta_3 - \alpha_2$ loops. It is possible that the 0.5μs scale simulations have allowed $\beta_2 - \alpha_1$ and $\beta_3 - \alpha_2$ loop regions to explore areas within the potential energy surface that were not accessible during the window of the EROS or the X-ray ensembles, particularly the higher energy regions of the conformational energy landscape that were not accessible to EROS/X-ray ensemble structures (see the discussion below on conformational sampling of various regions of the conformational energy landscape). Additionally, it is also possible that the force-field overestimates the flexibility of these two regions in the MD simulations or that the NMR investigations have underestimated the flexibility of the protein. Table 2.2 and the information in Figure 2.4b also indicate that the extent of conformational fluctuations improves with additional sampling between the 0.1 and 0.5μs ensembles. The comparison of the overlap between these ensembles (see the last column in Table 2.2) indicates convergence toward the conformational flexibility at the microsecond time scale. Additional sampling of the conformational landscape may lead to changes in the overlaps; however, as discussed below, these changes are not expected to be qualitatively much different.

2.3.3 Conformational Sub-states spanned by Low-frequency Modes

The ability of the slow modes to cover the conformational landscape was also characterized on the basis of the calculation of projections along the MD trajectories and NMR

ensemble. The projections from each of the ensembles were calculated using:

$$q_i(t) = (\mathbf{x}(t) - \langle \mathbf{x} \rangle).\mathbf{v}_i \tag{2.2}$$

where $q_i(t)$ is the projection of the tth snapshot in the trajectory, $x(t)$ is the corresponding positional vector, $\langle x \rangle$ is the mean positional vector, and v_i represents the ith eigenvector. The extent of overlap between the projections from the slow modes has been previously used in the literature to characterize the sampling of the conformational energy landscape in computational methods as well as the structural deviations observed in experimental techniques (including NMR and X-ray ensembles). Within the projections of the slowest two modes (see Fig. 2.6), it was observed that the EROS ensemble is entirely covered by the combined MD simulation. This indicates that 0.5μs aggregate MD simulations were sufficient to cover all of the states visited by EROS ensembles. In addition, MD simulations are also able to cover certain areas of the potential energy landscape not observed from the experimental ensembles (explaining the larger motions sampled by MD). This is especially evident at the right side of Fig. 2.6a, which is markedly devoid of any EROS related structures. Similarly, other slower modes also indicate that the MD simulation covers all of the conformations from the EROS ensemble (Fig. 2.6b), indicating a widespread agreement between the subspaces spanned by MD and EROS ensembles.

The use of eight separate MD simulations based on different starting structures allows characterization of microsecond flexibility with relatively short trajectories (62.5 ns each). As is depicted in Fig. 2.6, the slowest QHA$_{0.5\mu s}$ modes cover the part of the landscape that is also explored by the EROS ensemble. The projection of the slowest QHA$_{0.5\mu s}$ modes on the MD and NMR structures indicates that starting MD trajectories from different structures allows quick sampling of the neighboring area of the landscape, while it could be envisioned that the longer trajectories would allow sampling of other distant areas of the landscape. From these projections, up to six areas or clusters are identified and marked by the ellipses in Fig. 2.6. These clusters correspond to areas of the landscape visited by MD simulations. Note that some individual trajectories visited more than one cluster (for example, 1YIW and 2G45) with the intersection point or overlap between the ellipses corresponding to higher energy states or transition points. Any of the eight ubiquitin MD simulations less than 50 ns did not show the presence of these conformational states or transition points. The use of multiple MD simulations allows quick sampling of the different clusters, which would otherwise take much longer simulation times. An interesting observation from these plots indicates that the projections from the NMR ensemble are located in one or more clusters, which are also the most visited area by the various MD trajectories. This common area of the plot is the most visited area by multiple MD simulations. This could possibly indicate the presence of a more populated conformational substate of ubiquitin that is visited both during the computational (MD) and experimental

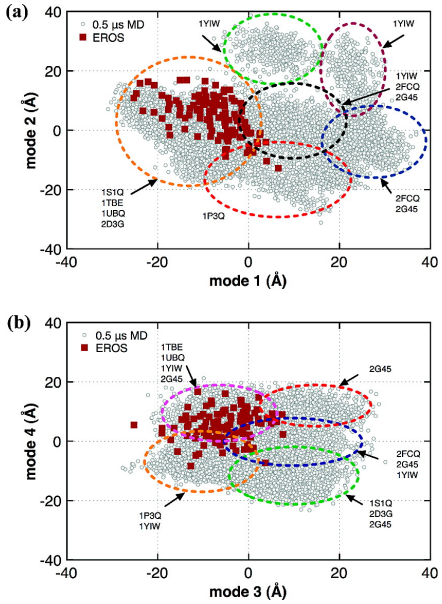

Figure 2.6: **Projections of the slowest modes show considerable overlap between QHA$_{0.5\mu s}$ and EROS ensembles.** Projections of the slowest modes from the microsecond MD ensemble and experimental EROS for (a) mode 1 vs mode 2 and (b) mode 3 vs mode 4. Gray circles represent the ensemble of structures from MD simulations, and red squares represent EROS structures. The slowest four modes computed with QHA$_{\mu s}$ were used calculating the projections for structures from all eight MD simulations and the NMR ensemble. As marked with ellipses, the projections can be separated into six and five clusters, one or more of which are sampled by individual trajectories. Note that the computed projections represent summation over all atoms in the protein.

(NMR) methods. These plots also provide an indication that the 0.5μs MD ensemble has covered the extent of conformational landscape accessible by ubiquitin at the microsecond time scale. Additionally, the coverage of a larger portion of the landscape (beyond what is explored by NMR) provides an explanation of higher flexibility indicated by the MD simulations. These observations are consistent with the overlaps listed in Table 2.2. Further MD sampling would lead to filling out the gaps in these plots (sampling of the higher energy transition points); therefore, qualitatively, any significant differences in the slow conformational fluctuations are not expected.

In order to further characterize the nature of conformational sub-states, we also looked at the distribution of internal energy of each conformation along the slowest three modes

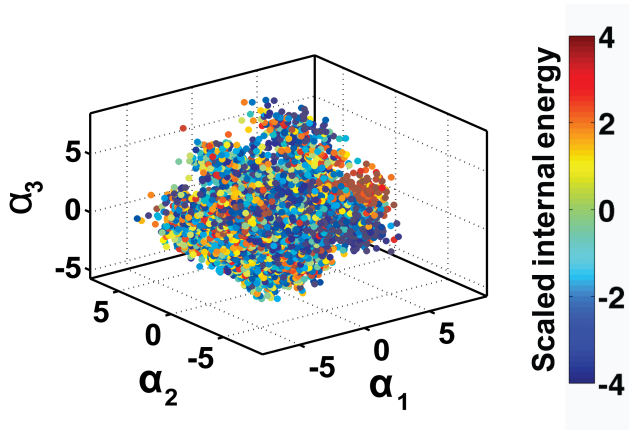

Figure 2.7: **QHA description of ubiquitin landscape as spanned by the top three basis vectors ($\vec{\alpha}$).** Note the lack of homogeneity in the internal energy distributions of QHA. The top three basis vectors from QHA ($\alpha_{1...3}$) were used to project the 10,000 conformations from the MD simulation. The projection of each conformation is colored by the scaled internal energy. Note the apparent lack of sub-states (clusters).

determined via QHA. This is illustrated in Fig. 2.7. Each conformation from the MD simulation is projected onto a space spanned by the top three basis vectors (slowest frequencies), which contribute to over 50% of the overall variance. Each conformation is also painted by the scaled internal energy of the protein. Although, one can clearly observe the presence of several conformational clusters, the distribution of the scaled internal energy is somewhat heterogenous.

2.3.4 Comparison of QHA$_{0.5\mu s}$ with NMA

NMA provides a quadratic approximation of the potential energy surface based on a reference structure, and is well suited to study motions close to a local energy minimum (see Fig. 2.1). NMA was computed for 12 structures, each separated by 5 ns for each of eight MD simulations; therefore, a total of 96 sets of NMA modes were obtained (see the Methods section for more details). The results are summarized in Fig. 2.8 and 2.9. As depicted in Figure 2.8, various ubiquitin regions show high flexibility; however, there are considerable differences in the regions of larger flexibility between different structures. The highlighted regions particularly show flexibility that is not reproduced in other structures. Even though the regions of flexibility are similar to the QHA$_{0.5\mu s}$, the magnitude

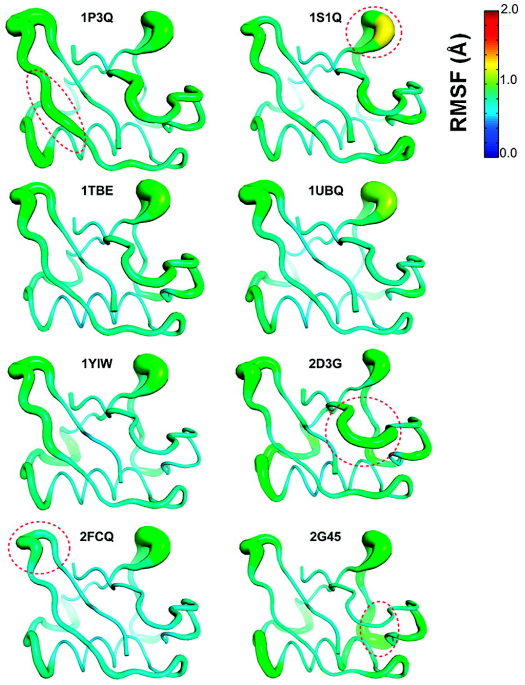

Figure 2.8: **Average RMSF determined on the basis of NMA of ubiquitin based on the eight structures show considerable variation in fluctuation profiles**. The displacement vectors for the slowest 10 modes were aggregated (by summing all atomic displacements in the modes) and averaged for 12 conformations. Note that the fluctuations are colored on the same scale as the results of the QHA analysis (Fig.2.4). The circled areas indicate regions with flexibility which is not reproduced in other structures. These results show lower amplitudes and less reliable modes than QHA.

of displacement is much smaller for NMA modes when compared to the QHA modes, as is visible by comparison with colored regions in Fig. 2.4 (note that the color bars are on same scale for the two figures). This difference possibly exists because NMA characterizes motions that explore the immediate vicinity of the local minimum of the reference structure (gray area in Fig. 2.1), whereas QHA of the 0.5μs ensemble indicates sampling of a larger portion of the conformation energy landscape [25] The conformational flexibility identified by NMA shows considerably lower agreement with the EROS ensemble as well (see Table 2.3); therefore, it is much less able to reproduce the experimentally observed ubiquitin flexibility.

Characterization of the correlation between the NMA modes indicates several interest-

	EROS	1P3Q	1TBE	1S1Q	1UBQ	1YIW	2D3G	2FCQ	2G45
$QHA_{0.5\mu s}$	0.713	0.641	0.490	0.540	0.721	0.457	0.600	0.441	0.401
EROS		0.607	0.624	0.601	0.690	0.502	0.562	0.625	0.436

Table 2.3: **Correlation of Positional Fluctuations between NMA, QHA, and EROS Modes.** The comparison represents the correlation between inverse frequency-weighted top 10 modes. NMA shows a lower agreement with EROS as compared to QHA results.

ing aspects of ubiquitin flexibility. As depicted in Figure 2.9, the NMA shows considerable variation during the MD trajectory, as indicated by considerably lower correlations with the other structures in the same MD simulation. For 1P3Q, 1YIW, and 2G45 in particular as the trajectory evolved, the lower correlation indicates changes in normal modes. This is also possibly an indication of the MD simulation sampling different areas of the energy landscape. However, in the case of 1TBE and 1UBQ, the normal modes show a larger correlation, possibly indicative of MD simulation sampling the neighboring areas of the energy landscape. It is interesting to note that these observations are similar to the one mentioned above for QHA.

2.4 Conclusions

2.4.1 Summary

The microsecond flexibility in ubiquitin was characterized using QHA of the conformations generated along eight MD trajectories starting from eight different crystal structures. Overall, the total ensemble corresponded to 0.5 μs sampling. In addition to significantly reducing the run time of the simulations by performing these eight runs simultaneously, this approach enables the sampling of separate areas of the energy surface due to starting from different crystal structures. The characterization of the slow conformations indicated that the top 10 modes accounted for over 78% of the flexibility observed at the microsecond time scale. The slowest mode indicated pincer-like motion, as was previously identified using microsecond NMR. This motion has been implicated in ubiquitin binding to other proteins in solution. The identified motions were compared to the structural deviations observed in the collection of ubiquitin structures in complex with other proteins, as well as the recently characterized NMR ensemble. The degree of similarity between NMR ensemble flexibility and the $QHA_{0.5\mu s}$ modes was observed to be close to 0.75 (for the top 10 modes), indicating a significant agreement in the nature of the slow motions

at the microsecond time scale. Further, the coverage of the conformational landscape by the computationally and experimentally identified flexibility was observed to be similar, even though individual trajectories sampled only 62.5 ns. The use of multiple trajectories allows for most of the conformational flexibility at the microsecond time scale to be reproduced. Therefore, we believe that combination of multiple trajectories provides an efficient method to explore long scale conformational fluctuations. Other investigations have also reported similar observations [57, 58].

2.4.2 Perspectives on using PCA and NMA

Perhaps, the most widely used techniques in understanding intrinsic motions of proteins are normal mode analysis (NMA) [25, 27] and principal component analysis (PCA) [142, 150, 15]. NMA based approaches are popular due to their inherent simplicity: beginning with a single X-ray crystal structure or an experimental ensemble of structures, it is possible to obtain fundamental insights into the internal motions and flexibility patterns of a protein. NMA based approaches have proven useful to understand the molecular basis of biophysical processes such as ligand-binding [24, 30]. Several studies have shown the ability of NMA to correlate with experimental B-factors or order parameters [28]. In the context of enzyme catalysis, NMA has been useful to assess the role of collective and intrinsic mobility of specific amino-acid residues in enzymes as a mechanistic requirement for function [287, 24].

A number of previous studies have been published, highlighting the usefulness and limitations of NMA based techniques. Ma has pointed out that while the directions as indicated by the low-frequency motions describe the overall trends of conformational changes, the motions of side-chains (which are at much higher frequencies) need to be interpreted with caution [185]. Further, a careful and extensive analysis of NMA based approaches by Kondrashov and co-workers [160] also showed that incorporating chemical specificity of residues can tremendously increase the accuracy of thermal B-factor prediction in NMA. A comparison of NMA with MD based approaches also showed that in general, a larger number of NMA modes were required to explain the same variance from MD [231].

While motions involved in ligand-binding and complex formation are the forte of NMA based approaches, it must also be assessed in the context of enzyme catalysis. Chemical catalysis involves a detailed understanding of chemical processes in the active site where atom-to-atom contacts and specific interactions play a critical role. Such chemical specificity is not accessible to coarse-grained NMA approaches and with all-atom NMA, the harmonic approximation typically breaks down since the formation/ deletion of a chemical bond is away from the equilibrium. Since NMA cannot explain structural deviations

away from equilibrium, techniques such as QHA are used to approximate the potential energy landscape as sampled from MD simulations [7]. QHA has been widely used to study biochemical processes such as enzyme catalysis and more recently, has also proven useful to interpret experimentally observed conformational transitions for the enzyme adenylate kinase (Adk) along the reaction trajectory [125]. It was shown that large-scale motions in Adk at millisecond time-scales (as accessed by NMR experiments) are infrequent, while faster and frequent motions at nano-seconds of smaller amplitude are along the direction of this large-scale conformational change.

The conformational fluctuations as described by QHA correlate well with experimentally determined large-scale motions. This is because QHA approximates the conformational landscape spanned by a protein by pursuing *variance*, a second order statistic. While variance is a good measure to describe the extent of conformational fluctuations, it should also be noted that variance is very sensitive to outliers in the underlying data. Thus, rare occurrences of extreme fluctuations in a MD simulation can have significant (and potentially adverse) effects on QHA and can lead to rather non-intuitive description of the landscape spanned by the protein. QHA has no intuition about the rich statistical diversity positional fluctuations may exhibit. Especially, it cannot differentiate if the positional fluctuations arose as a consequence of sampling uniformly along a particular direction or from rare and extreme deviations from the mean atomic position.

The consequence of such also note here that the sub-states identified via QHA do not exhibit conformational or energetic similarity. In chapter 3 of this thesis, we present a case for describing protein motions not only in terms of the variance but also higher-order moments determined from the MD simulation. As will be seen, the ability to describe the statistical regularities arising out of positional fluctuations will enhance not only the interpretability of the results, but also provide a natural means to organize the complex conformational landscape spanned by a protein. The higher-order moments, as will be seen, provide an intuitive and powerful means to describe large-scale conformational fluctuations associated with the landscape. This new method, as we will show, further strengthens the case for the role of intrinsic flexibility associated with the protein's fold and its function.

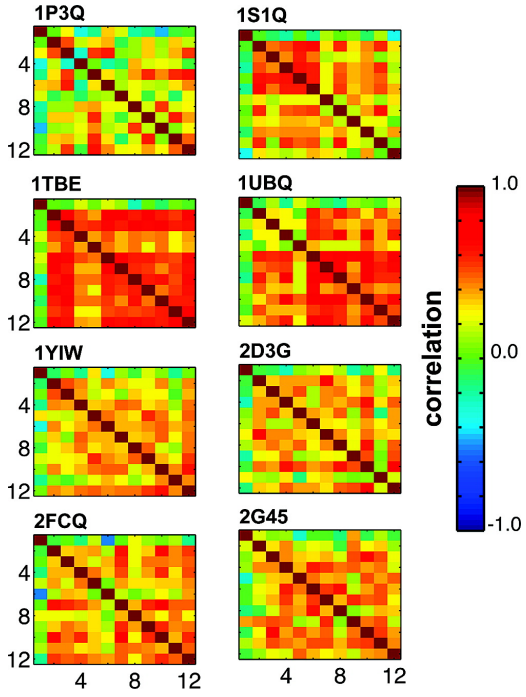

Figure 2.9: **Correlation of positional fluctuations as computed by NMA for structures along the eight MD trajectories.** The eight plots indicate the eight MD systems used in this study. NMA was performed on 12 structures in each trajectory separated by 5 ns (5 ns, 10 ns, ..., 60 ns), and the degree of correlations of normal modes between the 12 structures is shown as a matrix. The degree of correlation was obtained as taking a dot product of all eigenmodes followed by normalization. These results indicate a change in NMA modes over the course of the MD trajectory, except for 1TBE and 1UBQ that seem to be sampling in the nearby areas.

Chapter 3

Quasi-Anharmonic Analysis: Modeling Internal Motions and Energetics using Higher-order Correlations

A question that we examine in this chapter is how much of the protein exhibits anharmonic behavior under physiologically relevant temperatures at typically long time-scales. We use the example of ubiquitin, a widely characterized protein is used to understand the extent of anharmonic behavior in solution. We show that over 60% of the individual residues in ubiquitin exhibit predominantly anharmonic behavior and these residues are implicated in its binding. A novel linear computational model, quasi-anharmonic analysis (QAA) is then developed to characterize the anharmonic conformational landscape spanned by ubiquitin. We show that QAA elucidates the hierarchical nature of the equilibrium conformational landscape and identifies conformational sub-states in ubiquitin. These sub-states share significant conformational and energetic homogeneity. Further, the motions elucidated from QAA indicate that as ubiquitin "jumps" from one sub-state to the other, the binding regions are modulated in such a way that the binding regions can adapt themselves to accommodate multiple substrates. The conformational hierarchy also reveals that at every level, even though the intrinsic motions (or the shape) does not change, the amplitude of fluctuations in these binding regions can be quite different, leading to the natural plasticity observed in experiments.

3.1 Introduction

Native fluctuations allow a protein to access various conformational and energetic sub-states (of short durations [90]), which are closely linked to protein function including enzyme catalysis [47]. Experimental and computational investigations continue to provide fascinating insights into the spatial and temporal hierarchy of conformational fluctuations [47, 40, 225, 80, 92]. Analyses of internal energetics [283, 97, 292] have established a close tie between collective fluctuations, protein geometry, and energy transduction [176]. However, a grappling issue faced by biologists is the inability to intuitively visualize conformational sub-states as well as their inherent dynamics and energetics.

In Chapter 2 of this thesis, we have discussed the applicability of quasi-harmonic analysis (QHA) to characterize the long time-scale fluctuations in ubiquitin. It was shown that by accounting for the structural diversity of ubiquitin and using multiple molecular dynamics (MD) simulations, one can reliably estimate both long time-scale (slow) conformational fluctuations in terms of the variance in displacements and the directions of these displacements. Computationally, the observed RMSF and slow conformational fluctuations in human ubiquitin using a 0.5μs explicit solvent simulation were also validated [225].

However, the harmonic [128] or quasi-harmonic [180] approximation used, has met with limited success in isolating, visualizing and intuitively explaining conformational sub-states in a protein [192, 31, 169]. As examined in Chapter 2, although variance captures the extent of atomic positional deviations from a mean position, it is highly sensitive to outliers in the data. This blind pursuit of variance can in fact drastically affect the interpretation of results. Fig. 2.7 showed that the conformational sub-states spanned by the lowest frequency modes do not show significant conformational or energetic homogeneity, which can be a huge disadvantage when one wants to gather mechanistic insights into the working of a protein. At long (μs) time-scales, molecular dynamics (MD) simulations reveal a much richer statistical diversity of atomic fluctuations which can be conceptualized as mixtures of harmonic and anharmonic fluctuations [119] and this needs to be accounted for, when describing large-scale, collective fluctuations.

Anharmonic motions can arise when a protein visits multiple sub-states [93] or samples motions distant from equilibrium, leading to long-tailed, non-Gaussian distributions [208, 207, 135, 136]. A key question that we would like to answer in this chapter is how anharmonic are atomic fluctuations at long time-scales (μs and beyond) that are relevant to biological activity. It has been argued that anharmonic motions dominate the conformational landscape beyond the dynamical transition temperature [215, 93, 216]. Also, for any biological activity, anharmonic motions tend to contribute a significant proportion of

the observed RMSF [118]. In this context, the characterization of individual atomic fluctuations in terms of anharmonic behavior becomes extremely important.

RMSF, which measures the square root of variance, is valuable in identifying which parts of a protein are flexible; however, it does not indicate the underlying statistical regularity of the positional fluctuations. These statistical regularities arising from positional fluctuations allow the protein to be in multiple sub-states. To describe the regularities in the conformational landscape arising from anharmonic fluctuations requires the use of *higher-order statistics* [119]. Previous work characterizing higher-order moments in MD simulations used only picosecond length trajectories to quantify anharmonicity in individual atomic fluctuations [208, 135] and also to refine X-ray crystallographic data [136]. In this chapter, we extend and elaborate on the properties of anharmonic behavior in the protein ubiquitin by studying long time-scale simulations. We first focus on individual residues, since these constitute the functional units of any protein. To quantify the degree of anharmonicity, we examine the *kurtosis*, κ, which is defined as the fourth-order moment of a real-valued random variable z as:

$$\kappa(z) = E\{z^4\} - 3(E\{z^2\})^2 \text{ [214]}. \tag{3.1}$$

Intuitively, κ is a measure of "bulge" or "peakedness" in a probability distribution. An equivalent way of describing κ is the *excess* kurtosis, which is defined as $[E\{z^4\}/E\{z^2\}] - 3$. κ is specifically useful to characterize the 'tails' of a probability distribution. Intuitively, (as illustrated in Fig. 3.1) it explains if the variance in the underlying distribution is a result from infrequent yet extreme fluctuations or from frequent and modest deviations from the mean position. It is important to note that κ can be large just because the distribution is multi-modal (as illustrated in Fig. 3.1). Indeed, as will be seen in Fig. 3.5, κ provides one of the motivations to characterize anharmonic motions in a protein using higher-order moments.

We next ask the question if it is possible to exploit the higher-order dependencies (correlations) in intrinsic fluctuations and use that to build a computational model that describes the conformational landscape spanned by ubiquitin at long time-scales. We introduce a linear computational model, termed *quasi-anharmonic analysis* (QAA), which effectively decouples observed higher-order dependencies and efficiently summarizes internal protein motions. Our approach builds a novel representation for protein motions at long (μs) time-scales by explicitly pursuing higher-order statistics. Human ubiquitin is used as the primary example to understand the nature of anharmonic conformational fluctuations at the microsecond time-scales using higher-order statistics. We use the 0.5μs simulation of ubiquitin as described in Chapter 2 of this thesis to characterize the anharmonic landscape of ubiquitin.

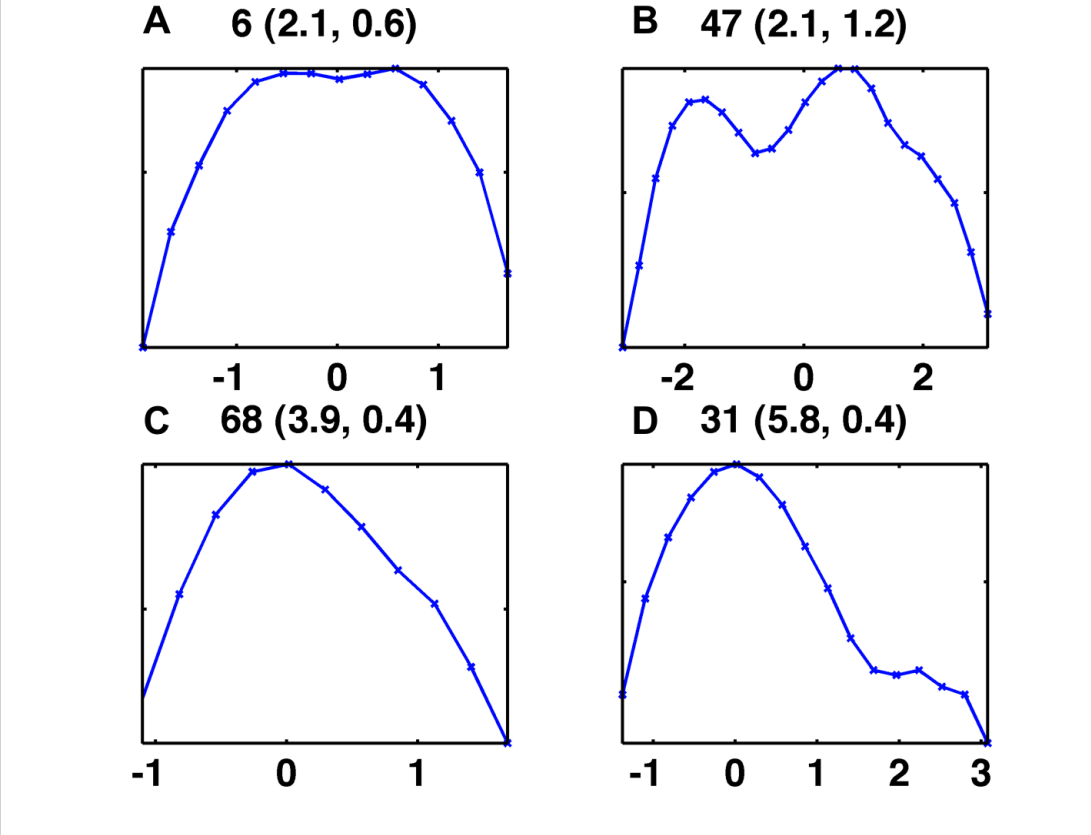

Figure 3.1: **Illustration of κ for atomic fluctuations reveals statistical diversity.** We consider ubiquitin to illustrate the typical statistical behavior of positional fluctuations in C^{α} atoms of some residues. The top of each panel indicates (κ, RMSF^2) values. It is important to note that κ qualifies the 'tails' of a distribution. For example, in panel (A) observe that the variance is about 0.6 Åand κ is quite small, where as in panel (B) even though κ is equal to that of panel (A), the variance is 1.2 Å. In panel (C), even though the variance is almost similar to that of panel (A), κ indicates that there is a noted peakedness towards the right had side of the distribution and in panel (D), there is a marked peakedness towards the right of the distribution. Thus, κ reveals richer statistical diversity in the underlying data.

3.2 Anharmonic behavior in protein motions

3.2.1 Quantifying anharmonicity in protein fluctuations

The N-terminal (residue 1) and C-terminal end (71-76) exhibit considerable degree of flexibility, and hence for this analysis, we will illustrate our results only on residues 2-70, which form the functional core of the protein. This approach of using only residues from 2-70 was done in the previous chapter (2) and elsewhere [170]. Anharmonicity (or non-Gaussianity) in positional fluctuations was characterized using the fourth-order statistic kurtosis, κ as explained in Eq. 3.1. $\kappa < 3.0$ and $\kappa > 3.0$ indicate that the underlying distribution is sub-Gaussian (G_s) and super-Gaussian (G^s), respectively. For a Gaussian (G) random variable with unit variance, κ is 3.0. We will use κ to study the anharmonicity observed in ubiquitin motions from both experimental ensembles (116 NMR structures revealing up to μs dynamics (pdb: 2K39) [170], and 44 X-ray crystallographic structures) and those obtained from 0.5 μs of MD simulation [225](Fig. 3.2).

A shown in Fig. 3.3, one can observe that both C^α (backbone) and all-atom positional deviations are anharmonic in the long-timescale MD data ($\kappa(C^\alpha) = 6.3$; $\kappa(\text{all atom}) = 8.2$), though anharmonicity is observed even at shorter time-scales . In comparison to C^α atoms, side-chains cause greater anharmonicity in the positional deviations, thus indicating their ability to move more freely in the protein. Interestingly, the NMR ensemble shows similar statistical behavior, suggesting the pre-disposition of ubiquitin to undergo anaharmonic fluctuations at long-time scales. The X-ray ensemble shows similar anharmonic behavior qualitatively; however, the tail of the C^α distribution is insufficiently sampled to arrive at a statistically meaningful estimate of κ.

For the rest of this chapter and to maintain clarity, we use C^α atoms to study the anharmonic fluctuations. Using a Gaussian fit to the C^α positional deviations from MD simulations, we compute how often each C^α atom is found three standard deviations or more away from the mean of the approximating Gaussian distribution. Ubiquitin's flexible loop regions $\beta_1 - \beta_2, \beta_3 - \beta_4$ (collectively referred to as R1), $\beta_2 - \alpha_1$, $\beta_4 - \alpha_2$, and the C-terminal tip of α_1 (R2) of ubiquitin consistently exhibit significantly anharmonic fluctuations (Fig. 3.2; bottom right panel). Anharmonicity is also associated with regions $\beta_4 - \alpha_2$ and also $\beta 2 - \alpha_1$. However, these regions spend less than 10% of the time exhibiting anharmonic fluctuations. It is interesting to note that while R1 forms the primary binding region along which most substrates bind to ubiquitin, R2 forms the secondary binding region. Given that these flexible regions have a functional role in ubiquitin binding [170], their associated anharmonic distributions warrant close study.

To this end, we examine the kurtosis of the positional deviations projected onto a

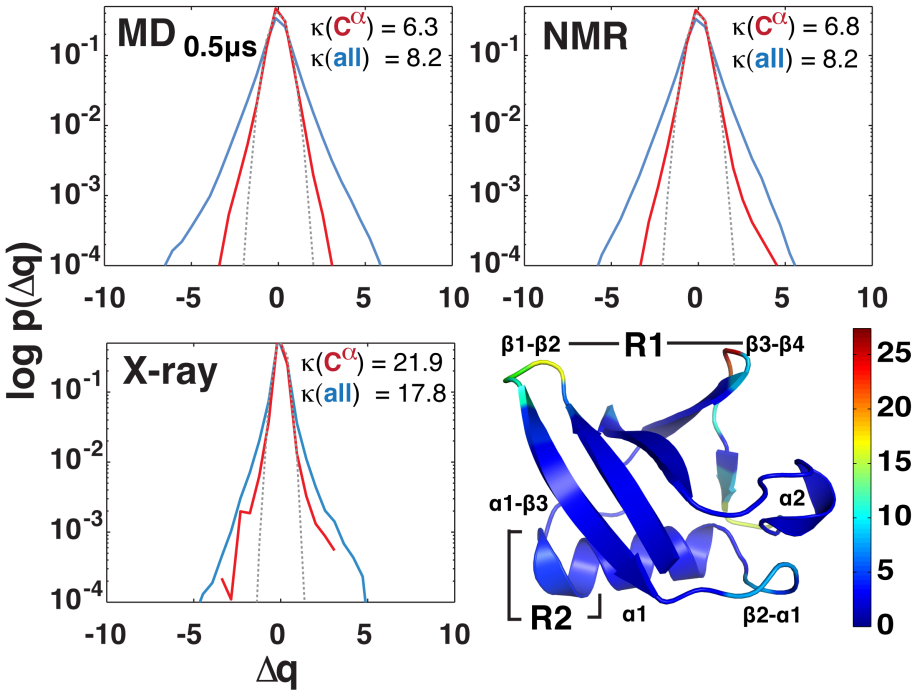

Figure 3.2: **Anharmonic distribution of positional deviations (Δq; units Å) in ubiquitin from 0.5 μs MD, NMR, and X-ray ensembles.** For each atom, the positional displacement from the time-averaged position was calculated at 50 ps intervals. The same bin size (0.54 Å) was used for all histograms. Dotted curve shows a Gaussian fit to the C^{α} distribution. Color map on the protein indicates the amount of simulation time spent (%) exhibiting anharmonic behavior. Note R1 represents the binding regions ($\beta_1 - \beta_2$, $\beta_3 - \beta_4$) and R2 represents $\alpha_1 - \beta_3$ regions in the protein. R1 and R2 represent primary and secondary binding interfaces respectively.

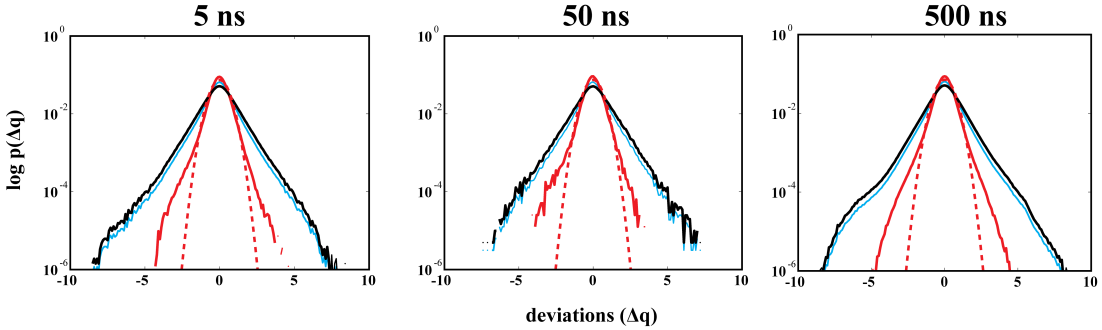

Figure 3.3: **Anharmonic distribution of positional deviations (Å) from MD simulations at 5 ns, 50 ns and 500 ns.** For each atom, the positional displacement from the time-averaged position was calculated at 50 ps intervals. The same bin size (0.54 Å) was used for all histograms. Distributions correspond to: C^{α} (red), Gaussian fit to C^{α} (dotted red), side-chains (light blue) and all-atoms (black). Side-chains cause greater anharmonicity than backbone motions. Observe that even at shorter time-scales there is considerable anharmonicity.

principal coordinate system built locally at each C^{α}. We observe that at least 41% of the C^{α} atoms are G^s, and 21% are G_s along all three principal components. It is further interesting to note that non-Gaussian (G_s and G^s) distributions are associated with R1 and R2, which are both involved in forming primary contacts with substrates [170]. Thus, atomic deviations at these functionally relevant protein regions are mixtures of G, G^s, and G_s distributions.

3.2.2 How do second-order statistics perform?

In the previous section, we have examined in detail the nature of anharmonic fluctuations ubiquitin can exhibit. One may also recall from Fig. 3.1, that κ can reveal characteristic 'shapes' or 'tails' in the distribution, even if the variance is almost similar. This ability to separate and quantify shapes of positional fluctuations can be seen as enriching the statistical content from MD simulations and will offer novel ways to characterize the conformational landscape. This will also primarily motivate why a computational model built using κ can be insightful.

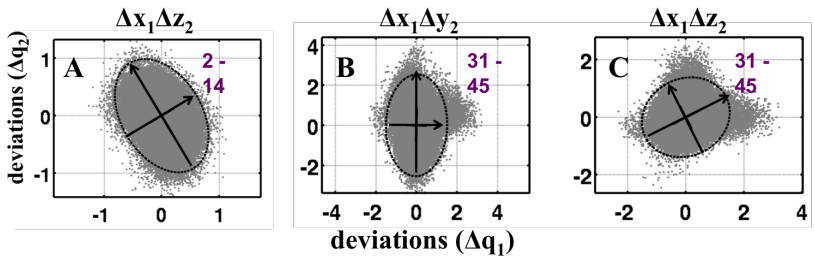

Figure 3.4: **Joint positional deviations of pairs of atoms and use of QHA to capture directions of dominant motions**. Residues 2 and 14 exhibit Gaussian-like fluctuations in the x and z directions respectively. When pairwise distributions are Gaussian like (panel A), QHA (black) basis vectors align well with the intrinsic orientation of the data. Residues 31 and 45 are anharmonic in (x, y) and (x, z) directions. When pairwise distributions are anharmonic the intrinsic orientation of the data can be non-orthogonal. QHA (black) does not recover this (panels B & C). All units are in Å.

Limitations of pursuing variance

As illustrated in Fig. 3.4A, examination of the joint positional fluctuations between two residues: 2 and 14, reveals that both residues have κ close to 3.0. It is evident that the joint positional fluctuations are also Gaussian (G) like - meaning that a QHA like model would capture the intrinsic directionality of fluctuations. The arrows in Fig. 3.4 illustrate the direction of the dominant PCA vector covering the variance where as the ellipse represents the shape of the motion covered by 2 standard deviations away from the mean positions of both residues. Observe that the eigenvectors determined from QHA align well with the intrinsic direction of the variance.

However, when the source distributions of residues are non-Gaussian (residues 31 and 45), the intrinsic orientations of the data are non-orthogonal, indicating that higher-order correlations exist within these distributions. Under these circumstances, QHA fails to capture the intrinsic motions in its blind pursuit of variance. As illustrated in Fig. 3.4B, QHA does not describe the fluctuations since the distributions are non-Gaussian. Further, when the distributions are as complex as the ones presented in Fig. 3.4C, the orientations of the vectors indicate clearly that the principal components do not align with the intrinsic orientations in the fluctuations. Therefore, for much higher dimensions, involving all of the $3N$ dimensions, QHA bases may not adequately capture the complex nature of positional deviations arising from mixtures of G, G_s, and G^s distributions. This is our first motivation towards using higher-order statistics to characterize the conformational landscape.

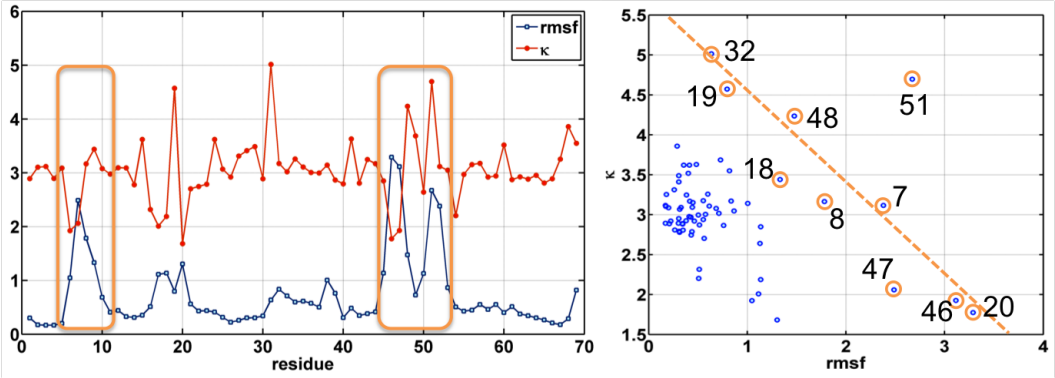

Figure 3.5: **RMSF is not correlated to κ.** A side-by-side comparison of RMSF with κ indicates that there is not much correlation between the two. But, as illustrated in the left hand panel, the regions from R1 and R2 show distinct behavior in the RMSF versus κ behavior. See the text for more explanation.

Are RMSF and κ correlated?

We consider this question to rationalize further the use of higher-order statistics is necessary to explain the conformational landscape of a protein. For ubiquitin, we observe (as illustrated in Fig. 3.5), the RMSF is not a good indicator of κ; indeed, the correlation between κ and RMSF is very low: $cc = 0.22$. This implies that the RMSF as a metric is not a good predictor for κ. However, if we turn our attention to the functionally relevant regions (R1 and R2) of ubiquitin, we do observe an intriguing behavior. Residues that line R1 (residues 7-11; 45-48) show large RMSF whereas show distinct G_s behavior; residues that line R2 (31-33), on the other hand, show small RMSF with G^s behavior. The $\beta_4 - \alpha_2$ (51-54) loops show the presence of substantial RMSF values and G^s behavior. The highlighted residues in the plot fall along a straight line showing negative correlation with RMSF, indicating that residues with a functional role in the protein tend to visit multiple sub-states and hence the RMSF values are negatively correlated to κ.

This observation offers the second motivation to characterize the anharmonic behavior of the protein in terms of the intrinsic orientations within the conformational landscape, rather than just pursue the variance. While the variance provides information regarding the extent (or spread) of the fluctuations, it does not explain if different atoms in the protein have a "predisposition" for movements in certain "wells". The notion of directionality in inter-atomic fluctuations is a strong reason for understanding the relevance for protein function.

3.3 Quasi-anharmonic analysis

We will build a *linear* computational model that can account for the intrinsic orientations in the protein's conformational landscape by pursuing higher-order statistics. We will then illustrate that by pursuing higher-order statistics, it is possible to obtain a natural description of the ubiquitin conformational landscape (see Sec. 3.4).

3.3.1 Extracting Anharmonic Modes of Motion

It is instructive to consider QHA, where the positional deviations $\vec{x}$ are modeled as a linear combination of harmonic sources $\vec{\alpha}$, given by:

$$\vec{x} = B\vec{\alpha}. \tag{3.2}$$

The harmonic modes B are conveniently expressed by the eigenvalues Σ and eigenvectors U of the covariance matrix given by:

$$C = E\{\vec{x}\vec{x}^T\} = U\Sigma U^T. \tag{3.3}$$

B is then set to $B = U\Sigma^{1/2}$. However, the covariance matrix C captures only second order correlations in atomic fluctuations and requires that the basis vectors to be orthogonal.

We propose quasi-anharmonic analysis (QAA), a linear model based on independent component analysis [37, 54, 16]. We model the observable positional deviation vector, $\vec{x}$, as a linear combination of anharmonic sources, $\vec{\gamma}$, such that:

$$\vec{x} = A\vec{\gamma}. \tag{3.4}$$

Here, A is an unknown coupling matrix where each column A_i encodes an anharmonic mode of motion describing the intrinsic higher-order correlations between different regions of the protein. To account for higher-order statistics, specifically fourth order correlations, we will estimate a fourth order cumulant tensor for the atomic fluctuations from which the anharmonic bases A will be derived. The excitation of the anharmonic modes can be quantified as:

$$\vec{\gamma} = A^{-1}\vec{x}. \tag{3.5}$$

Unlike in QHA, the basis matrix A can be non-orthogonal and hence the anharmonic modes can be intrinsically coupled. It is important to estimate both A and $\vec{\gamma}$ to suitably describe the anharmonic landsacpe.

We term this analysis *quasi-anharmonic* for two reasons: first, we study anharmonicity explicity whereby the sources are fully decorrelated and made as independent as possible; second, we impose a linear model which ignores any non-linear coupling that may exist in the fluctuations between parts of a protein.

To derive A The fourth order cumulant tensor $\mathcal{K}$, comprises of auto and cross-cumulants given by:

$$\begin{aligned} k(x_i) &= E\{x_i^4\} - 3E^2\{x_i^2\} \\ k(x_i, x_j, x_k, x_l) &= E\{x_i, x_j, x_k, x_l\} - E\{x_i, x_j\}E\{x_k, x_l\} \\ &\quad - E\{x_i, x_k\}E\{x_j, x_l\} - E\{x_i, x_l\}E\{x_k, x_j\}. \end{aligned} \tag{3.6}$$

To simplify the computation of $\mathcal{K}$, first we assume that the overall rotation/translation degrees of freedom have been removed and hence the positional fluctuations are centered around the origin. Second, the fluctuations are projected onto the QHA eigenbases and then normalized to have unit variance using:

$$\vec{y} = \Sigma^{-1/2} U^T \vec{x}. \tag{3.7}$$

The cumulant tensor, $\mathcal{K}$, is now built in the $\vec{y}$ space, where the normalization implies $E\{y_i y_j\} = 1$ when $i = j$ and 0 when $i \neq j$. $\mathcal{K}$ is now built in the $\vec{y}$ space, where the normalization implies $E\{y_i y_j\} = 1$ when $i = j$ and 0 when $i \neq j$.

$\mathcal{K}$ will have a total $3N \times (3N+1)/2$ matrices each of size $3N \times 3N$ accounting for auto- and cross-cumulant terms, where N is the total number of residues in the protein. These matrices can then be approximately diagonalized such that the sum of squares of cross-cumulant terms is minimized using efficient algebraic techniques like Jacobi rotations [104] to obtain a new rotation matrix D of size $3N \times 3N$. A public domain implementation of the joint diagonalization is available in [54]. Using the new rotation matrix D we can write:

$$\vec{\gamma} = D\vec{y} \tag{3.8}$$

and substituting for $\vec{y}$ from above (Eq. 3.7):

$$\vec{\gamma} = D\Sigma^{-1/2} U^T \vec{x}. \tag{3.9}$$

Thus, we obtain an expression for the anharmonic modes of motion, $\vec{\gamma}$ as:

$$\vec{\gamma} = A^{-1}\vec{x}, \tag{3.10}$$

implying that the anharmonic modes of motion A is just:

$$A = U\Sigma^{1/2} D^T. \tag{3.11}$$

The anharmonic modes of motion A_i, which are the columns of matrix A, are sorted in decreasing order of their amplitudes ($\|A_i\|$).

3.3.2 Relevant Work

Note, to estimate A, we used fourth-order statistics. However, A can be estimated using several other choices. Ichiye and Karplus [136], modeled $p(\vec{x})$ as nearly Gaussian and described $p(\vec{\gamma}) \propto p(A^{-1}\vec{x})$ with an Edgeworth expansion. However, they modeled these higher-order cumulants of positional fluctuations locally, but did not describe the conformational landscape globally. Indeed, the time-scales accessible to them was of the order of picoseconds. Given that MD simulations are able to scale regularly to μs time-scales, the current representation (QAA) provides a significant advantage in understanding internal dynamics of a protein.

A second and attractive choice to learn A is to maximize mutual information [67] $I(\vec{\gamma};\vec{x})$ between $\vec{\gamma}$ and $\vec{x}$. Lange and Grubmuller [168] use this approach to include both linear and non-linear correlations in the MD data and build a model of the conformational landscape. It must be noted that, while mutual information is an excellent metric estimating it accurately is quite difficult. In a MD simulation, given sampling inaccuracies and high dimensionality of the input data, estimating mutual information can be an extremely arduous task. Secondly, their choice of an orthogonal representation of the landscape is not quite justified, especially when we observe from Figs. 3.2 and 3.4 clearly that the intrinsic orientations in the positional deviations are non-orthogonal.

In this section, we have briefly summarized how one can model the anharmonic fluctuations by taking into account the higher-order statistics, particularly κ, or kurtosis. In the next section, we will examine how QAA characterizes the landscape spanned by a 0.5μs equilibrium simulation of ubiquitin [225].

3.4 Anharmonic Conformational Landscape of Ubiquitin

3.4.1 Anharmonicity in two dimensions

Before we examine the entire conformational landscape, we need to examine if QAA works well in two dimensions. We consider the same pairs of residues examined in Fig. 3.4. First, we note that when the fluctuations are Gaussian-like (Fig. 3.6A), the intrinsic orientations as determined by QAA align well with the QHA basis vectors. When the fluctuations are typically non-Gaussian, involving G^s sources, the intrinsic orientations determined from QAA align themselves along the directions of the 'arms' as observed from the figure (Fig. 3.6B and Fig. 3.6C).

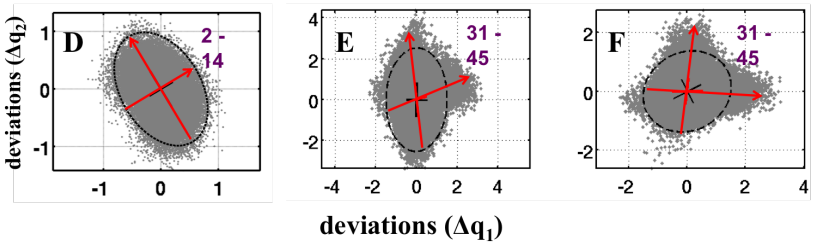

Figure 3.6: **Joint positional deviations of pairs of atoms and use of QAA to capture directions of dominant motions**. Residues 2 and 14 exhibit Gaussian-like fluctuations in the x and z directions respectively. When pairwise distributions are Gaussian like (panel A from Fig. 3.4; panel D), QAA (red) basis vectors align well with the intrinsic orientation of the data. Residues 31 and 45 are anharmonic in (x, y) and (x, z) directions. When pairwise distributions are anharmonic the intrinsic orientation of the data can be non-orthogonal. QAA effectively captures this behavior, since the basis vectors align themselves with the intrinsic orientation in the data (panels B and C from Fig. 3.4; panels E and F). All units are in Å.

3.4.2 Anharmonic Modes of motion in ubiquitin

We considered the C^{α} atoms for residues 2-70 ($N = 69$) and sampled 10,000 conformations spread evenly over $0.5\ \mu s$ MD. The $3N$ dimensional space was first projected onto the top 30 QHA dimensions (covering 96% of the overall variance). The projection onto this sub-space mitigates the effects of fast fluctuations (noise) and provides a more robust sub-space tractable for the convergence of QAA. Projecting the 30 QHA dimensions onto the top three anharmonic modes (γ_i), as shown in Fig. 3.7, we observe that the landscape separates into unique conformational wells. Using a mixture-of-Gaussian (MoG) [195] model, we identify four conformational wells (labeled I through IV) with their boundaries marked by ellipses drawn 3 standard deviations (σ) from the respective cluster centers.

The mean structures from each well reveal novel features of ubiquitin's plasticty, i.e., its ability to sample a wide range of conformations even at equilibrium. In cluster I, shown in blue (Figs. 3.7B & 3.7C) and consisting of over 8,000 structures, ubiquitin adopts a conformation whereby region R1 is constrained (13.6 Å), whereas $\beta_1 - \beta_2$ and R2 are far apart (11.5 Å). Observe that a majority of the NMR ensemble (43 conformers within 2σ and 78 within 3σ) and the X-ray ensemble (42 within 2σ and 44 within 3σ) fall within cluster I, indicating that MD sampling has indeed visited all of the bound/ unbound conformers observed in this three-dimensional space.

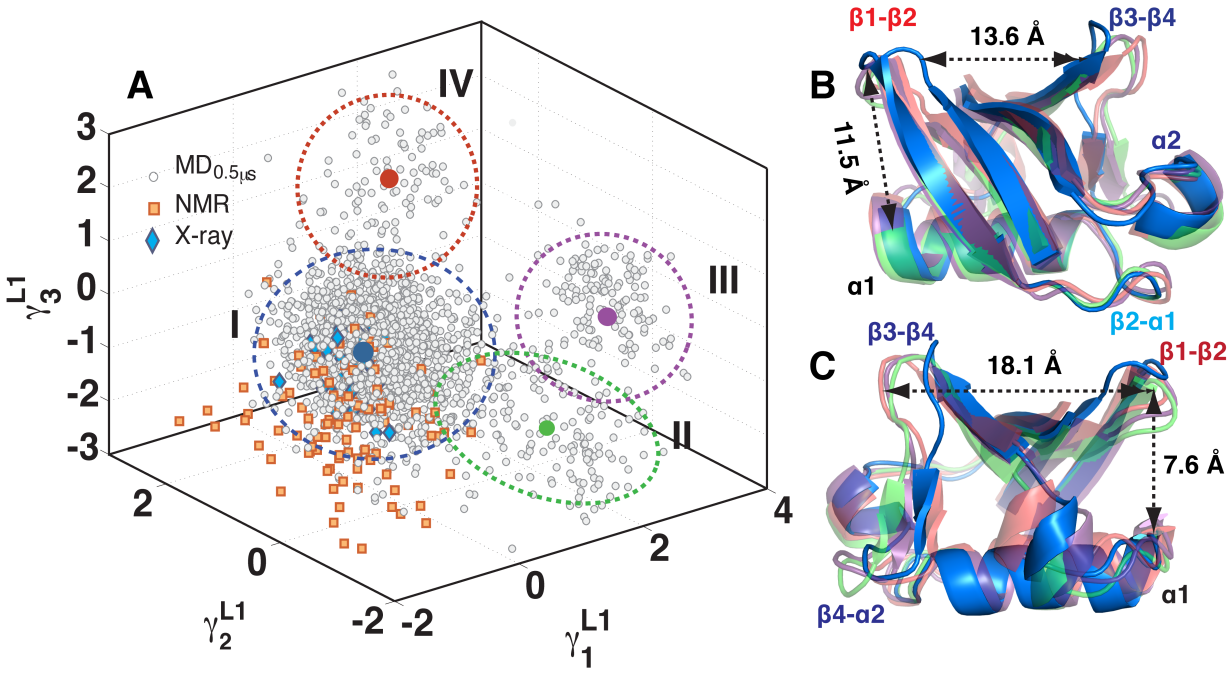

Figure 3.7: **Quasi-anharmonic analysis (QAA) of ubiquitin conformational landscape reveals conformational sub-states.** (A) The MD ensemble projected onto the top three anharmonic modes of motion. The projection (units Å) shows four distinct clusters (I-IV). The cluster centers are shown in blue (7,880 conformers; I) green (773; II), purple (692; III) and red (655; IV). (B) and (C) Two different view-points (rotated around y-axis by 180 °) of the mean conformations from each cluster (bold circles in A) show significant structural deviations in R1 and R2.

QAA reveals three other clusters (shown in purple, green and red respectively in Fig. 3.7A). They form the 'wings' of cluster I, exhibiting motions along $\beta_3 - \beta_4$ and $\beta_2 - \alpha_1$ regions, indicating motions that are complementary to R1 and R2 (Fig. 3.7B and Fig. 3.7C). In cluster IV, the mean structure shows an open conformation where region R1 is extended over 18 Å and R2 is close to $\beta_1 - \beta_2$ at 7.6 Å. Note that motions in both R1 and R2 are implicated in binding diverse substrates [170, 225].

We next examine if these clusters exhibit any similarity in terms of their internal energies, defined as the sum of van der Waals and electrostatic energy over all interactions (Eq. 1.5) in the protein and computed using the program MDEnergy [220]. We plot the scaled internal energy values on the data in Fig. 3.7 and illustrate it in Fig. 3.8 (Level 1). While cluster I shows considerable diversity in its internal energies, clusters II, III and IV are homogeneous. Note that clusters I and III are separated by high-energy structures possibly indicating a transition state between the two conformational wells. QAA is therefore able to succinctly capture the intrinsic fluctuations both in ubiquitin's motions and energetics.

3.4.3 Biological Perspective: Conformational sub-states in ubiquitin landscape

The largest sub-state (cluster I) is highly diverse with respect to its internal energy distributions and positional deviations (Fig. 3.8). Thus, we can examine the conformational diversity by performing QAA on this cluster to see if a subsequent decomposition might homogenize this landscape. Fig. 3.8 (Level 2) reveals that cluster I separates into 3 sub-states having unique structural and energetic properties. The largest structural deviations involve the motions of R1 (Fig. 3.2). The largest sub-state in level 2 comprises more than 6,000 conformations, and the internal energy distribution in this cluster is quite diverse (see Fig. 3.8; Level 2). Hence we use QAA to descend one more level in the conformational landscape. At Level 3 of QAA, we observe that the landscape splits into three sub-states. Compared to the preceding levels, the amplitude of motions in ubiquitin are less pronounced and this is indicative of higher conformational homogeneity.

This successive homogenization in terms of both positional fluctuations and energetics provides an intuitive understanding of the motions involved in ubiquitin binding, as illustrated on top of each panel in Fig. 3.8. At Level 1, the fluctuations are global involving the pincer regions: $\beta_1 - \beta_2$ (red), $\beta_2 - \alpha_1$ (cyan; R1), C-terminal tip of α_1 (R2; orange) and $\beta_3 - \beta_4 - \alpha_2$ (blue) regions. At Level 2 the motions become localized to the protein binding loops: R1 albeit with lower amplitudes (see movies in supporting information). At

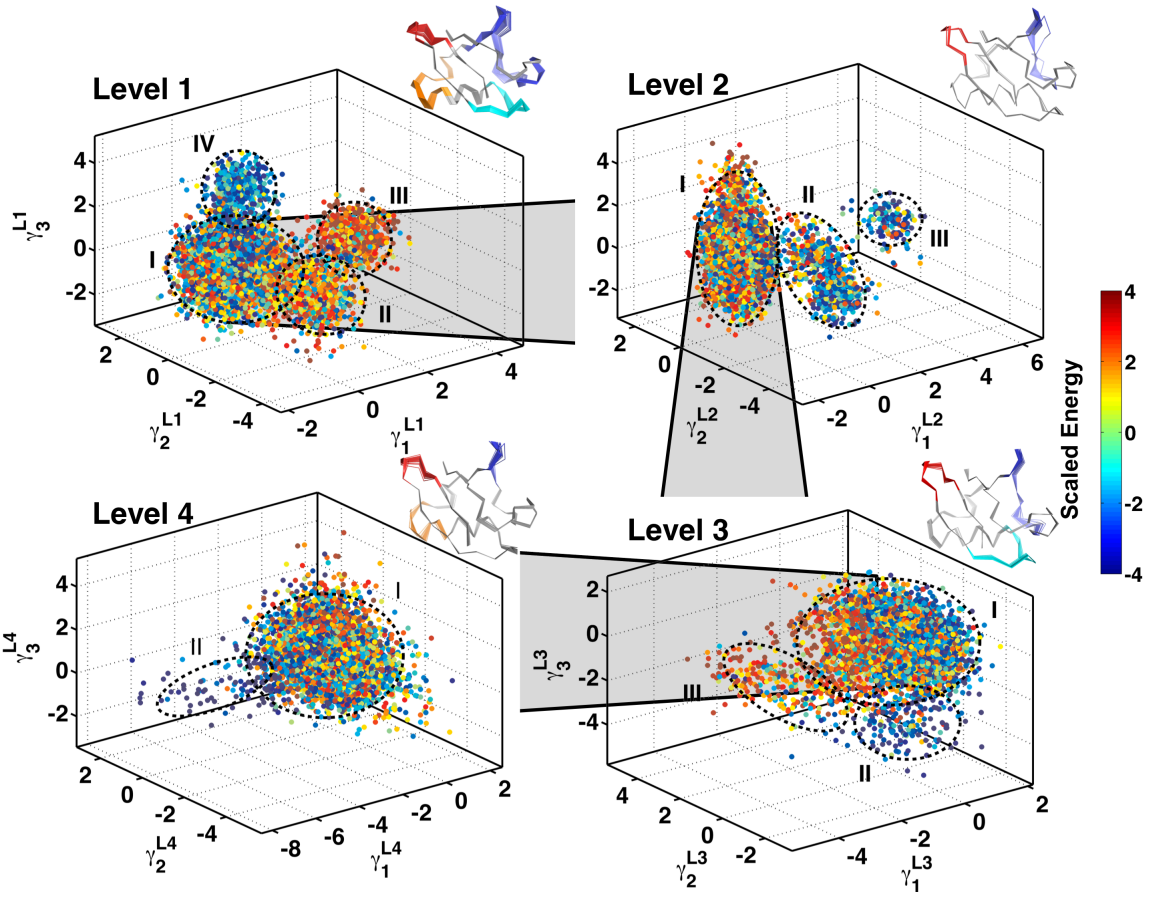

Figure 3.8: **Hierarchical organization of conformational sub-states in ubiquitin motions.** Level 1 decomposition identifies four sub-states (clusters). Each conformation is colored using the scaled internal energy. Levels 2, 3 and 4 are derived from the largest sub-state of the preceding level indicating more homogeneity in both positional deviations and internal energy. Motions along the top anharmonic mode are illustrated in each panel in a movie like representation.

Level 3 $\beta_2 - \alpha_1$ is coupled to R1 and at Level 4, R2 is coupled to R1. At a global level the motions of R1 (primary binding regions) dominate while at subsequent levels the motions in secondary binding regions ($\beta_2 - \alpha_1$ and R2) dominate. Thus, QAA reveals spatially localized anharmonic motions at successive levels of the conformational hierarchy.

Such progressive decomposition of the landscape combined with the homogeneity in energy space provides new insights into the binding motions of ubiquitin. First, observe that the top 3 anharmonic modes of motion can cover all of the conformational heterogeneity exhibited by the bound X-ray ensemble (Fig. 3.7; blue diamonds). Second, the hierarchy of motions in ubiquitin allow the protein to sample conformations that involve modulating the pincer regions (R1 and R2) to varying degrees. The subtle interplay between global conformational fluctuations (Level 1 motions) as well as its ability to modulate local motions (levels 2 through 4) can thus enhance ubiquitin's ability to target multiple substrates [170].

3.5 Conclusions

In summary, we find that the equilibrium motions of ubiquitin exhibit significant higher-order correlations in both individual and collective fluctuations (Fig. 3.2). The sub-state decomposition revealed a natural hierarchy of fluctuations that are important for ubiquitin in binding to diverse substrates. Chasing anharmonic fluctuations, QAA revealed the presence of conformational sub-states with different internal energies that are homogeneous within and heterogenous between sub-states. The unique structural features identified by QAA elucidate the mechanism of binding motions in ubiquitin (Fig. 3.7). A hierarchical decription of the sub-states allows the mapping of localized motions to the global fluctuations and provides insights into how ubiquitin modulates these motions to achieve effective binding (Fig. 3.8).

Chapter 4

An Online Approach to Mine Collective Motions from Molecular Dynamics Simulations

Collective behavior involving distally separate regions in a protein is known to widely affect its function. In this chapter, we present an online approach to study and characterize collective behavior in proteins as molecular dynamics simulations progress. Our representation of MD simulations as a stream of continuously evolving data allows us to succinctly capture spatial and temporal dependencies that may exist and analyze them efficiently using data mining techniques. By using tensor analysis we identify collective motions (i.e., dynamic couplings) and time-points during the simulation where the collective motions suddenly change. We demonstrate the applicability of this method on simulations for ubiquitin and barnase. We characterize the collective motions in these proteins using our method and analyze sudden changes in these motions. Taken together, our results indicate that tensor analysis is well suited to extracting information from molecular dynamics trajectories in an online fashion [1].

[1]This chapter is adapted from: Arvind Ramanathan, Pratul K. Agarwal, Maria G. Kurnikova, Christopher J. Langmead, An online approach to mine collective behavior from molecular dynamics simulations, J. Comp. Bio. (2010), 17:3 (in press).

4.1 Introduction

Molecular Dynamics (MD)/ Monte-Carlo (MC) simulations are perhaps the most commonly used computational techniques to simulate proteins (and bio-molecules in general) [235]. These methods sample from a complex energy landscape for the protein using either an all-atom or a coarse grained representation [18] of the protein. Solvent conditions, presence of ions and other physiologically relevant conditions, which may closely resemble *in vivo* conditions of the cell, can also be modeled using these techniques. Statistical sampling techniques can provide insights into rare events that may also play an important role in biologically relevant motions [81]. However, these methods can be computationally expensive and cannot sample enough of the energy landscape to observe functional motions at relevant time-scales (milliseconds-seconds) [64]. Apart from having high noise content (due to thermal fluctuations) [46], visualization and interpretation of collective fluctuations from simulation data in such high dimensions can be cumbersome.

Recent advances in MD/MC simulation methods have dramatically increased the time-scales accessible to examination. For example, Folding@Home [213] provides a unique resource to simulate protein folding landscape up to several microseconds [83]. Similarly, highly scalable MD codes are being developed such as NAMD [220] and Desmond [50] to run on both commercial clusters as well as super-computers [5] leading to a substantial increase in the time-scales achievable for conventional MD simulations [95]. Further, the development of customized hardware for MD simulations through Anton [242] and field-programmable gate arrays (FPGA) [11] have allowed all-atom simulations to scale to reach millisecond time-scales. This concomitant increase in the size of the resulting data sets, which can easily exceed several terabytes, presents new computational challenges. Beyond the practical issues of storage and retrieval, the scientific challenge of extracting information and detecting patterns of interest to the end-user from such massive data sets requires the development of new tools.

Most existing online methods for monitoring molecular dynamics trajectories track one or a handful of coarse-grained structural features, such as root-mean squared deviations (RMSD), radius of gyration, and secondary structure content. While useful, these metrics do not necessarily provide any information regarding how groups of atoms move in a co-ordinated fashion. In contrast, one would need a way of detecting and tracking fine-grained spatio-temporal patterns, including the *collective* (i.e., correlated or coordinated) motions of even widely separated regions [43]. Previously, such collective behaviors have been examined using techniques such as NMA applied to static structures [51, 25] or PCA (and its variants) applied in an offline fashion to MD trajectories [150, 15, 119, 118].

In the previous chapters, we have used QHA and QAA to analyze the protein's con-

formational landscape and characterize collective fluctuations. QHA [150] uses multi-dimensional Gaussian distributions to model the landscape and pursues variance to describe intrinsic fluctuations of the protein. QAA pursues higher-order correlations and analyzes the conformational landscape arising from a consequence of the intrinsic orientations in the fluctuations of constituent atoms. However, it must be pointed out that both QHA and QAA were applied *post-simulation*. This has some important consequences: (a) it is not possible to understand how collective behavior may change over time as simulations are progressing and (b) it is also not possible to relate how different groups of atoms or interactions (such as C^{α} atoms, side-chains, hydrogen bonds, hydrophobic interactions, etc.) collectively change as a function of time or if there are non-trivial correlations that exist as simulations are progressing.

In this chapter, we introduce a novel technique for automatically detecting spatio-temporal patterns and for detecting rare events and dynamic anomalies [239] in molecular dynamics trajectories. Our online approach enables one to monitor the emergence and dissolution of such collective behaviors through time, providing new insights into the relationship between dynamics and function. Moreover, a user-specified parameter selects the characteristic timescale of the detectable patterns, providing a framework for comparing and contrasting patterns at different scales. Finally, our method is based on a generalization of PCA to multi-dimensional (i.e., tensor) data, and is therefore capable of detecting higher-order patterns. One may thus summarize the contributions from this chapter as: (a) introduction of a novel representation of molecular dynamics trajectories as tensors, (b) the first application of an online algorithm for tensor analysis to MD data, and (c) the identification and analysis of collective motions in two proteins, ubiquitin and barnase. Taken together, our results indicate that it is possible to obtain biologically significant insights while tracking simulations online.

4.2 Tensor Representations of MD simulations

Tensors are an extension of matrices beyond two dimensions and provide a convenient way to capture multiple dependencies that may exist in the underlying data. Formally, a tensor $\mathcal{X}$ of M dimensions can be defined as a multi-dimensional array of real values,

$$\mathcal{X} \in \Re^{N_1 \times N_2 \times \ldots \times N_M}, \tag{4.1}$$

where N_i represents the i^{th} dimension for $(1 \leq i \leq M)$. Many operations on matrices can be extended to tensors. For example, in this chapter we utilize a generalization of principal

components analysis (see Sec. 4.3). A discussion of tensor representations and operations is presented in appendix. In what follows, tensors are represented with calligraphic letters (e.g. $\mathcal{X}$), matrices by bold capital letters (e.g. $\mathbf{X}$), vectors as bold small letters (e.g. $\mathbf{x}$) and scalars as normal text (e.g. x).

Our primary goal is to study the evolution of the collective behaviors of a protein's constituent atoms. We will do this by using a tensor to encode the dynamics of the relationships between atoms. There are a number of relationships one might wish to monitor, and each leads to a different kind of tensor. We will consider two representative encodings: the first is a dense encoding (one with few zero elements) exhibiting symmetries. The second representation is a sparse (one with few non-zero elements), asymmetric encoding. The choice of representation ultimately depends on the kinds of collective motions one wishes to identify.

Distance Map Tensors. The first kind of tensor encodes a molecular dynamics trajectory as a sequence of interatomic *distance maps*. As its name implies, a distance map is a symmetric matrix (or, equivalently, a second order tensor) encoding the pairwise Euclidean distances between atom. Thus,

$$\mathbf{X}_{ij} = ||\mathbf{r}_i - \mathbf{r}_j|| = \sqrt{\sum_{k=1}^{3} (\mathbf{r}_i(k) - \mathbf{r}_j(k))^2} \tag{4.2}$$

where $\mathbf{r}_i$ is the Cartesian position vector of atom i. An example distance map for the protein cyclophilin A (pdb id: 1AWQ) is shown in Fig. 4.1(A). We will consider distance maps between C^{α} atoms, and thus the total number of entries in $\mathbf{X}$ is n^2, where n is the number of residues in the protein. A molecular dynamics trajectory is essentially a sequence of t snapshots of atomic coordinates. By computing the distance map for each snapshot, we can encode the trajectory using a third order tensor with dimensions $n \times n \times t$.

We note that a distance map is isomorphic to a set of coordinates and it is straight forward to compute either mapping [117]. In comparison to a set of coordinates, a distance map has two representational advantages. First, it explicitly reveals both local and global information about the geometry of the system. We will use this property to examine both local and global collective motions in the protein. Second, a distance map is invariant to translations and rotations of the molecule, which is convenient when examining internal motions. At the same time, a distance map has a representational disadvantage because it has higher space complexity than a set of coordinates.

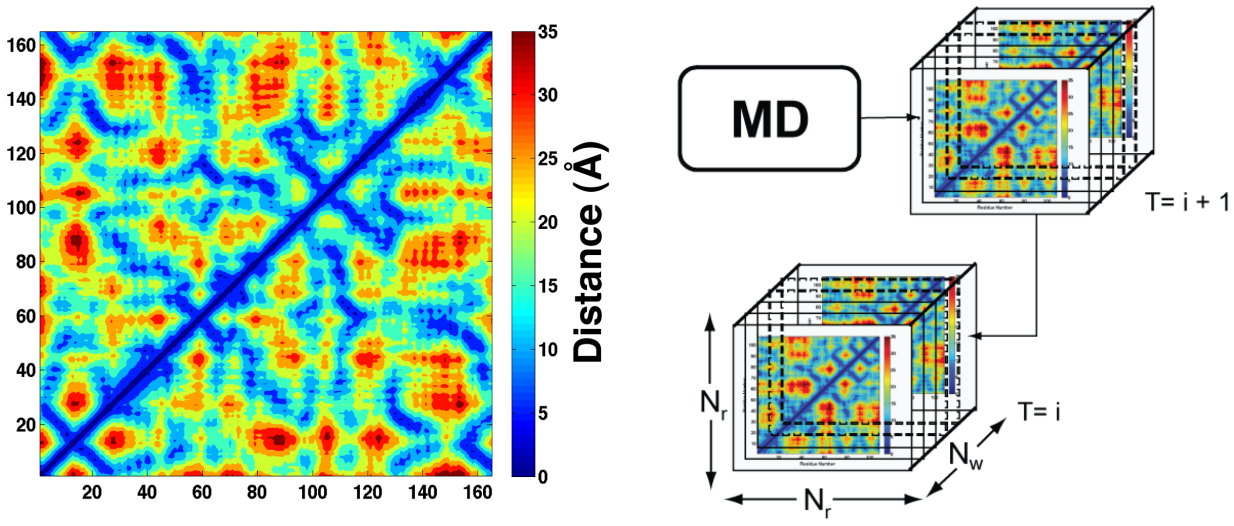

Figure 4.1: **Distance matrix and Tensor Representation of MD simulations**. (A) Distance map representation of the enzyme cyclophilin A (pdb: 1AWQ). Distant residues are identified using darker shades of red. (B) The tensor streaming representation of MD trajectories used in this paper. As new tensors keep arriving at every time interval $T = i + 1$, they are appended to the end of the current stream $T = i$. n is the number of residues, w represents the size of the window. This can be set by the end user depending on how often the user wants the analysis to run.

Hydrogen Bond Network Tensors. Our second kind of tensor encodes the dynamics of hydrogen bonds. Each hydrogen bond has a donor and an acceptor. A hydrogen bond can exist if a donor and acceptor are within about 3 Å. Thus, we can represent the set of hydrogen bonds present in a given MD snapshot with an $d \times a$ matrix where d is the number of hydrogen bond donors and a is the number hydrogen bond acceptors. In this matrix, element (i, j) is the distance between donor i and acceptor j if the two atoms are with 3 Å, and zero otherwise. Unlike a distance map, this matrix is sparse and asymmetric. By constructing this matrix for each snapshot in MD trajectory, we can encode that trajectory as a third order tensor with dimensions $d \times a \times t$.

Windowing. An alternative to constructing a single tensor encoding the entire MD trajectory is to partition the data into a collection of (overlapping or non-overlapping) windows of size w, where $w < t$. Here, each window is represented using a third order tensor

with dimensions $n \times n \times w$ (resp. $d \times a \times w$). An ordered sequence of such windows is called a *tensor stream*. Formally a tensor stream is a discrete collection of M^{th} order tensors $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T$, where $T = \lceil t/w \rceil$. A tensor stream representation used for MD simulation is illustrated in Fig. 4.1(B). Each tensor represents a discrete time window within the simulation. As the MD simulation progresses, more and more tensors become available, and the new tensors are appended to the end of the tensor stream.

There are three basic advantages associated with windowing. First, by adjusting the width of the window, we can selectively focus on collective motions occurring at specific time-scales. Molecules exhibit dynamics simultaneously at multiple time-scales. Our method enables scientists to identify patterns at specific time-scales and, if desired, compare and contrast the patterns occurring at different time scales by varying the size of the window. Second, windowing enables *online* analysis of trajectories. This is significant because MD simulations typically require days to months of computation, even using specialized hardware, and the specific events of interest (e.g., transitions between meta-stable states) may be rare. Using our method, a scientist can identify such events in an online fashion and fork off a new simulation to investigate the phenomenon in further detail (e.g., perturbation studies). Finally, windowing provides a practical advantage because the space and computational complexity of the tensor operations are, of course, proportional to the size of the tensor.

Before for proceeding to a discussion of our method, we note that the two kinds of tensors we defined in this section are simply two of many potentially interesting tensor streams worth analyzing. One can imagine, for example, constructing tensor streams over a number of structurally relevant features, including: hydrophobic interactions, force vectors, and electrostatic maps. Of course, it is also possible to construct tensors of higher dimension as needed.

4.3 Tensor Analysis of Molecular Dynamics Trajectories

In this section we present our method for analyzing molecular dynamics trajectories. We begin with a brief review of some concepts from principal components analysis (PCA) because our method relies on a generalization of PCA to tensor data. That generalization is commonly called *tensor analysis*. Next, we present an algorithm for performing tensor analysis in an online fashion, followed by discussion of several ways to extract biologically relevant information from tensor analyses of molecular dynamics trajectories. We conclude this section with a brief summary of related work.

4.3.1 Background

In two dimensions, PCA is a popular method for unsupervised pattern discovery and dimensionality reduction. Let $\mathbf{X}$ be an $m \times n$ data matrix with zero empirical mean representing n data points in an m dimensional space. PCA is a linear transformation that projects the data into a new coordinate system such that the direction with the greatest variance is the first coordinate (aka component), the direction with the second greatest variance is on the second coordinate, and so forth. Mathematically, this transformation can be computed by performing the singular value decomposition on $\mathbf{X}^T$ or, equivalently, by computing the eigenvectors and eigenvalues of the empirical covariance matrix $cov(\mathbf{X})$. In practice, the change of basis can reveal underlying patterns in the data. PCA is often used to perform dimensionality reduction. Here the data are projected onto a lower dimensional subspace by dropping low-variance dimensions. Such linear projections are optimal in the least-squares sense.

Tensor Analysis. As previously mentioned, tensor analysis generalizes PCA. It too can be used to perform pattern discovery and dimensionality reduction. In this chapter, we focus on pattern discovery. Formally, given a collection of tensors $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T$, each of dimension $n_1 \times n_2 \times \ldots \times n_M$, tensor analysis will find orthogonal matrices $\mathbf{U}_i$, one for each dimension, that minimizes the error of reconstruction, e, which is defined as follows:

$$e = \sum_{t=0}^{T} \|\mathcal{X}_t - \mathcal{X}_t \prod_{i=1}^{M} \times_i (\mathbf{U}_i \mathbf{U}_i^T)\|_F^2 \tag{4.3}$$

Here, $\|\mathcal{X}\|_F^2$ is the square of the *Frobenius norm* of tensor $\mathcal{X}$, which is defined as:

$$\|\mathcal{X}\|_F^2 = \sum_{i_1=1}^{n_1} \ldots \sum_{i_M=1}^{n_M} \mathcal{X}(i_1, \ldots i_M)^2, \tag{4.4}$$

and is equivalent to the sum of inner product operation in matrices.

Informally, Eq. 4.3 is the difference between the actual data, $\mathcal{X}_t$, and the approximation of $\mathcal{X}_t$ in the space spanned by orthogonal matrices $\mathbf{U}_i$, denoted by $\mathcal{X}_t \prod_{i=1}^{M} \times_i (\mathbf{U}_i \mathbf{U}_i^T)$. Here, $\mathcal{Y}_t = \mathcal{X}_t \prod_{i=1}^{M} \times_i \mathbf{U}_i$ is called the *core tensor*. The tensor-matrix multiplication operator, $\mathcal{X}_t \prod_{i=1}^{M} \times_i \mathbf{U}_i$, is defined as:

$$\mathcal{X} \prod_{i=1}^{M} \times_i \mathbf{U}_i = \mathcal{X} \times_1 \mathbf{U}_1 \ldots \times_M \mathbf{U}_M \tag{4.5}$$

Note that the eigenvector matrices $\mathbf{U}_i$ reveal any underlying patterns that may be present within the data.

4.3.2 Algorithms

Having defined tensor analysis as finding the orthogonal matrices that minimizes Eq. 4.3, we briefly discuss an offline algorithm for performing that task, and then present an online version. First, we note that unlike PCA, there is no closed form solution for tensor analysis. In an offline setting, several iterative algorithms exist for performing tensor analysis [165, 120, 284, 252]. The algorithm by [252], for example, unfolds the tensor along each dimension (see below), and then performs regular PCA to compute the projection matrix for that dimension.

Offline tensor analysis requires all tensors available before the algorithm can be applied. However, the size of molecular dynamics trajectories (which is unbounded, in principle) ultimately limits the utility of offline tensor analysis. Instead, we consider an online algorithm for tensor analysis, called *Dynamic Tensor Analysis* (DTA) [252], which has been used in the data-mining community to extract patterns from time-evolving data, although not molecular dynamics trajectories. The DTA algorithm is shown in Algorithm 1 and shown schematically in Fig. 4.2.

Algorithm 1 Dynamic Tensor Analysis

Input:
New tensor $\mathcal{X}^{(t)} \in \Re^{n_1 \times \ldots \times n_M}$;
Eigenvector matrices from time $(t-1)$, $\mathbf{U}_i^{(t-1)}|_{i=1}^{M} \in \Re^{n_i \times n_i}$;
Eigenvalue matrices from time $(t-1)$, $\mathbf{S}_i^{(t-1)}|_{i=1}^{M} \in \Re^{n_i \times n_i}$
Output:
Updated eigenvector matrices at t, $\mathbf{U}_i^{(t)}|_{i=1}^{M}$;
Updated eigenvalue matrices at t, $\mathbf{S}_i^{(t)}|_{i=1}^{M}$;
Core tensor $\mathcal{Y} \in \Re^{n_1 \times \ldots \times n_M}$
for $d = 1$ to M **do**
    Unfold $\mathcal{X}^{(t)}$ as $\mathbf{X}_d^{(t)} \in \Re^{(\prod_{i \neq d} n_i) \times n_d}$
    Reconstruct variance $\mathbf{C}_d^{(t-1)} \leftarrow \mathbf{U}_d^{(t-1)} \mathbf{S}_d^{(t-1)} (\mathbf{U}_d^{(t-1)})^T$
    Update variance matrix $\mathbf{C}_d^{(t)} \leftarrow \lambda \mathbf{C}_d^{(t-1)} + (\mathbf{X}_d^{(t)})^T \mathbf{X}_{(i)}^{(t)}$
    Diagonalize $\mathbf{C}_d^{(t)} = \mathbf{U}_d^{(t)} \mathbf{S}_d^{(t)} (\mathbf{U}_d^{(t)})^T$
end for
Calculate core tensor: $\mathcal{Y} = \mathcal{X}^{(t)} \prod_{i=1}^{M} \times \mathbf{U}_i^{(t)}$

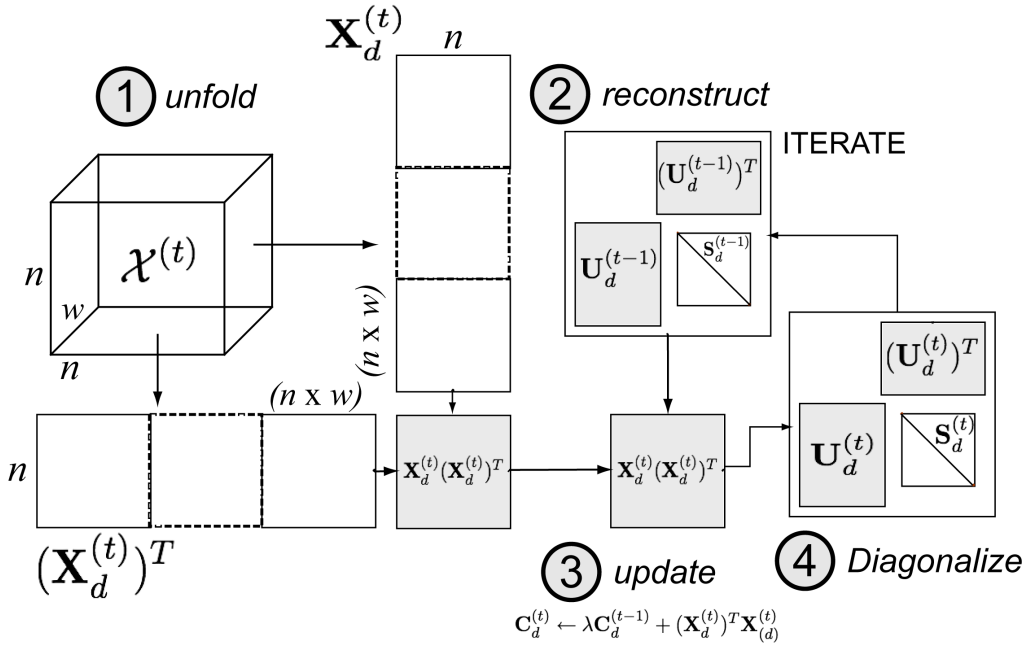

Figure 4.2: **Dynamic Tensor Analysis for online monitoring and analysis of protein dynamics**. A schematic representation of Algorithm (1). With the arrival of new data at time window t, the tensor $\mathcal{X}$ is (1) unfolded in dimension d, and the variance in the data is computed using an inner product of the unfolded tensor. The resulting variance matrix $\mathbf{X}_d\mathbf{X}_d^T$ is (3) updated to the existing variance matrix from time window $(t-1)$, followed by an (4) eigen-decomposition of the new variance matrix. The new eigen-basis is (2) stored to be used for the next iteration.

The algorithm takes as input the new incoming tensor $\mathcal{X}^{(t)}$ at time t, the eigenvalues $\mathbf{S}_i^{(t-1)}|_{i=1}^M$, and the eigenvectors $\mathbf{U}_i^{(t-1)}|_{i=1}^M$ computed from the preceding call to DTA at time $(t-1)$. If there are no previous eigenvalues/ eigenvectors (i.e., at $t=0$), then the only input is the first tensor, and the eigenvalues/ eigenvectors will be computed for use in subsequent calls to DTA.

Example: Assuming we are applying DTA to distance map tensors (Sec. 4.2), the inputs include tensors $\mathcal{X}^{(t)} \in \Re^{n\times n\times w}$, $\mathbf{S}_i^{(t-1)}|_{i=1}^3$, and $\mathbf{U}_i^{(t-1)}|_{i=1}^3$. Recall that the window size, w, is defined by the end-user.

We note that Algorithm 1 does not apply dimensionality reduction along. However,

dimensionality reduction can be applied in a straight forward manner. The primary difference being that the eigenvector and eigenvalue matrices will have dimension $\mathbf{U}_i^{(t-1)}|_{i=1}^{M} \in \Re^{n_i \times r_i}$, and $\mathbf{S}_i^{(t-1)}|_{i=1}^{M} \in \Re^{r_i \times r_i}$, respectively, where $r_i < n_i$ is the reduced size of the ith dimension.

The algorithm proceeds by minimizing the variance in every dimension i, ($1 \leq i \leq M$). Line 4 involves *unfolding* (or matricizing) the $\mathcal{X}^{(t)}$ in the selected dimension d. Given $\mathcal{X}^{(t)} \in \Re^{n_1 \times \cdots \times n_M}$, the matrix unfolding in dimension d involves constructing the $(\prod_{i \neq d} n_i) \times n_d$ matrix $\mathbf{X}_{(d)}$ such that each row is a vector in $\Re^{n_d}$ obtained by holding d fixed, and varying the remaining indices. In line 5, the variance matrix associated with dimension d from the previous call to DTA, $\mathbf{C}_d^{(t-1)}$, is reconstructed using the eigenvectors and eigenvalues from the predecessor window $(t-1)$. The variance of the unfolded incoming tensor is, by definition, $\mathbf{X}_d^T \mathbf{X}_d$. Line 6 is the key to the DTA algorithm. It involves updating our estimate of the variance according to the rule:

$$\mathbf{C}_d^{(t)} \leftarrow \lambda \mathbf{C}_d^{(t-1)} + \mathbf{X}_d^T \mathbf{X}_{(i)} \tag{4.6}$$

where, λ is referred to as the *forgetting factor*. This is a parameter controls the degree to which previous tensors influence the current estimate of the variance. When $\lambda = 0$, only the present tensor at time t is considered to be relevant and all the past tensors are ignored for updating the variance matrix. In our experiments, we have set $\lambda = 1$, giving equal weight to the previous and current estimates. Line 7 of the algorithm diagonalizes the variance matrix, resulting in an updated set of eigenvalues and eigenvectors that capture the dynamical behavior observed in the simulations observed thus far. Finally, the core tensor is computed as shown in line 9 of the algorithm.

Complexity. The space complexity for DTA includes storing costs involving the current tensor, the eigenvectors and eigenvalues computed and finally, the core-tensor. Note that the dominating factor is the storage of the current tensor $\mathcal{X}^{(t)}$, which amounts to $O(\prod_{i=1}^{M} n_i)$. In the distance map example, the storage is thus $O(w \times n^2)$. In comparison, an offline algorithm must store the entire trajectory, resulting in a storage cost of $O(t \times n^2)$.

The time complexity of the algorithm is dominated by lines 6 and 7 in Algorithm 1. The total cost is $O(\sum_{i=1}^{M} n_i^3 + \sum_{i=1}^{M} n_i \prod_{j=1}^{M} n_j)$ where the first term corresponds to the cost of diagonalization (line 7) and the second term is the cost of updating the variance matrix (line 6). This cost is dominated by the first term when the dimension of the input tensor is low ($M \leq 5$), and by the second term otherwise. In either case, the total cost is far less than an offline algorithm, which would require performing diagonalization of much larger matrices.

Finally, we note that [251] also describes an approximation to DTA, which they call streaming tensor analysis (STA), which adaptively updates the eigenvector matrices depending on the changes seen in the current tensor, instead of diagonalizing the variance matrix at every time window. This adaptive update of eigenvector matrices provides significant computational speedup at the cost of accuracy. However, in practice, we have found that DTA is superior to STA in the context of analyzing MD trajectories [226].

4.3.3 Extracting Information from MD data using Tensor Analsis

As previously mentioned, PCA is often used in the context of dimensionality reduction, and tensor analysis can also be used to perform the same task by dropping low-variance components. This may be useful for archiving molecular dynamics data, or to summarize the data in a compact form. We focus instead on the problems of spatio-temporal pattern discovery and anomaly detection.

Pattern Discovery. When applied to either of the tensor streams defined in Sec. 4.2, we will obtain three orthogonal matrices, $\mathbf{U}_1, \mathbf{U}_2$, and $\mathbf{U}_3$. While tracking C^α distance tensors, observe both $\mathbf{U}_1$ and $\mathbf{U}_2$ are identical, since the input distance matrix is symmetric. Both orthogonal matrices describe the collective spatial distance fluctuations in the protein. However, $\mathbf{U}_3$ describes correlations in the time dimension, which is useful to identify time-points along the trajectory where the collective behavior of a protein changed significantly. Given these matrices, we can identify correlations in the underlying distance fluctuations using $\mathbf{U}_1$ by performing k-means clustering [116]. Analogous to the use of clustering in the analysis of microarray data, which identifies correlated genes, the clusters we obtain identify clusters of dynamically coupled residues. The individual cluster centers identify the average distance fluctuations within each cluster; larger distance fluctuations imply flexible regions, whereas smaller distance fluctuations imply rigid regions in the protein. Further, high values along the columns of $\mathbf{U}_1$ or $\mathbf{U}_2$ indicate residues that showed high distance fluctuations with respect to other residues in the protein. Thus, apart from identifying clusters of residues that are dynamically coupled, we can also reason about individual residues, if they are constrained or flexible.

Anomaly Detection. DTA is continually updating its estimates of the orthogonal matrices, $\mathbf{U}_1, \mathbf{U}_2$, and $\mathbf{U}_3$, as new data arrive. These matrices are a compact description of the dynamical behaviors exhibited in the earlier data. Given these, we can use the reconstruction error metric defined in Eq. 4.3 to detect anomalies. That is, points where there are

significant deviations from the dynamical behavior encoded in the orthogonal matrices. Such changes in dynamical behavior may signal a change between microstates. To detect anomalies, we monitor the empirical mean and standard deviation of e as the simulation is running. Next, we define an error threshold error as follows:

$$e_t \geq mean(e_i|_{i=1}^{t}) + \alpha.std(e_i|_{i=1}^{t}) \tag{4.7}$$

where $e_i|_{i=1}^{t}$ refers to the error of reconstruction up to time t, and α is an arbitrary positive constant. In our experiments we set α to 2. That is, our threshold is an error that is two standard deviations, or more, above the empirical mean. If the incoming tensor exceeds this threshold, we declare that the new tensor is an anomaly. That is, an event of interest.

Apart from indicating changes in dynamical behavior, the events of interest can also provide insights into the structural deviations in the protein along two or more regions in the protein simultaneously. This change in collective dynamical behavior is complemented by one or more conformational changes along the trajectory, which may reflect how the protein (or its substrate) responded to changes in its dynamics. As we will show in the results section, such changes are important from a biological perspective, especially while studying enzyme reactions, where dynamical changes in the protein affect the conformation of the bound substrate.

4.3.4 Related Work

A number of techniques have been developed to analyze and reason about collective behavior from MD simulations. PCA based techniques (reviewed in [43]) are very popular in the MD community and are used to visualize collective behavior in proteins. Other techniques based on spectral analysis [63] as well as mutual information [178, 168] are also widely used in studying collective motions. However, there are two limitations to these approaches.

The first limitation relates to the fact that PCA can be done only in *two dimensions*; hence it is possible to obtain only insights into collective behavior from the perspective of spatial variance, no insight can be drawn from the temporal aspects of the simulation. Mutli-dimensional PCA and multi-way analysis [245], which have been applied to assign nuclear magnetic resonance (NMR) spectra [187, 248] and in cheminformatics [108], overcome this limitation, but have not been used to study molecular dynamics trajectories, to our knowledge. Tensor analysis has also been applied in a variety of problem domains including visual analysis of images from electroencephalogram [2] and systems biology [290]. The identification of rigid/ flexible sub-structures within a protein is a well

studied problem, especially from the viewpoint of rigidity theory [279, 140]. Application of these techniques to multiple snapshots of proteins proves to be unreliable [188] and similarly, it is not possible to obtain information about time-points where deviation in collective behavior is observed.

The second limitation arises from the fact that most traditional molecular dynamics analysis techniques are intended for offline use. Traditional online techniques track root mean squared deviation and similar aggregate statistics. Here, the end user is responsible for deciding whether sudden changes in these statistics represent an interesting change in behavior, or just a transient. In contrast, our method explicitly monitors spatio-temporal patterns, and alerts the user when such patterns change. Significantly, we will show in the next section that tensor analysis can detect significant changes in collective motions that do not necessarily correspond to large root mean squared deviations. That is, tensor analysis reveals information in molecular dynamics trajectories that cannot be obtained by visually inspecting snapshots, or by traditional online algorithms.

4.4 Implementation and Results

In this section, we analyze the equilibrium simulations of two proteins: (a) ubiquitin and (b) barnase. Both proteins have been studied extensively using experimental and theoretical/ computational techniques and are therefore ideal to test the utility of our method. Each trajectory was preprocessed to construct the tensors. These tensors were processed in MATLAB using the tensor toolkit libraries [20, 21] and an implementation of the DTA algorithm [252]. A python version of the code has also been made publicly available: http://pytensor.googlecode.com/ [291].

4.4.1 Equilibrium Simulations of ubiquitin

We test our algorithm on ubiquitin, a widely characterized protein. In chapter 2 we used QHA on ubiquitin to characterize the changes in collective motions at the μs-timescale. In chapter 3, we characterized the hierarchical nature of the equilibrium fluctuations in ubiquitin using a novel linear model that pursues higher-order statistics. In this chapter, we demonstrate that it is possible to obtain insights into changes in collective behavior as simulations progress. The equilibrium simulation of ubiquitin had a total of 10,000 conformations,spanning a total of 0.5μs [225]. Third order tensors were then constructed with every 10 snapshots ($w = 10$) aggregated into one tensor, thus providing a total of 1,000 tensors.

Collective dynamics in ubiquitin. When we examine the constrained residues from the eigenvector matrices $\mathbf{U}_1$, we find that these residues line the hydrophobic core of the protein. Thr7, Thr9, Thr14, Glu18, Ser20, Ile44, Gly53, Ser57, Tyr58 and His67 Fig. (4.3(a)) are identified to be restrained. The eigenvector matrices $\mathbf{U}_1$ representing the correlations in distance fluctuations from the MD trajectory for ubiquitin were clustered using k-means. Four dynamically coupled regions in ubiquitin were observed. The assignment of clusters obtained from k-means was then mapped onto the three-dimensional structure of the protein and visualized using PyMOL [73]. As shown in Fig. 4.3(b), the clusters show a clear separation of the β sheet from the rest of the protein. The loops, L1, L2, L3 and L4 (shown in yellow) exhibit most distance fluctuations, followed by the C-terminal tip of α_1 (shown in green). Both helices (shown in blue) exhibit the least distance fluctuations and the β sheet exhibits intermediate distance fluctuations with respect to the protein. The collective distance fluctuations indicate that different parts of ubiquitin exhibit unique dynamical behavior involved in binding multiple substrates. The fluctuations of loops L1 and L2 are coupled indicating similar fluctuations across these two regions. It is interesting to note that the C-terminal tip of α-helix seems to exhibit somewhat different collective dynamics than that of the loops L1-L4.

From a biological perspective, the coupling in fluctuations at both L1 and L2, although separated by over 15 $\mathring{A}$ indicates an innate ability of ubiquitin's binding regions (L1 and L2) to move in similar manner. Both L1 and L2 exhibit fast time-scale fluctuations and are involved in modulating the conformational landscape (as we have seen in Chapter 3; Fig. 3.7). Further, the distinct behavior of the C-terminal end of α_1 helix having an intermediate level of fluctuations between the β sheet and flexible loops (L1-L4), shows that this region can move independently of the other regions in the protein.

Anomaly detection in ubiquitin simulations. As previously noted, the simulations were collected from 8 different starting structures leading to extensively analyze the collective fluctuations of ubiquitin at the μs time-scales. Now, we pay close attention to the time-evolution of the simulations. We observe, as shown in Fig. 4.4, that of the 1,000 windows, 58 of these windows show e to be above the 2σ interval. Observe that the RMSD plot shown in Fig. 4.4B does not show any correlation with the error of reconstruction.

We illustrate the collective conformational changes from three different windows (11, 2140 and 6060). In window 11 (Fig. 4.4C), collective conformational changes are dominant along $\beta_1 - \beta_2$ and $\beta_3 - \beta_4$ as well as $\beta_1 - \beta_2$ and $\alpha_1 - \beta_3$ loop regions. Note that in this window, both loop regions $\beta_1 - \beta_2$ have moved far apart compared to the previous time-window. In window 2140 (Fig. 4.4D), we observe that the most dominant changes include α_1 and $\beta_1 - \beta_2$ regions where by, these regions have moved closer to each other. Note

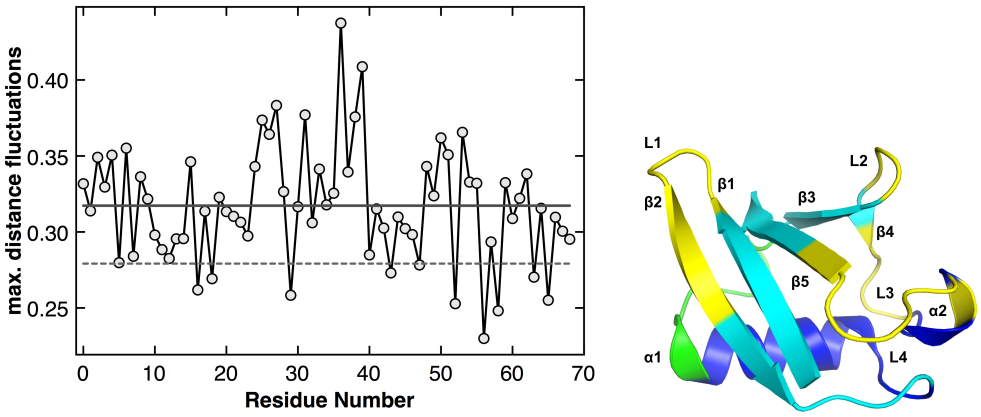

Figure 4.3: **Dynamically Coupled Regions in ubiquitin**. (A) Constrained residues in ubiquitin. A total of 13 residues lining the hydrophobic core of the protein are constrained. Four clusters are identified; (B) α_1 (shown in blue) and $\beta_1 - \beta_5$ (shown in cyan) undergo low distance fluctuations. $\alpha_1 - \beta_3$ (shown in green) undergo intermediate distance fluctuations where as L1-L2 shown in yellow undergo maximum distance fluctuations.

in both windows 11 and 2140, the overall RMSD between the structures is quite small, however the rearrangements of the loop regions is of importance since the modulation of fluctuations in these regions have some relevance to binding multiple substrates. In the third window (6060; Fig. 4.4C), the motions are substantial, involving α_1, $\beta_1 - \beta_2$, β_5 and $\beta_3 - \beta_4$; thus the change in collective motions observed span the entire protein.

4.4.2 Equilibrium Simulations of Barnase

Barnase is a 110 residue bacterial ribonuclease synthesized and secreted by *Bacillus amyloliquefaciens* [86]. It is known to be lethal to the cell without its strongly binding inhibitor barstar. This protein is often used to study protein-protein binding and protein folding [69]. Barnase consists of a long α-helix (α_1) and a 5-stranded β-sheet ($\beta_1 - \beta_5$), interspersed with two shorter α-helices (α_2 and α_3), connected by loops (L1-L5). The equilibrium simulation of inhibitor-free barnase had a total of 2,500 snapshots, constituting a total trajectory length of 10 nanoseconds [188]. Third order tensors were then constructed with every 10 snapshots ($w = 10$) aggregated into one tensor, thus providing a total of 250 tensors for barnase.

Collective dynamics in barnase. The eigenvector matrices $\mathbf{U}_1$ representing the correlations in distance fluctuations from the MD trajectory for barnase were clustered using k-means. Four dynamically coupled regions in barnase were observed. The selection of the best value of k for clustering was based on the mean value of cluster separation; for any value of k, if the mean value of the cluster separation was higher than that of $k-1$, another round of k-means clustering was applied, otherwise, k was chosen to be the optimal number of clusters.

The assignment of clusters obtained from k-means was then mapped onto the three-dimensional structure of the protein and visualized using PyMOL [73]. As shown in Figs. 4.5(b), α_1 and β-sheet ($\beta_1 - \beta_5$) form two separate clusters. These two regions, on average, show low distance fluctuations compared to the rest of the protein, indicating that both α_1 and β-sheet may be involved in slow time-scale motions. Residues 21-48, including α_2 and α_3, form another cluster, involving intermediate scale distance fluctuations. The loop regions (L3-L5; N- and C-termini) cluster together into one region , with highest distance fluctuations. Residues 50-52 are clustered into three different groups indicating an inherent difference in the flexibility of these residues which may play a role in the hinge site dynamics of barnase [285, 86]. We were also able to identify residues that were constrained or flexible during the simulation which correlate well with previous experimental and theoretical methods (Figs. 4.5(a) and 4.6).

A unique feature of DTA is its ability to characterize collective behavior in multiple tensor streams such as hydrogen bonds (HB) or hydrophobic (HP) interactions *simultaneously*. While distances between C^α atoms are symmetric ($\mathbf{X}_{ij} = \mathbf{X}_{ji}$), HB interactions involve tracking distances between two separate sets of atoms namely HB-donors (HBD) and HB-acceptors (HBA). The resulting distance matrix between HBD and HBA involves tracking a sparse, asymmetric tensor-stream over time.

DTA on the HBD and HBA results in two eigenvector matrices, $\mathbf{U}_{HBD}$ and $\mathbf{U}_{HBA}$ respectively: $\mathbf{U}_{HBD}$ determines correlations in donor-donor distance fluctuations and $\mathbf{U}_{HBA}$ determines correlations in acceptor-acceptor distance fluctuations. Using k-means, we were able to cluster HB-donors into 5 groups, and HB-acceptors into 3 groups. Grouping the HB-donors and HB-acceptors based on their respective cluster assignments, we observed that α_1 and $\beta_1 - \beta_5$ form the largest cluster (Fig. 4.7A), indicating the dynamical behavior of hydrogen bonds in these two regions are similar. To examine their similarity, we considered the average life-times [188] of these interactions during the simulation. We found that these HB interactions were extremely stable during the course of the simulation with an average life-time of about 70% of the simulation. The second largest cluster of hydrogen bonds identified via DTA is localized to residues 21-48 (Fig. 4.7B), concurring with the behavior identified in C^α distance fluctuations. Subsequent clusters of HB are

smaller and localized along the flexible loop regions in barnase. We found the lifetimes of these HB interactions to be more transient with life-times averaging around 50% of the simulation.

By studying the collective behavior across multiple tensor-streams, one may gain insights into the dynamical origins of collective motions in barnase. First, from our analysis of C^{α} distance fluctuations, we note correlation patterns in the trajectories indicate coupling between distally located residues. Loops L3, L4 and L5 that are widely separated by an average distance of 10 Åshow similar dynamics. Second, the collective fluctuations in C^{α}-atoms and HB interactions separate residues 21-48 from the rest of the protein. This indicates the dynamical motions of these residues to be decoupled from the rest of the protein, forming a separate dynamical domain, agreeing with previous experimental and computational work [206, 298]. Third, the distinct assignment of residues 50-52 into three clusters is also relevant to barnase function since the inherent difference in flexibility along these three residues may be essential for inter-domain movements [206].

Anomaly detection in equilibrium simulations of barnase. Our method can identify time-points during which collective behavior changed significantly. That is, time-points where the incoming tensor exhibits patterns that are significantly different than those from the past. This is accomplished by monitoring the error of reconstruction (Eq. 4.3) and evaluating Eq. 4.7. A plot of the error of reconstruction against the tensor windows is shown Fig. 4.8A. Beyond the second standard deviation threshold, we find 14 tensor windows. Fig. 4.8B illustrates the average RMSD of the structures within the ith tensor to those in the preceding tensor. Notice that the average RMSD is all very small ($\approx$ 1 Å), indicating that the error of reconstruction, as expected, is detecting changes in the collective behaviors of the residues, and not merely significant structural changes. Moreover, the curves in Fig. 4.8-(A) and Fig. 4.8-(B) are not clearly not correlated, further supporting the claim that these two statistics are revealing different kinds of information about the MD trajectory.

In Fig. 4.8C we show two structures to illustrate changes in collective dynamics. In the tensor window 40 (highlighted by a dotted circle), we detect the collective movement of the flexible loops (L1-L2) in residues 21-48 towards L3, L4 and L5. This movement may be of functional significance, since it involves the residues involved in the binding pocket of barnase coming closer by about 3 Å. In another window (215), loop L1 had moved with respect to $\alpha 1$ helix compared to the predecessor window (Fig. 5C bottom panel). This collective displacement of these loops with respect to residues 21-48 is of importance since they represent inter-domain movements [206].

4.5 Conclusions

We have shown that it is possible to obtain biologically relevant information about flexible and rigid substructures within a protein as the simulations progress. This was illustrated on two entirely different sets of simulations and in both cases, we were able to identify dynamically coupled regions and to identify sudden changes in those couplings. While preliminary, our findings are consistent with those obtained via previous computational and experimental studies.

Our experiments were limited to tensors encoding the dynamics of distance maps and hydrogen bond networks. These are by no means the only tensors which may provide biologically useful information. For example, it may be beneficial to track forces or velocities over different atoms to detect patterns of energy flow in protein structures. Tensors also provide a natural means for examining higher-order relationships, such as the correlations between non-covalent interactions and how they account for large-scale motions within a protein. This is of fundamental interest to chemists and biologists alike in domains such as protein and drug design where networks of covalent and non-covalent interactions are known to widely affect protein function [8, 6, 298, 250].

The core tensor $\mathcal{X} = \mathcal{C} \prod_{i=1}^{d} \times \mathbf{U}_i$ can be used as an efficient means to approximate the dynamics thus far observed in a simulation, and provides a convenient means for summarizing an otherwise massive data set, or to compare and cluster trajectories say, across species. For example, imagine partitioning the trajectories of two different trajectories into a sequence of core tensors that correspond to regions flanked by spikes in the error of reconstruction metric. Such core tensors represent quasi-stable states, and so one can imagine comparing two trajectories by determining whether they visit similar sets of quasi-stable states and, if so, whether they do so in roughly the same order. Alternatively, because the tensor analysis can be tuned to identify patterns at particular time-scales, one might compare the core tensors observed at one timescale with those obtained at another, potentially leading to a hierarchy of spatio-temporal pattern.

From the perspective of analyzing simulations in an online fashion, a direct extension to our method will be to fork simulations from the time-points where significant deviation in dynamics is observed in order to sample a larger conformational space. This, in turn, is particularly useful to drive simulations involving chemical reactions or folding pathways which are known to be hard to simulate. We are investigating the use of tensor analysis for analyzing protein folding trajectories as part of ongoing work.

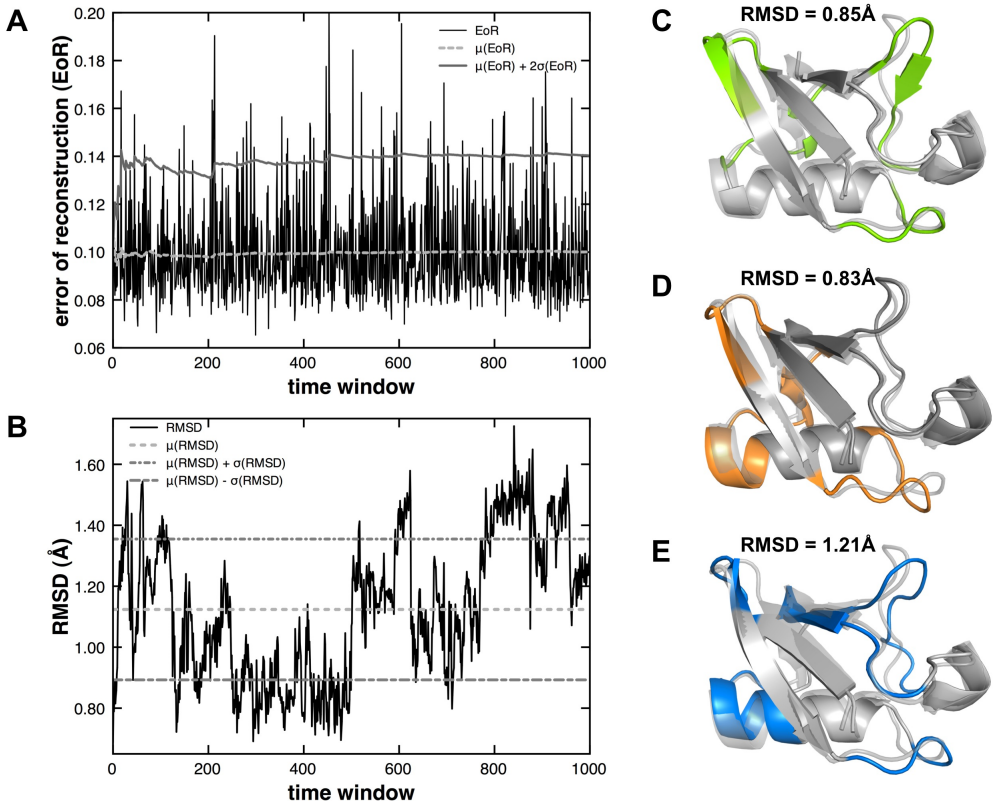

Figure 4.4: **Error of reconstruction (EoR) for ubiquitin compared to root mean squared deviations (RMSD)**. (A) shows EoR plotted as a function of tensor window. The dotted line indicates the mean reconstruction error as per Eq.(4) and the second standard deviation interval (gray solid line) is also plotted. Dotted circles are used to highlight tensor windows shown in the adjacent panel. (B) average window RMSD plotted for ubiquitin, showing the average (gray dotted line) and standard deviation interval (gray line) for the simulation. (C) Structural changes associated with $w = 11$ (top panel) showing movements in L1 and L2 overall RMSD 0.85 Å; (D) $w = 214$ (bottom) showing movements associated with $\alpha 1$ and L1 with an overall RMSD of 0.83 Åwith the preceding window. Note in both these cases, we see changes in L1 and L2 and α_1. (E) shows another window $w = 606$ where motions in β sheet as well as α_1 are evident, with an overall RMSD of 1.21 Å.

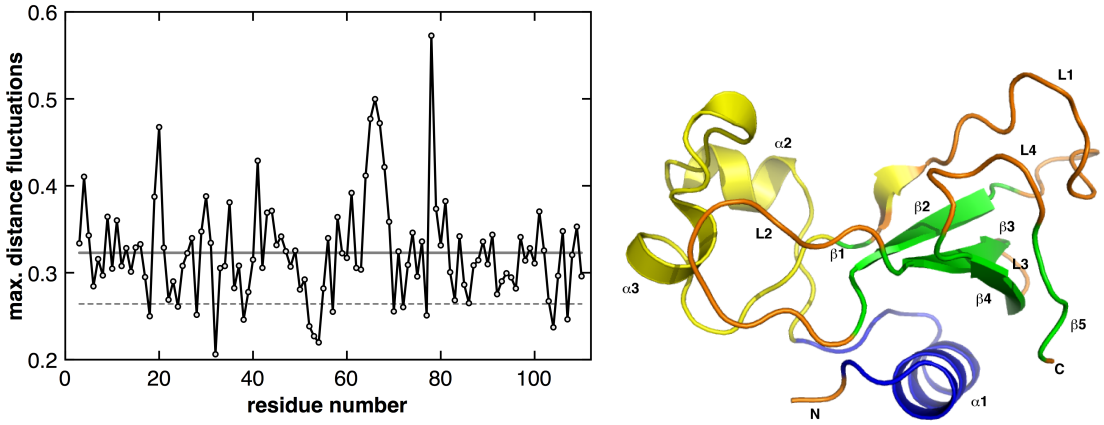

Figure 4.5: **Dynamically Coupled Regions in barnase**.(A) A total of 15 residues were identified to be constrained - below the first standard deviation interval (shown as gray dotted line). It is interesting to note that the hinge site residues Gly52 and Gly53 are constrained. Further residues important for the structural integrity of the protein Tyr24, Ala32, Arg87 and Thr103 are also observed to be constrained. (B) Four clusters are identified; α_1 undergoes the least distance fluctuations (shown in dark blue) and $\beta_1 - \beta_5$ form two clusters shown in the top panel showing slightly higher distance fluctuations (shown in green and yellow). $\alpha_2 - \alpha_3$ and L1-L2 (residues 21-48) separate into a cluster (yellow). The N- and C-termini and loops L3-L5 are form a cluster (shown in orange) undergoing the largest distance fluctuations.

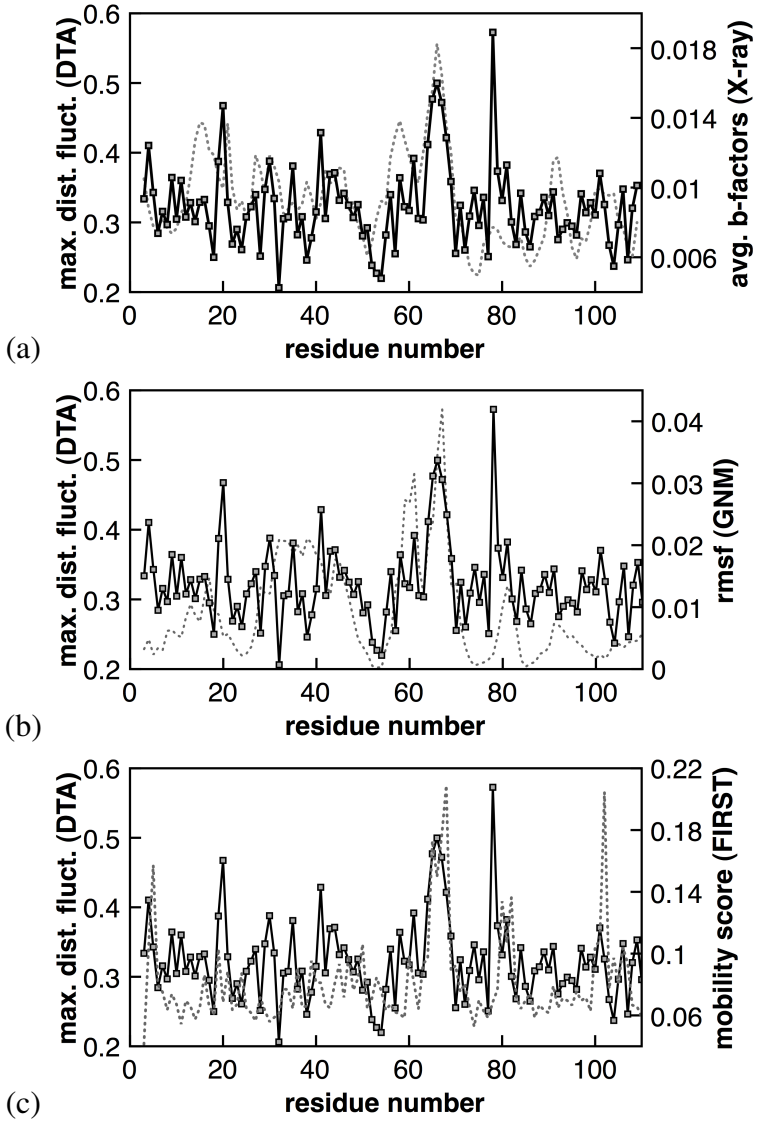

Figure 4.6: **Comparing flexibility profiles of barnase using experimental and theoretical techniques**. (a) DTA distance fluctuations are plotted (black solid line) against the average root mean square fluctuations (RMSF) determined from 6 different crystal structures of barnase. (b) Distance fluctuations determined from DTA against that of RMSF determined via guassian network model (GNM) [22], shows good agreement with respect to the hinge sites (Gly52-Gly53) as well as the flexible regions in the protein. (c) Mobility scores determined from FIRST [140] compared with DTA determined distance fluctuations; note good agreement between flexible regions determined via FIRST and DTA, however, there are some differences at the hinge site.

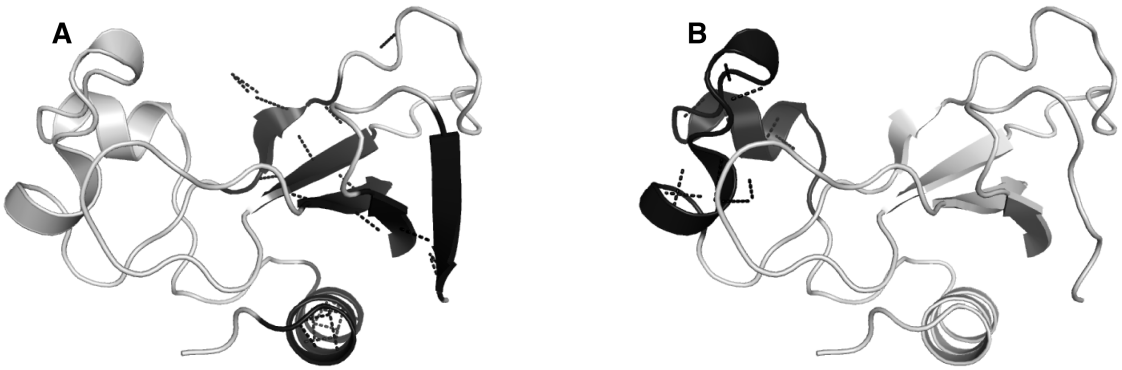

Figure 4.7: **Clustering hydrogen bond (HB) interactions based on similar dynamics**. (A) Largest cluster formed by stable secondary structure interactions along barnase, including interactions in α_1 and $\beta_1 - \beta_5$. (B) The second largest cluster consists of HB interactions in residues 21-48. Loops (L1-L5) form HB interactions that have different characteristics based on their location in the protein and exhibit more transient behavior.

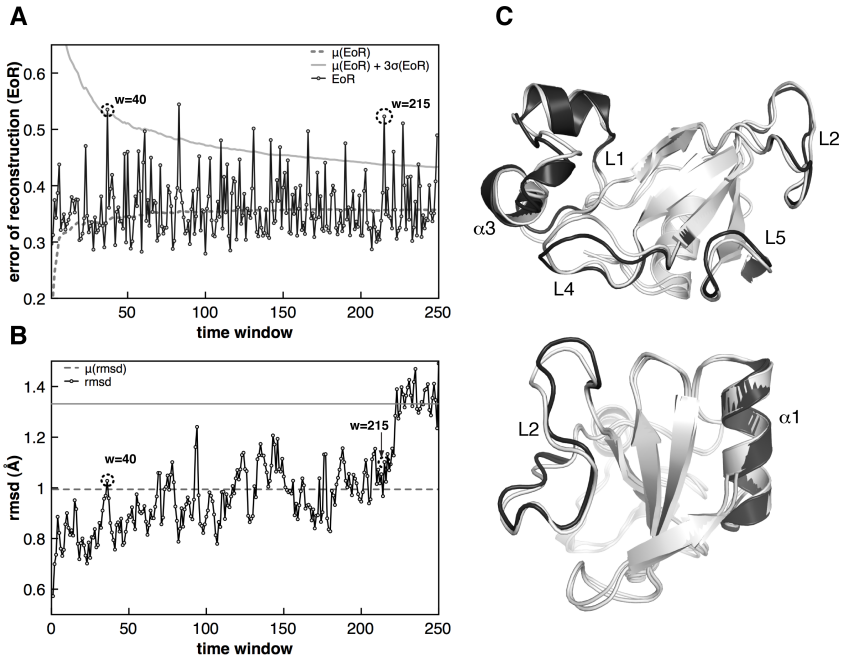

Figure 4.8: **Error of reconstruction (EoR) for barnase compared to root mean squared deviations (RMSD)**. (A) shows EoR plotted as a function of tensor window. The dotted line indicates the mean reconstruction error as per Eqn.(4) and the second standard deviation interval (gray solid line) is also plotted. Dotted circles are used to highlight tensor windows shown in the adjacent panel. (B) average window RMSD plotted for barnase, showing the average (gray dotted line) and second standard deviation interval (gray line) for the simulation. Note the two snapshots selected for analysis are highlighted in dotted circles. (c) Structural changes associated with $w = 40$ (top panel) showing movements in L1 and L2 along with the functional domain 21-48, overall RMSD 1.03 Å; $w = 215$ (bottom) showing movements associated with $\alpha 1$ and L1 with an overall RMSD of 0.612 Åwith the preceding window. In both cases, the regions shown in a darker shade of gray represent predecessor windows whereas lighter shade shows the current window.

Chapter 5

Comparing the Intrinsic Dynamics of Cyclophilin A Before, During and After Catalysis

Enzymes are perhaps the most widely studied proteins. In this chapter, we set the background for the two questions outlined in the introduction (Chapter 1). We investigate, in this chapter, the intrinsic dynamics of the enzyme cyclophilin A (CypA), en extensively characterized enzyme. CypA catalyzes the isomerization of prolyl-peptidyl bonds in a variety of substrates including small peptides and large proteins. In this chapter, we characterize the internal dynamics of CypA under three scenarios: (a) when CypA is in its unbound state - without any substrate in its active site, (b) when CypA is bound to a substrate peptide (in reactant state and product state) and finally, (c) when CypA catalyzes the isomerization of the bound peptide. Our analysis will be based on the techniques that were developed in chapters 2 and 4. This chapter will serve as an illustration of how a combination of these techniques could elucidate the complex behavior of CypA under these scenarios. Further, it will also test critically the hypothesis where intrinsic dynamics catalysis are closely associated with one another [47, 49, 77, 78, 125] or if there are subtle distinctions in the way CypA "alters" its dynamics to perform its function[1].

[1]This chapter is adapted from:

- Arvind Ramanathan, Pratul K. Agarwal, Maria G. Kurnikova, Christopher J. Langmead, An online approach to mine collective behavior from molecular dynamics simulations, J. Comp. Bio. (2010), 17:3 (in press)
- Arvind Ramanathan, Pratul K. Agarwal, *Computational Identification of Slow Conformational Fluctuations in Proteins*, J. Phys. Chem. B., 2009, 113 (52), pp 16669–16680.

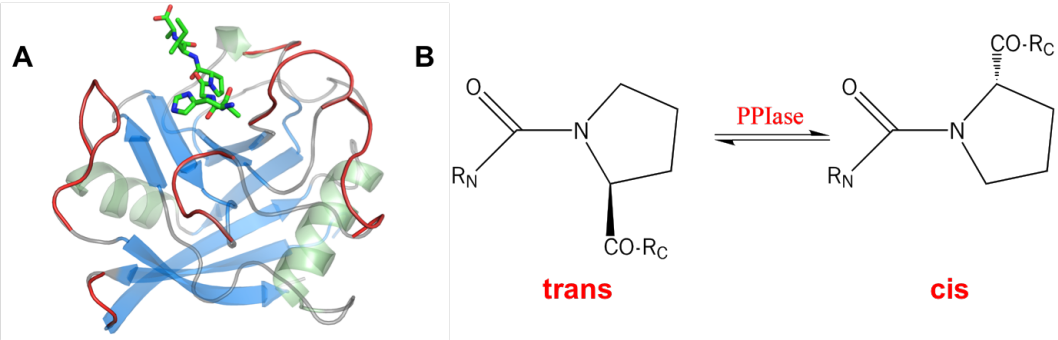

Figure 5.1: **Reaction catalyzed by peptidyl-prolyl isomerases (PPIases) including cyclophilin A**. The $trans$ conformer from the peptide around the amide bond is rotated by 180° into the cis conformation, as depicted in this 2D plot.

5.1 Introduction

Enzyme catalysis has intrigued biochemists for over a century: even today, the origins of the remarkable catalytic power in naturally occurring enzymes is unknown. Although there is considerable speculation in the literature about how enzymes function, a recent proposal gaining significant momentum is that perhaps enzymes are intrinsically dynamic molecules that have evolved to precisely move in directions that enable function [24, 124]. In this context, one would need to examine critically the evidence presented that either support or disagree with the dynamical hypothesis. In this part of the thesis, we are primarily concerned with garnering support for the dynamical hypothesis. Before answering these questions though, it becomes necessary to examine if the intrinsic dynamics of an enzyme, in the absence of its natural substrates is similar at all to the functional dynamics as the enzyme catalyzes its reaction.

In this chapter, we examine the similarity in internal dynamics of an enzyme Cyclophilin A (CypA) in three different scenarios: (a) when CypA is in its unbound state - without any substrate in its active site, (b) when CypA is bound to a substrate peptide (in reactant state and product state) and finally, (c) when CypA catalyzes the isomerization of the bound peptide. CypA is an excellent example case study for testing the dynamical hypothesis since there is a wealth of experimental and computational evidence that have extensively investigated the enzyme under several conditions. It is a fairly small protein that does not require the presence of co-factors or ions to function. Further, CypA is an

important drug target since it is known to be involved in HIV-1 pathway.

5.1.1 Cyclophilin A: Intrinsic dynamics and its impact on the catalytic process

CypA (see Fig. 5.1A) is ubquitously expressed protein in both prokaryotes and eukaryotes, and is a major intracellular receptor for the immunosuppressive drug cyclosporin A. It belongs to the class of enzymes called cyclophilins which play a critical role in several biochemical pathways, including protein folding, cellular transport and biological signaling. The reaction catalyzed by CypA is that of isomerization peptidyl-prolyl amide bonds that are N-terminal to proline residues. Human CypA is a single chain polypeptide consisting of 165 amino acid residues. Its molecular structure reveals the presence of eight β-strands ($\beta_1 - \beta_8$; Fig. 5.1A), forming a β-barrel, with the hydrophobic residues lining the inner core of the protein. Two α helices (α_1 and α_2) flank the hydrophobic core of the protein. The active site is a shallow cavity on one of the faces of the protein, surrounded by several solvent exposed surface loops which typically exhibit high beta-factors as observed from X-ray crystallographic studies. The reaction mechanism catalyzed by CypA is illustrated in Fig. 5.1B. Note that in this chapter, we study the $cis/trans$ isomerization reaction of the bound peptide in CypA.

NMR studies on CypA by Kern and co-workers [77] have suggested a close link between the enzyme's internal dynamics and the substrate isomerization step (catalysis). It was observed that internal dynamics of several regions along the protein were strongly correlated with the catalytic step. Several active site residues and surface loop residues: Arg55, Lys82, Leu98, Ser99, Ala101, Asn102, Ala103, and Gly109, exhibited considerable motions in the presence of substrate alone. It was also shown that certain residues: Phe67, Asn71, Gly74, Ser77 and Ser110 showed motions in both presence and absence of the substrate, where as Thr68 and Gly72 showed increased motions only during the catalytic step of the enzyme. This evidence and follow up work [78], highlighted the importance of CypA's innate ability to exhibit dynamics that were somehow related to catalysis.

Agarwal and co-workers [9, 4], using MD simulations modeled the free-energy profile (see Sec. 5.2.3 for more details) for CypA with multiple susbtrates. These detailed investigation of the free-energy landscape revealed that over the course of the reaction, important hydrophilic (Arg55, Asn102) and hydrophobic (Phe60, Phe113, Leu122 and His126) residues played a critical role in stabilizing the substrate peptide bound at the active site. While the target proline is held rigid at the active site, the carbonyl oxygen from

the preceding residue in the substrate rotates a 180°. Quantum mechanical modeling of the active site also revealed that this mechanism was in agreement with observations from experimental studies. Analysis using QHA (see Chapter 2), revealed three vibrational modes that were most coupled to the catalytic step of the reaction. All the modes associated with the catalytic step exhibited concerted motions along the surface loop regions of the enzyme showing large displacements. The correlated motions from the vibrational modes as determined from QHA revealed an interesting picture about CypA's function. A series of interconnections extending from the flexible surface regions of the protein all the way to the active site of the enzyme were responsible for the observed correlated motions in these vibrational modes. These interconnections included hydrogen bonds and hydrophobic interactions amongst highly conserved residues and the dynamics in these networks altered the crucial interactions between the enzyme and substrate during the reaction process.

The vibrational modes associated with the catalytic step were also found to be promoting the progress of the reaction [5]. In fact addition of kinetic energy to the flexible surface loop regions implicated in these vibrational modes indicated that more trajectories were successful in crossing over the transition state into the product side from the reactant side. Addition of kinetic energy to modes not coupled to the reaction, however, did not result in producing successful crossovers from reactant to product state. It was also shown that addition of kinetic energy to the first solvation shell of the protein enhanced the reaction to proceed into the product state. Thus, the motions in the enzyme can be thought of playing a critical role in promoting the reaction to proceed from the reactant to the transition state. Mutations to the network or alteration of the dynamics within these network could in fact drastically change the outcomes of how CypA would function.

While the aforementioned studies are extremely important, it should also be noted that both the computational studies considered the dynamics during the catalytic process. In this chapter, we specifically ask the question if the dynamics before and after catalysis (with the substrate peptide bound at the active site) and the dynamics of the substrate-free enzyme are similar at all. We analyze this problem by carrying out extensive simulations on the substrate-free CypA molecule as well as the substrate bound CypA molecule in the three scenarios outlined above. The internal dynamics, as quantified by the modes of motion determined via QHA, in the course of the reaction pathway as well as the end-state simulations and substrate free simulations are compared and contrasted. The comparison shows that the large-scale motions are not only similar, but share a large overlap in terms of the conformational sub-states they visit.

5.1.2 Simulating enzyme catalysis: Umbrella Sampling along the reaction pathway

Simulating enzyme catalysis can be a significant challenge. Traditionally done via quantum mechanics/ molecular mechanics (QM/MM) techniques [258], there has been tremendous growth in the area of developing semi-empirical methods such as empirical valence bond (EVB) [276] to understand the mechanistic aspects of enzyme catalysis. EVB employs a standard molecular mechanics force-field (see Chapter 1) and maps the potential surface over the course of a catalytic reaction by using standard free-energy simulation techniques such as thermodynamic integration [96] or umbrella sampling [260].

MD simulations in combination with techniques such as umbrella sampling [260] can be used to model the free energy landscape associated with the enzyme reaction pathway as a function of a reaction coordinate. As depicted in Fig. 5.2 the entire set of conformations sampled along the various sections of the reaction coordinates allows identification of conformational fluctuations at the reaction time scale. The idea behind umbrella sampling is that these simulations would inherently help one overcome poor sampling between two states that are divided by a large free energy barrier by sampling those regions that have a naturally low probability to be sampled using a standard equilibrium simulation. Umbrella sampling is a means of "bridging" the gap that exists between the two states by using a biasing potential that can, in effect, cancel the influence of the energy barrier. Once the umbrella sampling is performed, the free-energy profile (FEP) of the reaction can then be generated using the weighted histogram averaging method (WHAM) [166].

5.2 MD Simulations for Cyclophilin A

In this section, we briefly outline the MD simulations for CypA that were carried out to investigate the relationship in the intrinsic dynamics and catalytic step of the enzyme. Each of the simulations were carried out with the same protocol described in Sec. 2.2.2, involving multiple steps of energy minimization, followed by a careful equilibration procedure that allowed the protein (or system) to equilibrate at 300K and finally full production runs lasting for however long the simulation was scheduled for.

5.2.1 Equilibrium Simulations of Cyclophilin A

The motions in the free enzyme were characterized on the basis of five MD trajectories. The starting points for four of these trajectories were selected from the NMR ensemble

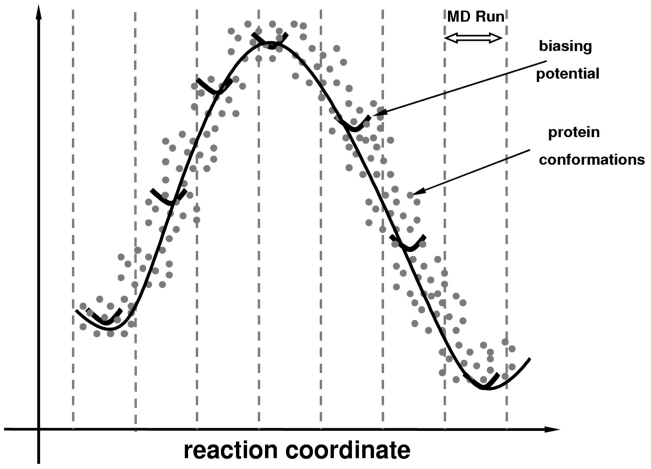

Figure 5.2: **Methodology for identification of slow conformational fluctuations associated with an enzyme reaction.** A number of MD runs are used to sample the conformations along the reaction pathway by using a suitable description of the reaction coordinate and umbrella sampling method. The biasing potentials or umbrella potentials (marked dark black curves) allow sampling of the higher energy regions. The set of conformations (each conformation is indicated by a gray dot) sampled in all MD simulations is used for generation of the free energy profile (black curve) as well as the construction of the covariance matrix for QHA.

(PDB code 1OCA) [211] and an X-ray crystal structure (PDB code 1AWQ) [262] with the substrate excluded. AMBERs parm98 force-field was used for simulations. Each of these structures was simulated in explicit solvent and equilibrated following the procedure described previously (Sec. 2.2.2). A total of 10,000 conformations corresponding to a total of 10 ns of sampling (stored every 1 ps over 2.0 ns for each MD trajectory) were used for further analysis.

5.2.2 CypA End-States Only

The reactant and product states of the CypA reaction (Fig. 5.1A) were modeled based on the X-ray crystal structure (PDB code 1AWQ) [262]. Note, in the reactant state, the bound substrate was held in the $trans$ conformation, where as in the product state, the bound substrate was modeled to be in the cis conformation. Extended MD trajectories of 5 ns each for the reactant and product states were generated (storing snapshots every ps for

both the states). Finally, a total of 8,000 equally spaced snapshots were used for further analysis.

5.2.3 Modeling the CypA reaction pathway

For CypA, the reaction progress is a measure of how much the bound substrate has rotated (isomerized). To track this, the easiest and most accessible measure is the ω angle of the amide bond dihedral angle of the bound substrate. The change in the ω angle is usually referred to as the reaction coordinate. Once a suitable reaction coordinate is identified, it is necessary to sample along the variation of the reaction coordinate to understand how the reaction proceeds.

The motions within cyclophilin A associated with the enzyme reaction are known to occur on the microsecond to millisecond time scale. A combination of 39 simultaneous MD simulations was used to characterize the conformational fluctuations associated with the $cis/trans$ isomerization reaction catalyzed by the enzyme. For umbrella sampling, the amide bond dihedral (ω) of the peptide bound at the active site was restrained harmonically with a force-constant of 0.01 kcal/mol.deg^2, centered at 5° rotations of ω. The procedure for generating the ensemble along the reaction pathway is shown in Fig. 5.3. For each window along the reaction pathway, we performed a total of 20 ps equilibration using the harmonic restraints and the production run from each window lasted another 400 ps. A total of 18,000 conformations were collected as part of sampling the reaction pathway. The free energy profile was computed with a convergence criterion of 1×10^{-4} kcal/mol.

5.2.4 Analyzing collective dynamics in CypA

For each of the trajectories generated, we analyzed the trajectories using QHA [150]. The subspace overlap from each trajectories were also compared using Hess's metric given by Eq. 2.1. Further, we also computed the projections of the snapshots along individual modes using Eq. 2.2. The NMR ensemble (PDB code 1OCA [211]) were also analyzed as a separate ensemble to compare and contrast the motions in the substrate-free simulation. Dynamic tensor analysis (DTA) was performed by extracting 1800 evenly spaced conformations along the reaction pathway and constructing tensors of size $w = 10$, resulting in a total of 180 windows.

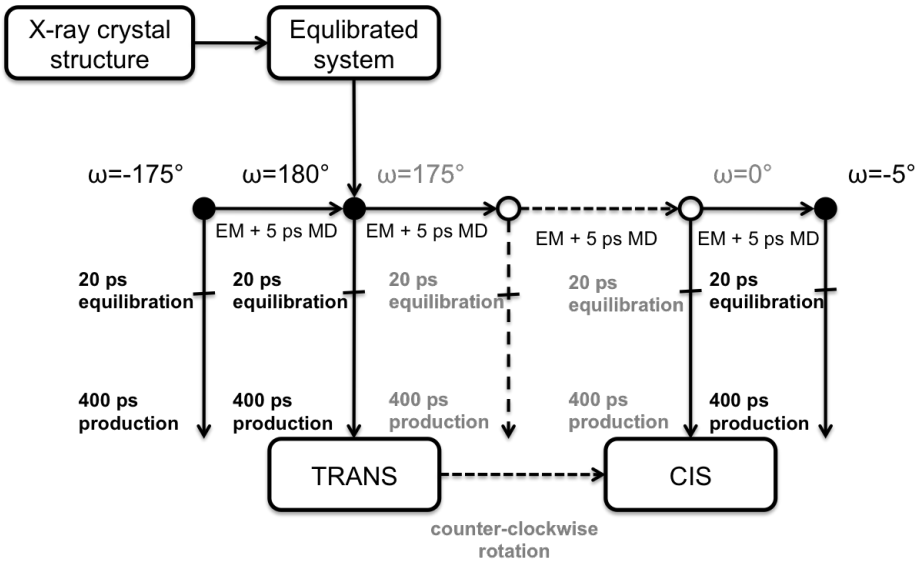

Figure 5.3: **Umbrella sampling methodology for CypA reaction pathway generation**.

5.3 Results

5.3.1 Similarity of Motions in CypA along its functional cycle

We begin by examining the large-scale fluctuations in CypA under three different conditions: (a) reaction pathway, (b) end-state simulations and (c) substrate-free simulations. as illustrated in Fig. 5.4, the fluctuations as determined from the top ten modes show considerable similarity. Observe in Figs. 5.4A, 5.4B, 5.4C that there is inherent flexibility in the regions 12-15, 68-76, 78-84, 101-107, 118-125, and 136-152. While both the reaction pathway and end-state simulations show increased flexibility in regions 12-15 and from 136-152, the substrate-free simulations show enhanced flexibility in the loop 68-76 due to the absence of the substrate in the binding cavity. The increased fluctuations in the regions when the substrate is not bound to the enzyme corresponds to the increase in the degree of freedom available when the binding site is empty. Surprisingly, when the substrate is bound (along the reaction pathway or the end-state simulations), there is a marked increase in the fluctuations of the loops 12-15. The correlation in the fluctuations, as determined from each of the simulations is 0.8, indicating that the internal dynamics as compared in these three scenarios are very similar.

Qualitatively, the regions of flexibility agree with NMR spin relaxation studies as re-

ported by Kern and co-workers [78]. An NMR ensemble for human CypA corresponding to the cis/trans isomerization reaction is not available to perform an analysis similar to ubiquitin as reported above. However, an NMR ensemble for substrate-free CypA is available with 20 structures (PDB code: 1OCA) [211]. Therefore, a comparison of flexibility in substrate-free CypA was performed. The substrate-free simulations of CypA and the NMR ensemble show similar conformational flexibility. As depicted in Figure 5.4D, the regions showing large conformational fluctuations include residues 12-15, 43-51, 66-84, 101-106, 120-126, and 146-151. The QHA modes were computed with five MD simulations, each 2.0 ns in duration; therefore, the total conformational sampling only corresponds to 10.0 ns. However, even at the nanosecond time scale (left), the results qualitatively show agreement with the NMR ensemble (right). On the basis of the insights from the ubiquitin results presented in Chapter 2, it can be envisioned that with more sampling there would be quantitative agreement in the observed flexibility as well.

The extent of conformation fluctuations observed within CypA during the course of the reaction pathway was further characterized by computing the projections of the slowest modes. The projections from the slow modes of QHA from reaction path sampling and NMR ensembles indicate that the structural heterogeneity in the reaction path sampling as well as the substrate-free MD covered and even exceeded the NMR ensemble (see Fig. 5.5). The presence of additional structures within the reaction path sampling ensemble indicates that it covered a larger area within the potential energy landscape, which includes the NMR ensemble as a subset. Even though the NMR projections are localized to a section of the projection maps, it provides an indication that the reaction-coupled modes are sampled by the NMR ensemble. Note that the NMR ensemble does not have the substrate bound; however, as previously reported, these reactions promoting motions are intrinsically associated with the CypA structure and mechanism [78, 9]. In our calculations, the overlap between the reaction path sampling and substrate-free MD was found to be 0.65.

Identification of the reaction-coupled flexibility is computationally expensive, as it requires sampling the entire reaction pathway; moreover, this methodology can be considered limiting if a suitable description of the reaction coordinate is not available. As an alternate methodology, we investigated the QHA modes obtained from the set of CypA-substrate conformations in the end-states only (reactant and product states). Comparison of the aggregated flexibility based on the slowest 10 reaction-coupled modes (see Figure 5.4) as well as the projections of the modes computed from the reaction pathway and the end-states only (reactant + product) indicate that the results are qualitatively similar. A detailed comparison of the individual motions from the two simulations as well as the overlap between the modes further indicated that the slowest reaction-coupled modes from the reaction path sampling and end-state simulations were similar. These results indicate

that the intrinsic flexibility of CypA does not change; however, sampling over the reaction pathway provides the complete extent of conformational fluctuations that are sampled by the enzyme.

It is interesting to note that the end-state MD showed two densely sampled regions within the landscape, whereas the reaction path sampling was more uniform. As could be somewhat expected, the conformations sampled in the end states did not entirely overlap with the reaction path sampling. On the basis of these observations, it appears that the information contained in the end-state simulations is only qualitatively representative of the conformational flexibility within the protein during the course of the reaction step. However, this information is not sufficient to provide quantitative estimates of the actual motions (displacements) along the course of the reaction. Nonetheless, the end-state QHA modes provide a quick way for identification of the reaction-coupled flexibility. It should be cautioned that if the end-states differ considerably, and there is structural rearrangement in the protein and the end-states sample different parts of the phase space, then the qualitative insight obtained may not be representative of the reaction pathway.

5.3.2 Analyzing Spatio-temporal collective dynamics using DTA

Collective behavior in CypA

Using DTA (see Chapter 4), we were able to identify six dynamically coupled regions within CypA (Fig. 5.6(b)). The β-sheet in the protein clustered into two different regions; the first cluster (Fig. 5.6(b)); dark blue; β_{1-2}, β_8) represents the hydrophobic core of the protein. The second cluster (shown in cyan; β_{3-7}) represents the surface region of the protein in close contact with the substrate. The α-helices form two clusters; one constrained via hydrophobic interactions (α_1) and held rigid, whereas the second helix (α_3) exhibits much more flexibility and is coupled with the loops at the active site of the protein (L4, L8). Note that the loops L3, L5, L6 (shown in orange), in front of the active site are all coupled. Similarly, L5 and L7 at the back of the active site belong to the same cluster. The substrate and loop L1 (red) show coupled motions. These observations are in good agreement with previous computational investigations [9, 4].

The biological relevance of our findings can be explained on the basis of the dynamic coupling observed. Based on the clustering, we find loops L3, L5 and L6 (Fig. 5.6(b); loops highlighted in orange) are dynamically coupled. Mutations to loop L5 are known to affect the collective fluctuations in both L3 and L6 [78]. Similar observations can be made about other spatially separated loops. For example, despite the fact that residues from L1 and the substrate (Fig. 5.6(b)) are separated by over 18 Å, clustering indicates that these

two regions are coupled dynamically. Similarly, L4 and L8, which are separated by approximately 12 Å, also exhibit coupled motions. The residues identified to be constrained (Fig. 5.6(a)) are part of a network of conserved residues inter-connected via hydrogen bonds/ hydrophobic interactions. This network has previously been identified via post-processing of this trajectory [9, 4] using other methods. Thus, the clusters of coupled residues identified using DTA provide insights into to the catalytic function of CypA.

Anomaly detection along cypA reaction pathway

The simulation for CypA consisted of generating a free-energy profile using the amide-bond dihedral angle of the bound substrate as the reaction coordinate. We identified a total of 17 out of 180 tensors showing error of reconstructions above the second standard deviation threshold (Fig. 5.7A). The average RMSD was about 1.45 Å. Once again, the error of reconstruction is not correlated with the RMSD (Fig. 5.7B).

In Fig. 5.7C we show two structures to illustrate changes in collective dynamics. In tensor window 10 (top panel), we detect the changes in the collective movements involving L1, L3, L5, L6, L8 and α_2. In tensor window 40 (bottom panel), we found that the substrate molecule showed large deviations with respect to its collective motion with interacting partners, namely β_3 and L6-L8, yet showed an RMSD of only 0.65 Åto its predecessor window. The motions associated with these regions of the enzyme are known to be correlated with the catalytic process [9, 4] and therefore seem to be of functional importance. It is important to note that even though the overall RMSD for CypA in the two windows is relatively small, the substrate peptide highlighted in Fig. 5.7C (bottom panel) shows a distinct conformational rearrangement which could not have been detected by using RMS deviations alone. Since our method incorporates deviations in distance fluctuations, we can keep track of changes that may relate two distally located residues and reason how collective motions along each stage of the reaction pathway changed.

5.4 Conclusions

Our objective in this chapter was to compare and contrast the intrinsic dynamics in cyclophilin A using three scenarios: (a) dynamics along the reaction pathway from $trans$ to cis configuration of the bound substrate peptide, (b) the reactant and product states and (c) just the free enzyme. In all of the three cases, we found undoubtedly a remarkable similarity in terms of the large-scale fluctuations observed. Notably, the topmost (low frequency) modes of motion determined via QHA coupled to the reaction shows substantial overlap

with respect to the large scale fluctuations in the reactant and product state simulations and just the free enzyme itself. This can be seen from the regions exhibiting substantial fluctuations (as illustrated in Fig. 5.4) and the projections spanned by these modes.

Does this mean that the intrinsic dynamics does not change in the three scenarios? As evidenced in Fig. 5.4, there are subtle variations in the fluctuations observed from the top ten modes. For one, the region 68-76 exhibits considerable fluctuations in unbound simulations. This is to be expected since the binding of the substrate to the active site causes specific changes that seem to restrain the additional degrees of freedom in and around this region of the enzyme. A second interesting observation is that the residues 12-15 in the reaction pathway as well as the end-state simulations show considerably higher fluctuations compared to that of the unbound simulations. Further enhanced are fluctuations in the regions 146-154, all of which are involved in 'promoting' the progress of the catalytic step [77, 78, 9, 5].

There is substantial overlap between the reaction coordinate simulation and just the end-states alone. Although the simulations of the substrate-unbound simulations also show substantial overlaps in terms of the large-scale motions, the end-state (and the substrate-free) simulations do not "cover" all of the landscape seen from the free-energy profile. The smaller overlaps between the full free-energy reaction profile versus 0.67 for end-state and 0.65 for substrate-free, further allows one to understand the subtle changes the enzyme undergoes upon substrate binding and during catalysis. We also repeated our experiments with QAA (not shown) and find that the reaction pathway samples regions in and around the transition states which are not accessible to the substrate-free or end-state simulations.

The analysis from DTA leads to characterize the dynamical coupling that exists between different regions of CypA. An interesting observation here is that the loop L1 (see Fig. 5.6(b)) and substrate are coupled in terms of their dynamical behavior in spite of being separated by over 20 Å. Further more, collective dynamical changes occur along the network regions implicated in previous studies [9] and therefore it highlights the importance of having access to such states during the course of a simulation. These states might not correlate with states showing high RMSD values.

In summary, reaction pathway sampling such as umbrella sampling in combination with QHA and DTA provides a unique method of identifying slow conformational flexibility coupled to the enzyme mechanism. Similar to the case of ubiquitin, these modes also show agreement with the experimentally determined protein flexibility. For enzyme catalyzed reactions, the reaction-coupled flexibility can be qualitatively understood by QHA on a set of conformations obtained by a combination of the reactant and product states only. This allows overcoming a limitation of this methodology, as it requires sampling along the reaction pathway. The slow conformation fluctuations based on the entire re-

action pathway and the end-states only were qualitatively similar for the CypA catalyzed reaction. The coverage of conformational landscape was similar for the first four modes coupled to the reaction. From a mechanistic point of view, it could be rationalized that when the reactant and product states of the enzyme do not show large conformational changes involving rearrangements of secondary structures, sampling just the reactant and product states provided qualitative information about the large-scale motions involved in the catalytic substep. Overall, the characterization of microsecond to millisecond conformational fluctuations of ubiquitin and CypA indicates that flexibility is closely related to their designated function. Even though the role of protein flexibility in function is the topic of an ongoing debate [210] similar to the observations reported here, increasing evidence from other groups also suggests a close link between protein conformational fluctuations and enzyme catalysis [40, 47, 156].

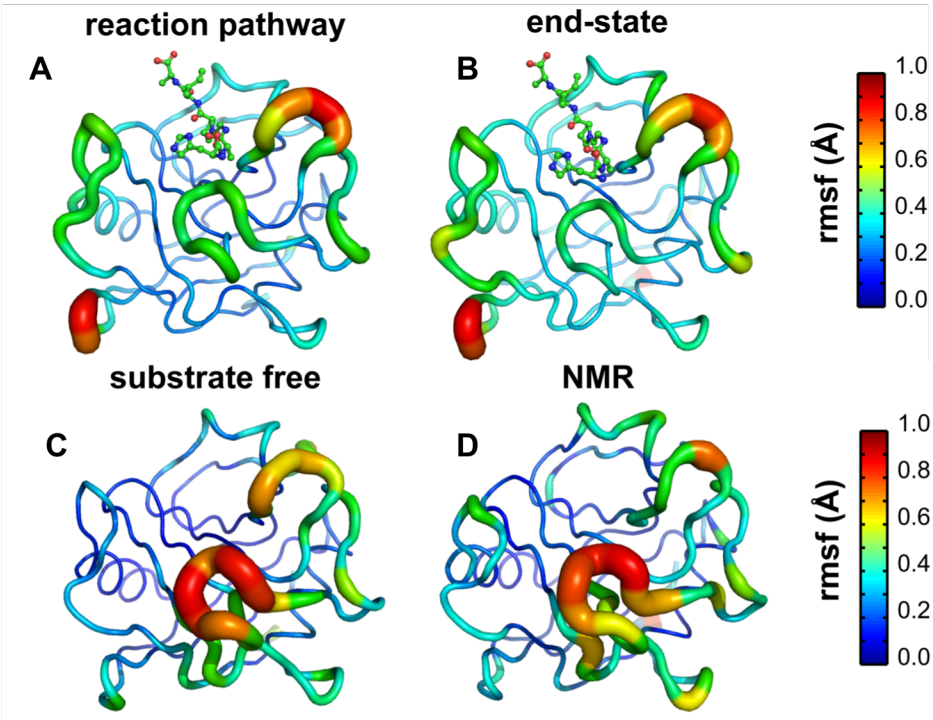

Figure 5.4: **Reaction-coupled flexibility in the enzyme CypA shows similarity to motions in the end-states and substrate free enzyme**. Inverse frequency-weighted positional fluctuations for the top 10 reaction-coupled modes based on QHA of conformations from (A) the entire reaction pathway and (B) end-states only (reactant + product). The substrate is shown in a ball-and-stick representation, while the enzyme is shown as a cartoon. Similarity in fluctuations is also seen in (C) substrate-free simulations and (D) NMR ensemble 1OCA [211] which qualitatively agree with the full reaction pathway.

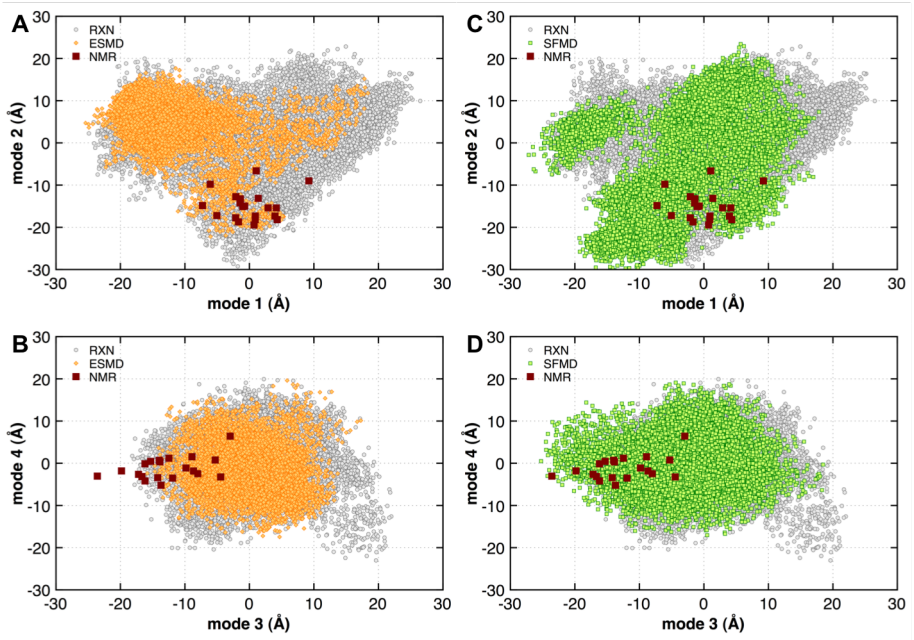

Figure 5.5: **Projection of MD and NMR structures on the modes coupled to the reaction catalyzed by CypA show large overlaps.** (A) mode 1 vs mode 2; (B) mode 3 vs mode 4. The gray open circles correspond to the projections from the set of 18,500 conformations sampled along the reaction pathway; red squares correspond to the NMR structures. The yellow filled circles correspond to the end-states MD only. (C) mode 1 vs mode 2; (D) mode 3 vs mode 4. The green squares correspond to substrate-free simulation for CypA. The large extent of overlaps in each of these simulations indicate how reaction path sampling and the other simulations share substantial similarity in individual motions. Note that the computed projections represent summation over all atoms in the protein.

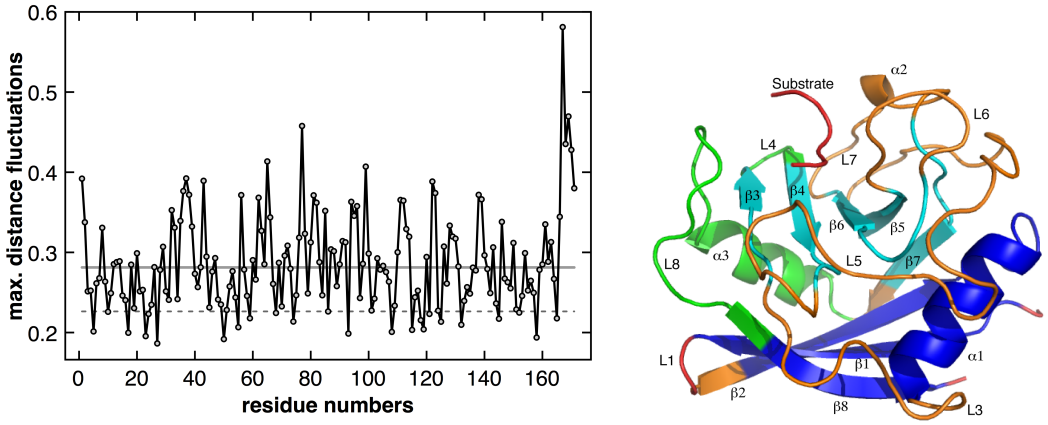

Figure 5.6: **Dynamically coupled regions in Cyclophilin A**. The left-hand panel identifies a total of 26 residues that are constrained. These residues are conserved and form a network of interactions stretching from the outside of the enzyme to the active site, as proposed by Agarwal and co-workers [9, 4]. Six regions of the protein dynamically coupled; the hydrophobic part formed by $\beta_{1-2,8}$ and α_1 experience low distance fluctuations (shown in dark blue); β_{3-7} (top middle panel) undergo slightly larger distance fluctuations (shown in cyan); α_3, L4 and L8 are grouped into one cluster (green); α_2, L5 and L7 are regions behind the substrate, L3, L5 and L6 cluster together into one group (front of the substrate shown in orange) and L1 and the substrate form the most flexible parts of the protein (shown in red).

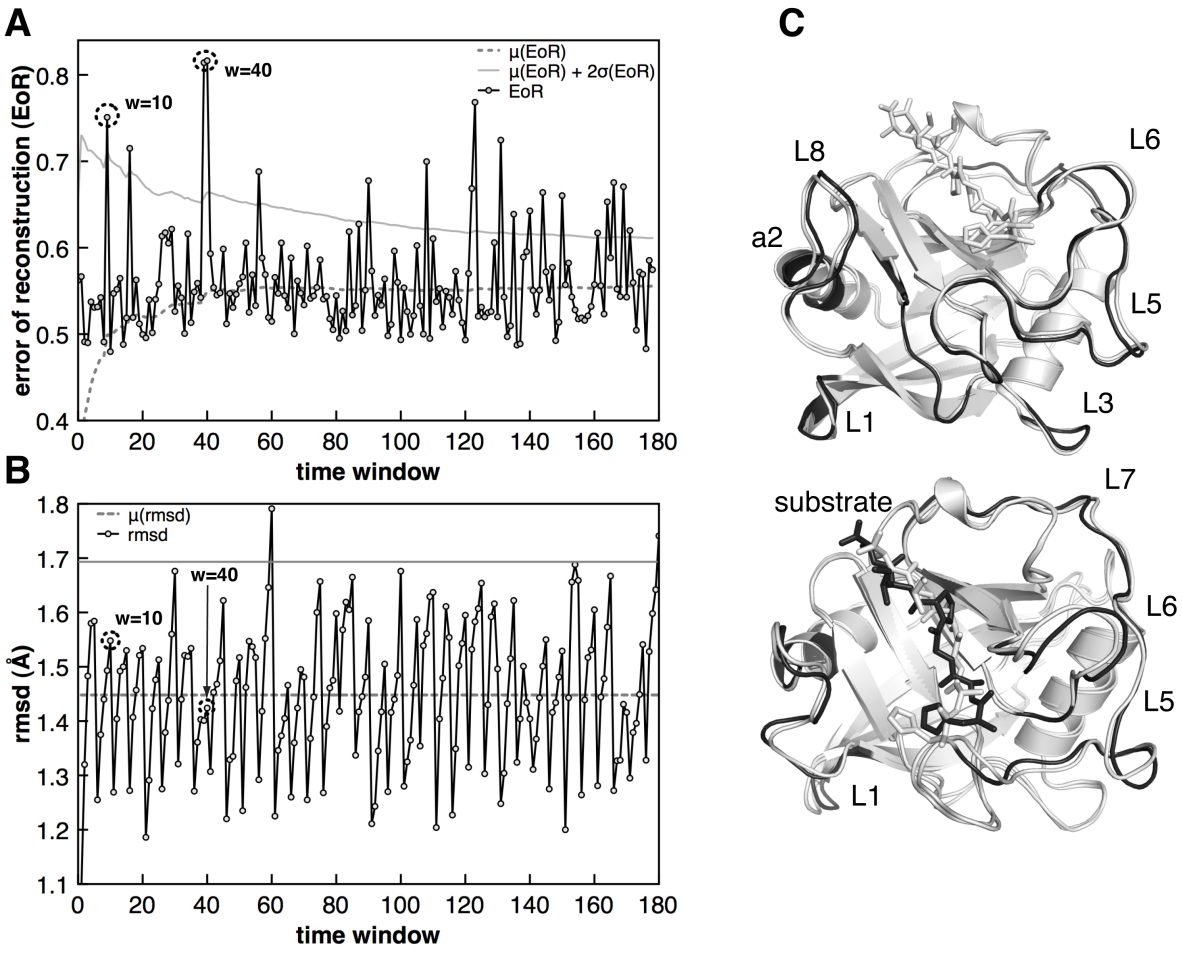

Figure 5.7: **Error of Reconstruction along Cyclophilin A reaction pathway compared against the root mean squared deviations (RMSD)**. (A) shows the reconstruction error plotted as a function of tensor window. Gray dotted line: mean reconstruction error, gray solid line: second standard deviation interval. Dotted circles highlight those tensor windows with higher reconstruction errors shown in the adjacent panel. (B) RMSD showing structural changes along the reaction coordinate defined by Agarwal and co-workers [9]. (C) Structural changes associated with CypA for two windows, namely $w = 10$ (top; overall RMSD 0.617 Å) and $w = 40$ (bottom; overall RMSD 0.650 Å). The structure from the predecessor window is shown in dark gray and the current window is shown in light gray; regions involved in collective movements highlighted in green. Note the large movement in the substrate molecule, shown as sticks in the bottom panel. This cannot be picked up using traditional metrics such as RMSD since it is only an average measure of structural deviations. However, tracking distance variations, we note a significant difference in the placement of the substrate.

Chapter 6

Evolutionary Linkage between Enzyme Fold, Flexibility and Catalysis

Enzymes are intrinsically flexible molecules. Emerging evidence has linked internal protein motions to the mechanism of enzyme catalysis; networks of coupled promoting vibrations/motions have been discovered for several enzyme systems. In this report, we investigate the linkage between the protein flexibility, enzyme function and the overall shape of the enzyme (enzyme fold). Three chemical reactions are studied: (a) isomerization, (b) hydride transfer (reduction) and (c) hydrolysis of single strand RNA molecules. For each reaction mechanism, we ask if two different folds catalyzing the same reaction show any similarity in terms of the intrinsic dynamics coupled to the catalytic step. Our studies show that there is a remarkable similarity in the dynamics of enzymes even if they do not share significant structure or homology. Three enzymes in particular, cyclophilin A, dihydrofolate reductase and ribonuclease A show the conservation of reaction coupled slow conformational fluctuations across multiple species, particularly in the surface loop regions. These highly flexible regions in the enzyme are interconnected by conserved networks, which connect these regions to the crucial active-site residues. Detailed characterization indicates reaction coupled flexibility as well as the network residues, along with the crucial structural interaction in the active-site, are important aspects of the enzyme fold. These findings also explain why there is drastic impact on the enzyme activity when these crucial linkages, even located far away from the active-site, are altered.

6.1 Introduction

Proteins are intrinsically flexible molecules. The relevance of conformational flexibility or multiple conformations of proteins, with small deviations from native state, in proteins designated function is the topic on an ongoing and interesting debate [98, 41, 210]. The role of structure in protein function such as enzyme catalysis is well established. Techniques including X-ray crystallography and nuclear magnetic resonance (NMR) have been widely used to obtain information about the protein structure, thereby providing insights into the mechanism of function. The obtained information reveals the functioning protein present in slightly different but related conformations, with some areas of the protein (particularly the surface loops) being more flexible than others. Given the success of structural effects in explaining many aspects of function, the observed fluctuations in the structure have largely been ignored. More recently, however, it has been proposed that the protein function may involve multiple conformations; the observed deviations are not just inconsequential random thermodynamic fluctuations but flexibility may be closely linked to protein function including the catalytic efficiency of enzymes [41, 113, 6, 48, 229].

Internal proteins motions span a wide range of length and time-scales. The dynamical landscape of a protein and the associated energy landscape have been challenging to characterize, as the internal motions and the associated structural deviations over a broad range of time-scales [52]. Experimental techniques including, but not limited to, NMR spin relaxation and single molecule experiments have provided insights into the protein motions that occur at time- scale of the enzyme catalytic step in several enzyme systems including dihydrofolate reductase (DHFR) [47, 236] and cyclophilin A (CypA) [77, 78]. The time-scales for the slow conformational changes and the chemical step catalyzed by the enzyme are similar, therefore, raising the question whether they are inter-related or not [47].

Preliminary evidence has suggested the possibility of protein dynamics playing a promoting role in the biophysical mechanism of enzymes [40, 115]. Conformational dynamics of enzymes have been associated with substrate (and cofactor) binding and product release for some time now; however, the interconnection between flexibility and the substrate turnover step still remains the topic of debate. In the enzyme CypA, Kern and coworkers monitored the motions of several surface loop residues only in the presence of substrate [77]. Agarwal and coworkers, performed computational studies of CypA, identified a network of protein residues that influenced the reactive trajectories in the active-site [9, 4, 5]. For the hydride transfer catalyzed by DHFR, groups of Benkovic, Wright, Hammes among others have indicated the movement of surface loops Met20 and βF-βG in associated with hydride transfer [236, 7, 8].Using computational methods, Hammes-

Schiffer and coworkers have identified a network of coupled protein motions linked to enzyme function in DHFR. These networks formed by conserved residues both in and distal to the active- site have been implicated in promoting the catalytic step. These networks extend from flexible surface loop regions, showing high conformational flexibility, all the way to the active-site residues that directly participate in the chemical step. It has been hypothesized that the solvent shows motion and energetic coupling with flexible surface loops which eventually transfer the required kinetic energy to the active-site through the conserved network interactions [6]. Evidence from computational studies and Mossbauer and neutron scattering investigations also supports that the thermo-dynamical fluctuation in hydration-shell and bulk solvent control the behavior of reactive trajectories; [5, 85]as well as the motions within the active-site residues control the chemical environment, which favors the reaction crossing the energy barrier [40].

Conservation of structural features across species has provided vital clues to their role in protein function. In particular, it has been argued that the enzyme active-site residues are optimally arranged to provide complementary environment to the transition state to allow for its stabilization [217, 281]. The overall enzyme shape or the enzyme fold has been suggested as a scaffold that orients the active-site residues that are conserved features of the enzyme structure. This notion has lead to the structure-encodes-function paradigm for understanding enzyme catalysis. As a result, a number of theories for understanding enzyme catalysis have been proposed with strong emphasis on the structural interactions between enzyme and the substrate in the active-site.

More recently, investigations have offered new insights into the linkage between enzyme flexibility and designated function. In particular, Klinman and coworkers have reported the reciprocal flexibility in enzyme fold: the active-site shows considerable rigidity while the external regions show flexibility in linkage with enzyme function [156]. We hypothesize that the argument of conservation of important structural features can also be extended to identification of protein flexibility in interconnection to enzyme function. Similar to the secondary structure elements as well as individual residues that are conserved for their structural role, we suggest that the chemistry promoting dynamical regions of enzymes are also conserved as a part of the enzyme fold.

In this chapter, we describe our investigations of protein dynamics connected to enzyme catalysis to test the evolutionary preserved interconnection between the enzyme fold, conformational flexibility and function. We have investigated three well characterized enzymes catalyzing different chemical reactions with distinct folds and reaction mechanisms: CypA catalyzing cis/trans isomerization of peptidyl-prolyl amide bond in peptides/proteins; DHFR catalyzing hydride transfer and ribonucelase A (RNaseA) catalyzing the hydrolysis of single-stranded RNA. For each enzyme fold, computational in-

vestigation of enzyme structures from several species with different sequences has been performed. Slow conformation fluctuations at the time-scale of the reaction and spanning the entire enzyme structure have been characterized. One may argue that this result might not be surprising given the high structural homology between these proteins.

In order to mitigate the effects of structural homology, we compare the reaction coupled flexibility identified across the three reaction mechanisms across multiple folds: for peptidyl-prolyl isomerization, we consider CypA and Pin1 catalytic domain; for hydride transfer, we compare Ec DHFR with R67-DHFR (an older mitochondrial enzyme) and finally for hydrolysis, we compare RNaseA with the function of angiogenin. Note that in isomerization and hydride transfer, the two enzymes considered do not share sequence/ structural homology. In the case of hydrolysis, both proteins RNaseA and angiogenin share 33% sequence homology and a significant structural homology [3]; however, angiogenin is over 6 orders of magnitude less efficient than RNaseA [240, 241]. In all the three cases, we find that reaction coupled dynamics exhibits remarkable similarity, even if structures do not share significant homology. However, when the network of interacting residues are not present, the activity of the enzyme is severely affected. Thus, the presence of this networks along with intrinsic motions associated with the protein fold are important for catalytic activity.

6.2 Methods

The starting structures for the enzymes were obtained from the protein data bank, according to accession codes mentioned in the main text. Enzyme-substrate complex were modeled using molecular mechanics under explicit solvent conditions as previously described [9]. AMBER simulation package was used for model building and simulations. AMBERs parm98 force-field and SPC water model were used. After the model preparation enzyme-substrate in explicit water, the system was equilibrated based on protocol described in Chapter 2. Briefly the model was minimized to remove bad contacts and slowly heated to 300K. All production runs were performed at 300K under NVE conditions. Note, previously we have verified the suitability of parm98 force-field for dynamics modeling with comparison with other popular force-fields [9].

6.2.1 Reaction profiles

cis/trans Isomerization: The cis/trans isomerization catalyzed by CypA was modeled for three structures from Plasmodeum yoelli (PDB code: 1Z81 [267]), Bos taurus (1IHG [256])

and Homo sapiens (1AWQ [262]). The human cyclophilin was modeled as previously described with peptide substrate His-Ala-Gly-Pro-Ile-Ala. For the B. taurus cyclophilin, only the residues 2-185 (corresponding to the CypA fold) in the PDB file were used for the model; a substrate peptide Ala-Gly-Pro-Phe was modeled on alignment of active-site residues from human CypA. For P. yoelii cyclophilin, only the residues 40-210 in the PDB file were used for the model; a substrate peptide His-Val-Gly-Pro-Ile-Ala was modeled on alignment of active-site residues from human CypA.

The reaction pathway was modeled with amide bond dihedral angle (ω) as reaction coordinate; 37 windows (in 5° decrements) were used to map the reaction from the reactant state (ω = 180°) the product state (ω = 0°). Reaction profiles were generated for the cis/trans isomerization reaction catalyzed by CypA as described previously [9] and in Chapter 5. The amide bond dihedral angle of the isomerized peptide bond was selected as the reaction coordinate and umbrella sampling was performed on the entire section of the reaction pathway. No covalent bonds are broken or formed during this reaction mechanism. Previously, we have used this methodology to investigate isomerization in several human CypA structures [4]. Each window was simulated for 200 ps and 500 structures from each MD simulation were collected for the QHA and correlated motion analysis. 18,500 conformations were used for computing the QHA modes and doing DTA.

Hydride Transfer: For the hydride transfer catalyzed by DHFR, which involves breakage and formation of C-H bonds, empirical valence bond (EVB) approach [276, 272] with a two state model as implemented in the AMBER package was used. The simulation protocol, including the parameters for the Morse potential for the C-H bonds, was same as the previous studies of hydride transfer in DHFR simulated using EVB [7, 8]. The hydride transfer catalyzed by DHFR was simulated for four structures including the Escherichia coli (PDB code: 1RX2 [234]), Mycobacterium tuberculosis (1DG5 [183]), Candida albicans (1AI9 [280]) and Homo sapiens (1KMV [157]). Based on previous evidence the associated proton transfer was assumed to preeced the hydride transfer step. The substrate was modeled as protonated dihydrofolate with the cofactor nicotinamide dinucleotide phosphate (NADPH) as the hydride donor as reported in previous study.

For modeling the hydride transfer step, we used protonated substrate and empirical valence bond (EVB) method. The procedure for EVB is described previously, was used for mapping the reaction from the reactant state (hydride on NADPH) to product state (hydride on DHF). Morse potential was used for the breaking C-H and forming C-H, with the following values of the C-H Morse parameters D_e= 103 kcal/mol, R_e =1.09 and α=1.285 Å$^{-}1$, based on previous studies [7, 8]. The parameter α was adjusted according to the relation $\alpha = \sqrt{k/2D_e}$ for the parm98 force-field harmonic bond force constant (k) to

reproduce the C-H-C geometrical parameters as previously reported. A mapping potential in the EVB framework was used to gradually map the reaction from the reactant state (mapping parameter λ=0) to the product state (λ=1), with 21 simulations (λ=0, 0.05, 0.10, . . ., 0.95 and 1). Each of these simulations was equilibrated by energy minimization and followed by 70 picoseconds (ps) of molecular dynamics (MD). For production runs,100 ps of MD under constant energy conditions (300K and 1 atm) were performed. Other conditions were similar to previous study mentioned above. For each of the 21 bins, 1000 conformations were collected representing the enzyme-substrate conformations sampled along the reaction pathway. Thus, 21,000 conformations were used for QHA and DTA.

Single-strand RNA hydrolysis: One may observe that the reactant and product state for RNase A is quite similar. In this scenario, we decided to model only the reactant and product states (5 nanoseconds of MD simulation in each state) for the structures from Bos taurus (PDB code: 1U1B [35]), Rana catesbeiana (1KVZ [133]) and Rattus norvegicus (1RRA [107]). A summary of the sequence identity and the structural similarity for the enzyme structures used in this study is given in Table 1. Further details of the simulation protocol are available in the supporting information.

	CypA			**DHFR**				**RNase A**	
	1IHG	1Z81		1DG5	1AI9	1KMV		1KVZ	1RRA
1AWQ	63[a]/ 29.5[b]/ 1.1[c]	54/ 28.1/ 1.1	1RX2	36/ 23.6/ 1.5	31/ 20.2/ 2.2	30/ 21.1/ 1.8	1U1B	26/ 11.4/ 2.4	67/ 22.2/ 1.1
1IHG		58/ 29.1/ 1.4			28/ 18.9/ 2.6	31/ 21.2/ 1.7	1KVZ		27/ 11.3/ 2.4
			1AI9			35/ 22.1/ 2.3			

Table 6.1: **Sequence and structural comparison the enzymes investigated**. Alignments performed with DaliLite pair wise comparison web tool: http://www.ebi.ac.uk/DaliLite/ [132]. See text for the PDB accession codes. a=sequence identity(%), b=Z-score, c=RMSD in the reference PDB structures (Å)

Reaction coupled flexibility: Protein flexibility at long time-scales was identified using the quasi-harmonic analysis (QHA) [150] as described previously in Chapter 2. QHA of entire set of enzyme-complex conformations sampled along the reaction pathway allows identification of conformational fluctuations occurring at the time-scale of the reaction. As identification of the slow conformational fluctuation is central to our investigations, we have verified the ability of QHA to reproduce experimentally observed protein conformational fluctuations at the microsecond-millisecond time-scales [225]. In case of CypA, we used 18,500 conformations while in case of DHFR we used 21,000 conformations collected along the entire reaction profile. This approach provides reproducible modes associated with long time-scales, as described in our recent study. For RNaseA, only the reactant and product states were used for QHA with a total of 20,000 conformations. Our detailed analysis indicates that for reactive states where the protein conformations do not differ considerably, the slow modes can be reasonably approximated by QHA of the conformations from just the reactive and product states only as observed from Chapter 5. In all cases we analyzed the first 50 modes from QHA. In CypA, the modes coupled to the reaction were defined by computing the variation in the amide bond dihedral angle. The largest fluctuations of the angle in the slowest modes were assigned the largest coupling to the reaction. In DHFR, the coupling was defined as the dot product of the hydride transfer displacement vector in the eigen-modes with the donor carbon (CD) and the acceptor carbon (CA) distance vector. In the case of RNaseA, only the top 10 slowest modes were analyzed.

Network of protein vibrations/motions were identified by characterization of enzyme regions displaying large movements in the QHA modes, by investigating the dynamic cross correlation maps, and monitoring the distances of correlated regions over the course of reaction as reported previously [9]. Genomic analysis was performed using Clustal W [257] and the structural analysis was aided by the PyMOL program [73].

6.3 Intrinsic Dynamics and Catalysis across Multiple species

6.3.1 Cyclophilin A

CypA is a prolyl-peptidyl isomerase (PPIase) catalyzing the cis/trans isomerization of peptide bonds in small peptides and proteins. In the previous chapter (Chapter 5, we described in detail the molecular architecture and the reaction coupled flexibility in CypA. We noted that the reaction coupled dynamics with multiple substrates bound to CypA [9] and just the unbound enzyme showed high overlaps in terms of the large-scale fluctuations observed

from simulations. In this chapter, we examine the large-scale collective dynamics coupled to the reaction coordinate across three species first, and then in a completely different enzyme, Pin1, which shares neither structural/ sequence homology with CypA.

Characterization of the reaction coupled conformational flexibility from *Bos taurus* and *Plasmodeum yoelli* have revealed remarkably similar flexibility for CypA across species including human CypA (see Fig. 6.1). The motions of the enzyme residues Ala101-Asn102-Ala103 and nearby loop 105-108 as well as Arg55 in human CypA alter the crucial enzyme-substrate interactions during the reaction pathway. Additionally, motions of Phe60 in the active-site and the associated loop region 57-60 also make important contributions of the reaction mechanism. In *B. taurus* the equivalent residues Ala121-Asn122-Ala-123 and the loop 125-127 also shows large movement during the enzyme catalysis, with conserved active-site residues Arg75 and Phe80 also displaying motions along the reaction pathway. Similarly, in *P. yoelii*, the residues Ala143-Asn144-Ser145 and the loop 147-149 also display large movements during the course of the reaction, with active-site residues Arg97 and Phe102 controlling the crucial enzyme-substrate interactions.

Several regions distal to the active-site also display similar type of dynamical motions in the slowest modes coupled to the isomerization reaction. These regions highlighted in Figure 6.1 include the highly flexible surface loops (for e.g. the region 82-88 in human CypA). Careful comparison of the regions with high flexibility in these reaction coupled modes has led to the discovery that the slowest 3 protein vibrational modes coupled to the cis/trans isomerization are conserved over evolution (see Figure 6.1A). These three reaction promoting modes show absolute similarity in the location of large flexibility in the distant areas of the enzyme, even though the protein structures are from different species.

To rule out the possibility of biasing the active-site dynamics, 3 different substrates were used in these simulations. Further, in previous investigation of human CypA with a biologically relevant substrate also provided the same vibrational modes coupled to the reaction [4]. The regions of the protein showing displacement within these modes have been indicated to show increased fluctuations in presence of substrate as measured by N15 spin relaxation experiments [77, 78]. In spite of having different substrates bound at the active site, it was also observed that the correlated motions across the three different species were extremely similar (see Fig. 6.2).

Characterization of intrinsic CypA flexibility in the structures from different species shows the presence of regions of similar dynamical fluctuations, as indicated by a k-means clustering analysis of protein regions (see Fig. 6.3). The proteins residues separate into 6 clusters based on their characteristic flexibility over the course of isomerization; these include the β-sheet separating into two clusters, the two large helices in separate clusters, while the flexible loop regions forming two additional clusters. Structural analyses of the

Region	H. sapiens	B. taurus	P. yeolii
I1	29-33/ 85-86	40-44 / 104-105	27-31/ 88-89
I2	34-36 / 77-78	45-47 / 96-97	32-34 / 80-81
I3	56-57 / 142-150	75-76 / 155-162	60-61 / 143-151
I4	82-85 / 104-105	101-104 / 123-124	85-87 / 107-108

Table 6.2: **Regions in CypA that exhibit correlated motions.**

I	D13N-K155O	N35N$_{\delta 2}$-G109O	I56N-G150O	A101N-Q111O	F83N-N108O
II	G21N-D164O	N43N$_{\delta 2}$-G117O	V64N-D159O	A109N-Q119O	F91N-N116O
III	D16N-D167O	N38N$_{\delta 2}$-G120O	V67N-N162O	A112N-Q122O	F94N-N119O
IV	G25N-L175O	N47N$_{\delta 2}$-G129O	I76N-E170O	A121N-Q131O	F103N-N128O
V	N7N-I156O	N26N$_{\delta 2}$-T95O	V44N-D149O	A86N-Q97O	I68N-A94O

Table 6.3: **Network Residues in CypA and the conserved linkages.** The corresponding hydrogen bond linkages are shown for each of the five species: human CypA (I) PDB code: 1RMH; human cyclophilin B (II) PDB code: 1CYN; B. Malayi (III) PDB code: 1A33; B. Taurus (IV) PDB code: 1IHG; E. coli (V) PDB code: 2NUL.

PPIase fold indicates the regions showing reaction coupled flexibility are interconnected by network of hydrogen bonds; the residues and interaction (hydrogen-bonds) forming these networks are conserved over evolution (Fig. 6.1C). These interactions originate in the highly flexible surface loop regions on opposite sides of the protein (Phe83N-Asn108O, Asn35$_{N\delta 2}$-Gly109O and Asp13N- Lys155O in human CyPA), and pass through internal regions (Ala101N-Gln111O and Ile56N-Gly156O) to eventually connect to the residues involved in structural contacts with the substrate (Arg55, Phe60, Asn102 and Ala103). It is interesting to note that even though the exact residue is not conserved, the linkage at the location is conserved keeping the network intact. The presence of regions with similar dynamical characteristics in distal areas of the protein, the preservations of these clusters as well as conservation of the linkages connecting the flexible clusters to the active-site regions form an important feature of the enzyme fold.

6.3.2 Dihydrofolate reductase

DHFR catalyzes the catalyzes the reduction of 7,8- dihydrofolate (DHF) to 5,6,7,8-tetrahydrofolate (THF) using nicotinamide adenine dinucleotide phosphate (NADPH) as a coenzyme. DHFR

belongs is a member of family of proteins sharing the nucleotide binding Rossmann fold, characterized by a central core formed β-sheet surrounded by α-helices; in DHFR there are two occurrences of the Rossmann fold fused in the same peptide chain. Previously, a network of coupled motions promoting hydride transfer in DHFR has also been identified using detailed theoretical and computational modeling. Similar to the network in CypA, this network is also formed by surface residues present on the flexible loop regions (particularly the βF- βG and the Met20 loop) interacting with other conserved residues all the way to the active-site. The detailed characterization of this network had lead to the identification of a chain of residues, as a dynamical contributor to hydride transfer reaction. Previous investigations had also connected the dynamics of some of these residues to the chemical step during enzyme function. The presence of this network has also been confirmed by several studies including experimental investigations.

Conformational flexibility linked to the hydride transfer catalyzed by DHFR from E. coli, M. tuberculosis, C. albicans, and humans show considerable movements in the surface loop regions. The k-means clustering method indicates that over the course of hydride transfer reaction the dynamical motions of the β-strands separate into 2 clusters. There are 3 additional clusters: the Met-20 and βF-βG loops; adenosine binding domain; and the substrate binding pocket. This clustering was same in all the 4 species investigated. The slow protein vibrational modes coupled to the hydride transfer reaction show conservation across species with different structures (see Figure 6.5A). The most characteristic feature of the slowest vibrational modes are the large activity in the surface loop Met20 and βF-βG loops as well as the substrate binding pocket which is close to the DHF aromatic ring. These regions have impact on the reaction by positioning the nicotinamide ring of the cofactor in close proximity of the substrate ring. Similar to CypA, DHFR from different species also shows the complete similarly of motions in the equivalent regions (see Fig. 6.5C). These regions contain the residues that form the network of coupled protein motions including the Tyr100, Ile14 and the Phe31 in E. coli DHFR. Additionally residue Arg57 shows concerted movement with DHF tail. Structural analysis indicated that these residues and the interactions are also conserved over evolution and display same motions along the reaction pathway in M. tuberculosis (Tyr100, Ile14, Phe31, Arg60), C. albicans (Tyr118, Ile19, Phe36, Arg72) and human (Tyr121, Ile16, Phe34, Arg70) enzyme (Fig. 6.8).

The interconnection between the DHFR fold and the reaction coupled flexibility is similar to CypA. The clusters of flexible surface loops are connected to the active-site residues through the preserved linkages (Fig. 6.5C). Particularly the surface hydrogen bond Asp122-Gly15 in E. coli, Asp126-Gly15 in M. tuberculosis, Asp146- Gly20 in C. albicans and Asp145-Gly15 in human is conserved. Therefore, these regions also form

Region	E. coli	M. tuberculosis	C. albicans	H. sapiens
I1	15-22 / 116-125	15-22 / 116-125	16-26 / 140-150	14-24 / 142-149
I2	31-36 / 142-150	31-36 / 142-150	36-45 / 178-186	34-43 / 170-176
I3	64-72 / 142-150	66-74 / 142-150	78-90 / 178-186	76-87 / 170-176

Table 6.4: **Regions in DHFR that exhibit correlated motions.**

E. coli	D122N-G15O	I14-Y100	Y100-NADPH	F31-DHP
M. tuberculosis	D122N-G15O	I14-Y100	Y100-NADPH	F31-DHP
C. albicans	D146N-G20O	I19-Y118	Y118-NADPH	F36-DHP
H. sapiens	D145N-G19O	I16-Y121	Y121-NADPH	F34-DHP

Table 6.5: **Network Interactions conserved in DHFR.**

characteristic feature of the enzyme fold with implications for catalysis. Note that the importance of Ile14, a non-active-site residue but a part of the network contributing to the enzyme mechanism, has been confirmed by NMR and mutation studies [47]. Mutation in the network residues is known to alter the reaction rates for the hydride transfer. Further, Met20 loop of E. coli DHFR, particularly is known to exist in 2 conformations, occluded and closed and has been implicated in the catalytic step, mutations around these regions lead to considerable decrease in catalytic efficiency. Single molecule experiments have also suggested the concerted movement of this loop with the hydride transfer.

Examining the patterns of correlations across the multiple DHFRs (as shown in Fig. 6.6), further reveals how similar these motions are during the reaction pathway. As clearly observed, motions along βF-βG are highly correlated with motions in the Met20 loop, where as motions in residues 64-72 are negatively correlated with motions in βG-βH. It is significant to note that residues 64-72 and βG-βH are separated by over 30 Å, implying some form of long-range connectivity in the dynamical sense. Further, it is also possible to compare the dynamical behavior by examining the clusters as examined via DTA (see Chapter 4). DHFR is comprised of five dynamically coupled regions (see Fig. 6.7). The core β-sheet separates into two clusters (shown in blue and cyan respectively) where as the Met20 and βF-βG loops are clustered together (shown in orange). α-helix consisting of the residue Phe31 forms a distinct cluster (yellow) as well as αD (consisting of residue Tyr100; shown in green) indicate that these regions exhibit different dynamics compared with the rest of the protein. While the hydrophobic core of the protein (shown in blue and cyan) exhibit slower motions, the αA, αB and αD show higher levels of fluctuations. This pattern of coupling is also seen across all the four members examined.

Region	B. taurus	R. norvegicus	R. catesbeiana
I1	15-19/ 79-83	15-19/ 79-83	14-17/ 63-67
I2	62-76/ 40-51	62-76/ 40-51	49-56/ 31-38
I3	86-97/ 100-105	86-97/ 100-105	73-78/ 82-86

Table 6.6: **Regions in RNaseA that exhibit correlated motions.**

6.3.3 Ribonuclease A

RNaseA is secreted by pancreas and catalyzes the hydrolysis single strand RNA (see Fig. 6.9A). RNaseAs characteristic shape is formed by β-sheet in the center surrounded by several flexible loop regions and α-helices. NMR experiments have suggested the link between flexibility and function [65, 162, 163]. The active-site is located at the bottom of the inverted β-sheet. A distinctive feature of this fold, different from CypA and DHFR, is the linkage of the flexible surface loops through disulphide linkages (Fig. 6.9C). The dynamical clustering shows the separation of the residues into 4 clusters (the β-sheet forming two clusters and the two loop areas forming two additional clusters). Reaction coupled flexibility is also preserved over evolution (Fig. 6.11). In particular, the slowest modes across the 3 species investigated show the displacement in the surface loop areas near the active-site as well as distal to the active-site. Further, one can also note that there is distinct similarity in the motions as noted by the correlated motions (Fig. 6.10).

A network of interactions coupled to catalysis and connecting the regions of high flexibility is also present in RNaseA (Fig. 6.9C). Similar to the CypA and DHFR network, this network is formed by connection of the surface loop regions all the way to the active-site. In RNaseA, the highly flexible surface regions are linked to other loops through disulphide linkages (Cys26-Cys84, Cys40-Cys95, Cys58-Cys110 in B. taurus) and a hydrogen-bond (Tyr97OH- Lys41O). Conserved residues in active-site (His12 and His119) mediate these network motions between the enzyme and the substrate. These linkages are conserved over evolution as a part of the enzyme fold (Fig. 6.12).

6.4 Diverse Enzyme folds catalyzing same Chemistry

Enzyme structures with high degree of sequence similarity are expected to show similar intrinsic flexibility due to similar molecular architecture. It could potentially be argued, against the findings reported here, that the conservation of flexibility could purely be a coincidence due to the similarity in shape. Note that the conserved flexibility discussed in

this report focuses not only on the slowest conformational fluctuations but on the global conformational fluctuations that are coupled to the enzyme catalysis step, which may not necessarily be the same as intrinsic slow movements of the enzyme folds. More importantly, a test of the hypothesis of inter-connection between the flexibility and function can come from comparison of enzyme folds catalyzing the same chemistry but different structural folds. Therefore, we have closely inspected the reaction linked flexibility, if any, for 3 enzyme systems that catalyze the same chemistry as discussed above.

6.4.1 Pin1 PPIase

Pin1 is also a PPIase that catalyzes the isomerization of the peptidyl-prolyl bonds. Although Pin1 catalyzes the same reaction as CypA, a difference between the two enzymes is that Pin1 is preferential to the isomerization of phosphorylated substrates (pSer-Pro or pThr-Pro motifs). NMR studies have provided preliminary indications that the intrinsic flexibility for both enzyme folds during catalytic turnover are very similar, Pin1 computational modeling of the isomerization provides vital insights that the catalytic residues, which form critical interactions with the bound substrate, are interconnected to flexible surface, loop regions (see Figure 6.13A), similar to that of CypA [167].

The location and impact of reaction coupled flexibility in CypA and Pin1 is remarkably similar, even though there is no sequence or structural similarity. While Pin1 active-site residues show considerable rigidity (providing the hydrophobic pocket for the target proline residue); however, the loops in proximity to the active-site show significant flexibility coupled to the reaction pathway. In particular the surface loops 47-56 and 117-132 show large displacements and are interconnected through hydrogen bond (see Fig. 6.13A). On the other side, the catalytically important residue R68, is also connected to surface regions of large flexibility through hydrogen bonds.

Note that all the flexible regions in Pin1 consist of residues with long side-chains similar to that of CypA loops that play a critical role in catalysis. It is also interesting to note that both CypA and Pin1 show the presence of networks that extend from the flexible surface loop regions all the way to the active site of these proteins. Mutations of some of these network residues have previously indicated significant decrease in the catalytic activity [36].

6.4.2 R67 DHFR

The structure of this plasmid encoded DHFR consists of a homo-tetramer (each subunit 76 amino acids in length). While the reaction is the same as that catalyzed by the chromosomally encoded DHFR, R67 shows neither sequence nor structural homology with chromosomal DHFRs. This type II DHFR was discovered due to its ability to confer trimethoprim resistance upon host bacteria. Both DHFR and R67 DHFR catalyze the hydride transfer from cofactor NADPH to substrate DHF. Recent computational and experimental investigations have revealed interesting similarities in the reaction coupled flexibility for the two enzyme folds. While R67 DHFR uses an endo transition state and EcDHFR uses an exo transition state, the computational modeling show similarities as to how these two enzymes provide crucial structural interactions in the active-site (see Fig. 6.14A).

There are similarities with regards to the relative motions (see Fig.6.14), both in the center and on the edge of the active-site, that alter the chemical environment making it suitable for catalysis to occur. In particular, puckering of the NADPH ring and a change in the DHF-tail angle coupled with the hydride transfer are observed in both enzyme systems. In DHFR, the nicotinamide ring puckering motion has been suggested to be induced by a network of coupled motion originating from Asp122 on the surface and terminating in Tyr100 in the active-site, positioned behind the CD. In R67 DHFR, the Gln67 side-chain appears to provide similar motions to the CD atom, also positioned behind the cofactor ring. The most interesting difference in the enzyme mechanism appears to impact the chemical environment at the CA (see Figure 6.13A). In DHFR, it has been suggested that the motions of the Phe31 side chain provide promoting motions that alter the DHF-tail angle, thereby making the CA more suitable for the incoming hydride. In R67 DHFR, the present computational studies predict that the same change in the chemical environment results from the p-ABG tail movement. Sampling of the DHF-tail angle is made possible by the large flexibility of the tail located at the edge of the pore surrounded by bulk-solvent. The two extreme states for the conformations are ion-pairs between the α-γ carboxylate groups of DHF interacting with symmetry related lysines from two different subunits (K232 and K332).

6.4.3 Human Angiogenin

The human angiogenin protein provides an interesting comparison for the chemistry catalyzed by RNaseA. This protein also catalyzes hydrolysis of single stranded RNA; however, with low catalytic efficiency (at rates $10^5 - 10^6$ less) than RNaseA [240, 241]. Angiogenin is structurally similar to RNaseA, with the active-site showing similar contacts with

the substrate, however, the major difference with RNaseA is the presence of a truncated surface loop (see Fig. 6.13C). This surface loop forms an important part of the network in RNAseA; however, angiogenin is missing portion of the network due to a truncated flexible loop near RNaseAs Asp121. Interestingly, the dynamical motions of Asp 121, a network residue that interacts with Lys66 on the truncated loop, have been implicated in catalysis and mutation of this residue results in 90% activity loss in RNaseA [65, 35].

6.5 Conclusions

An integrated view of protein structure, flexibility and function is emerging for better understanding of the detailed biophysical mechanism of enzyme catalysis.6 The role of conserved structural interactions between active-site residues and substrate has been understood for some time; however, the role of overall enzyme fold still remains a mystery, particularly the conserved residues that are located far away from the active-site. On the other hand, increasing evidence continues to link protein motions with designated function. The intrinsic flexibility of a protein is related to the overall shape that is the protein fold, as well as local organization of dynamical regions. Does all the emerging evidence suggest that the overall enzyme fold is optimized for structural as well as dynamical effects to carry out the protein function?

Careful characterization of the networks discovered in the three enzymes described in this study, leading to certain common features (see Fig. 6.13). The networks discovered connect surface loop regions to conserved active-site residues that make direct contacts with the substrates. The surface loop regions show large flexibility as observed in X-ray and NMR investigations as well as computational studies. These regions are exposed to the solvent and contain non-conserved residues that share a common feature of large side-chains, suggesting coupling with the solvent fluctuations. Another common feature observed in these networks is the connection of these flexible loops through a conserved hydrogen-bond with another region of the protein in the direction of the active- site. It has been reported that bulk solvent fluctuations drive internal protein dynamics, therefore, impacting protein function [84, 5]. Previous investigations have also suggested the existence of energy pathways as a part of protein structure [184, 250]. Further, evidence supporting the interconnection of the network surface loops to catalysis comes from human angiogenin protein [65, 163].

The interconnection between enzyme fold, flexibility and function presented here suggests the conventional emphasis structure-encodes-function may need revision to better understand the fundamental mechanism of how enzymes work. Conservation of reaction

coupled conformational flexibility as an important characteristic of the enzyme fold suggest that structure encodes dynamics and structure-dynamics encode function. It is entirely possible that specific enzymes have evolved to utilize the structural interactions with flexibility making only minor contributions.3 While in other systems such as CypA, DHFR and RNaseA the contributions of flexibility could be closely related to the enzyme mechanism. This emerging view of proteins provides some basis for understanding allosteric and cooperative effects and also has wide implications for drug design as well as protein engineering.

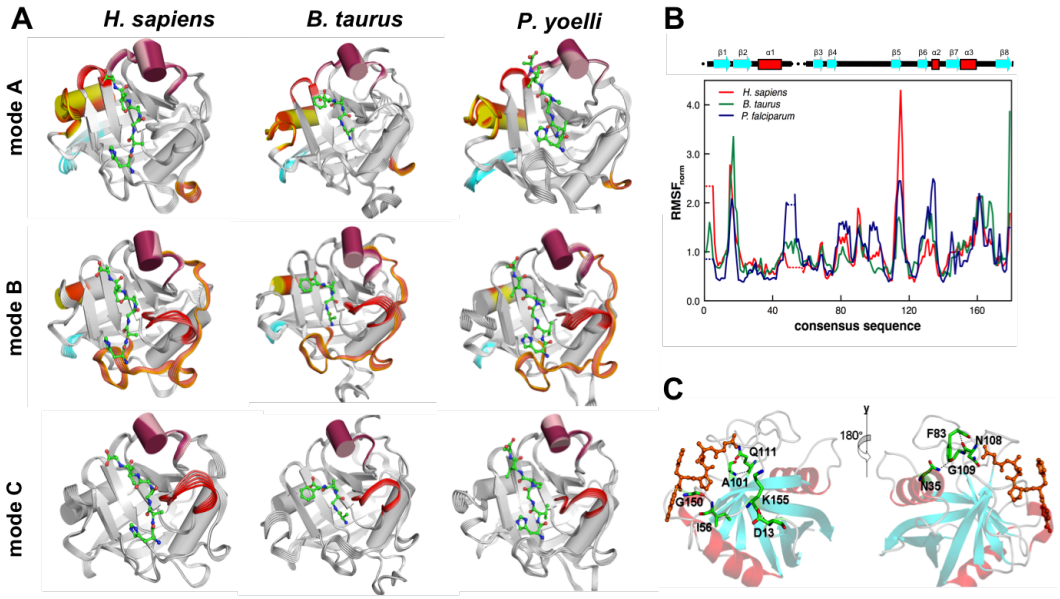

Figure 6.1: **Conservation of reaction coupled flexibility in enzyme CypA across 3 different species.** (a) Top 3 slowest modes coupled to the cis/trans isomerization reaction show large fluctuations in identical regions (near and away from the active-site). Multiple snapshots are shown to indicate movements along the modes, and the regions with high flexibility are shown in color. (b) Enzyme back-bone flexibility depicted as root mean square fluctuations (RMSF); computed by aggregating the C^{α} displacement magnitude in the top 10 modes coupled to the reaction. For comparison consensus sequence has been used and RMSF has been normalized by dividing by the average C^{α} flexibility of all residues in the enzyme. (c) Conservation of the network interactions connecting the flexible regions as a part of CypA fold (only human CypA is shown; however these interactions are conserved in human cyclophilin B, CypA from B. Malayi, B. Taurus and E. coli as well). See supporting information for the animated movies.

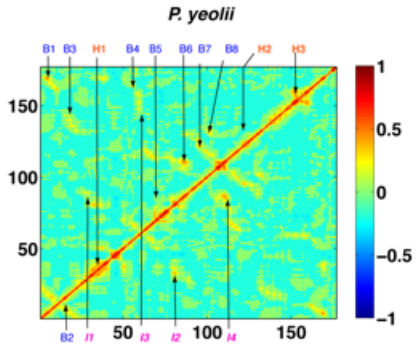

Figure 6.2: **Cross-correlations observed along the reaction profile for cyclophilin A**. B1-B8 correspond to the correlations along the β-sheet of the enzyme. H1-H3 correspond to the 3 α-helices. Regions marked I1-I4 correspond to distal correlations observed along loop structures. I1: residues 29-33 with 85-86, I2: 34-36 with 77-78, I3: 56-57 with 142-150 and I4: residues 82-85 with 104-108. Note, residue numbers refer to H. sapiens as the reference species; corresponding residue numbers for the two species are available in table: 6.2.

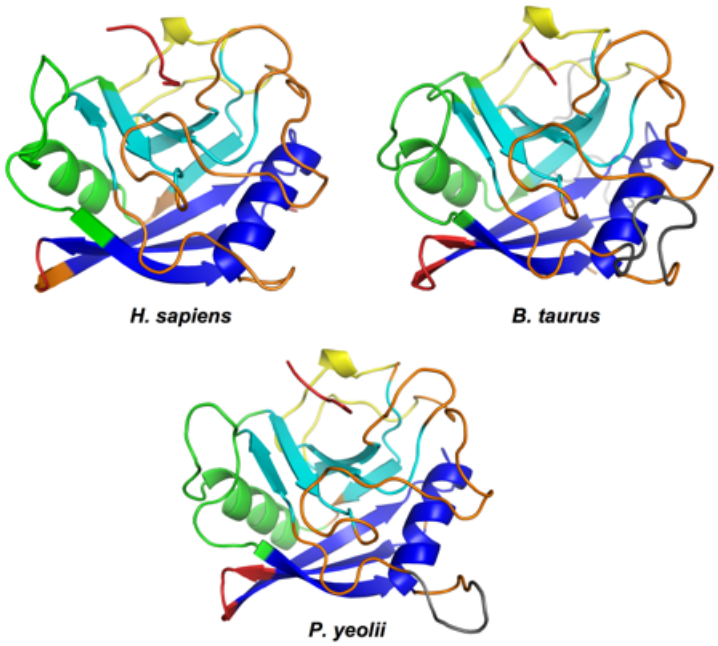

Figure 6.3: **Dynamical clusters in enzyme cyclophilin A**. Six clusters were identified (marked in different colors), that were identically across the three species. The substrate (shown in red stick representation) and the β-hairpin formed by residues 13-16 (H. sapiens; red cartoon) exhibit large-scale fluctuations. The hydrophobic core of the protein (dark blue) and the active site regions (cyan) show similar motions across all the three species. Flexible surface loops along the outer edge of the active site (orange) are coupled across all the three species. The flexible loops behind (yellow) and adjacent (green) to the active site region exhibit coupled motions that are conserved features of this enzyme fold. Note, regions that are insertions in the other two species (B. taurus and P. yeolii) are shown in dark gray color. Regions of similar dynamical fluctuations are conserved, indicating that dynamics coupled to the catalytic mechanism are conserved across multiple species regardless of sequence homology.

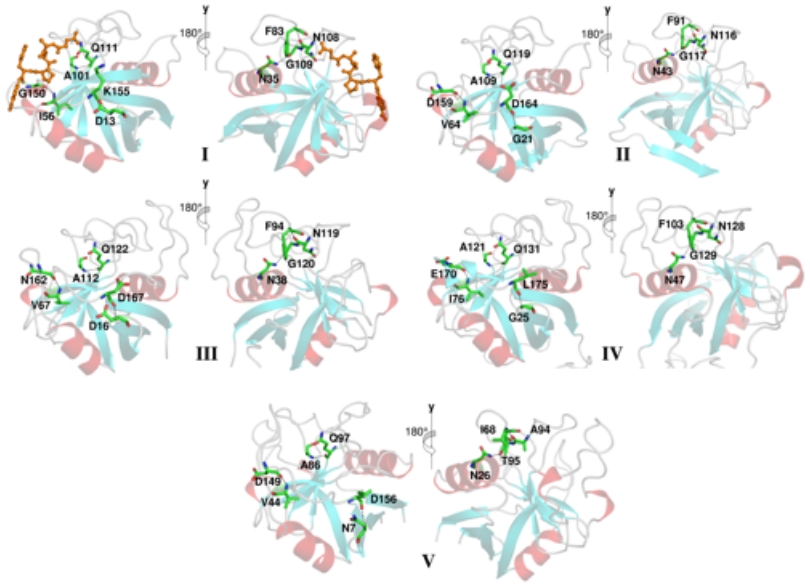

Figure 6.4: **Conservation of the network interactions as a part of PPIase fold**. human CypA (I) PDB code: 1RMH; human cyclophilin B (II) PDB code: 1CYN; B. Malayi (III) PDB code: 1A33; B. Taurus (IV) PDB code: 1IHG; E. coli (V) PDB code: 2NUL. The equivalent hydrogen bonds are listed in the table 6.3.

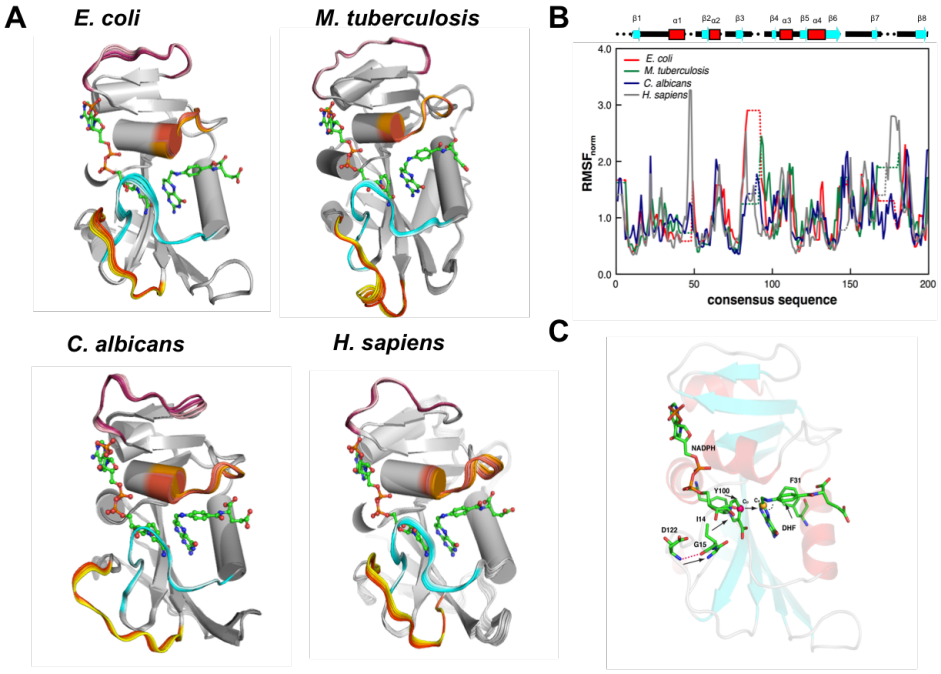

Figure 6.5: **Conservation of reaction coupled flexibility in enzyme chromosomal DHFR across 4 species**. (a) Slowest mode coupled to hydride transfer show large fluctuations in same regions (near and away from the active-site) of the enzyme from 4 species. (b) Enzyme back-bone flexibility depicted as normalized root mean square fluctuations (RMSF). (c) Conservation of the network interactions connecting the flexible regions as a part of DHFR fold (only E. coli DHFR is shown). The modes are depicted/colored and the RMSF is normalized in a similar way to the CypA results.

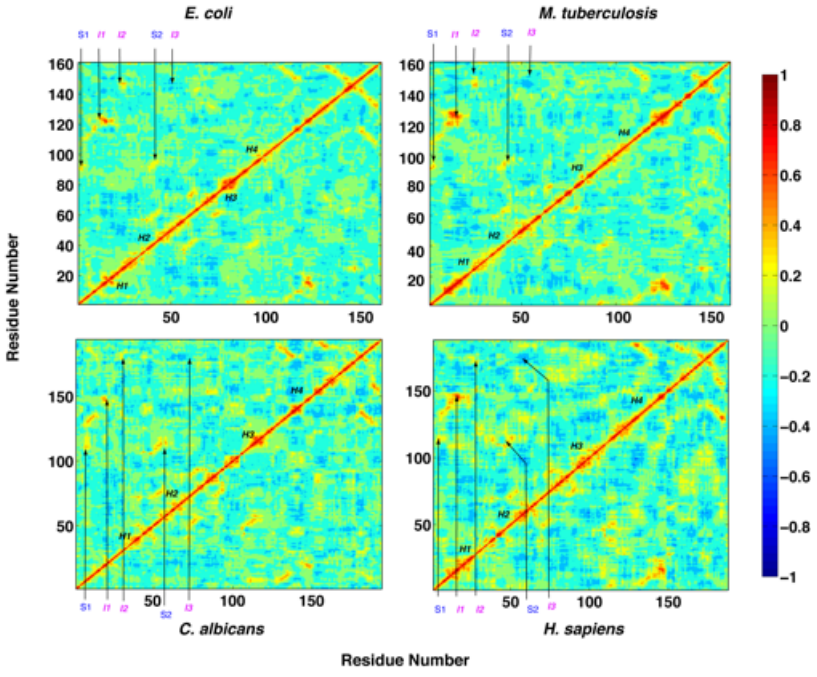

Figure 6.6: **Cross-correlations in enzyme DHFR along the hydride-transfer**. Regions marked S1-S2, H1-H4 represent the correlated dynamics of the secondary structural elements in DHFR. Regions I1-I3 however correspond to distal correlations observed from the reaction profile. I1: residues 15-22 correlated with 116-125 (Met20 and βF-βG loops), I2: 31-36 (αA) correlated with 142-150 (βG-βH). I3: residues 64-72 (βG-βH) negatively correlated with residues 142-150 (βG-βH). Note, we have used the reference structure as E. coli (1RX2). Corresponding regions from other species are shown in the table 6.4.

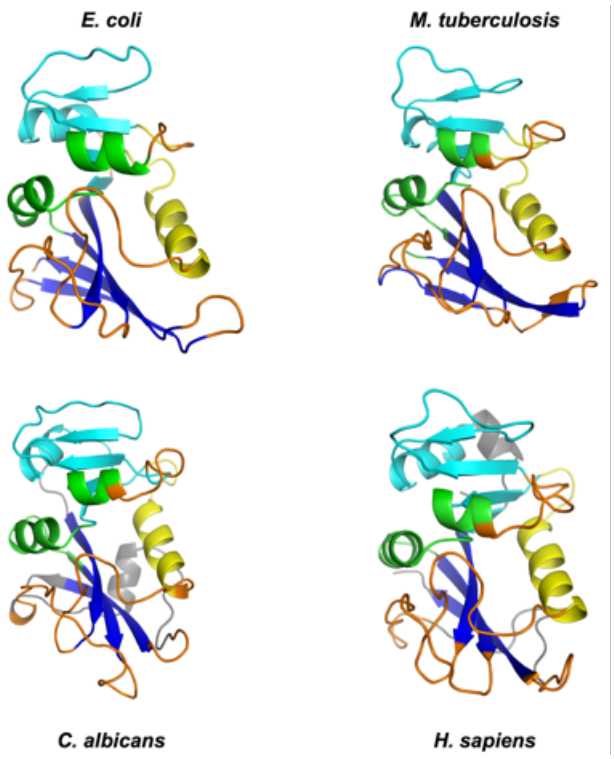

Figure 6.7: **Dynamical clusters of residues in enzyme DHFR**. Five dynamical clusters were identified across the four species studied, indicating a identical behavior of flexibility coupled to hydride transfer. The Met-20, βF-βG, βG-βH and the substrate binding loops (shown in orange) exhibit large-scale fluctuations. The flexibility of these regions is a conserved feature of this enzyme fold. The central β-sheet is split into two clusters (cyan and dark blue) which is consistent with the observation by Sawaya and Kraut [234] regarding the intrinsic twist in the β-sheet. Further, loops shown in yellow are coupled to the substrate-binding region. Regions shown in dark gray (in M. tuberculosis and H. sapiens) are additional inserts not found in the other species.

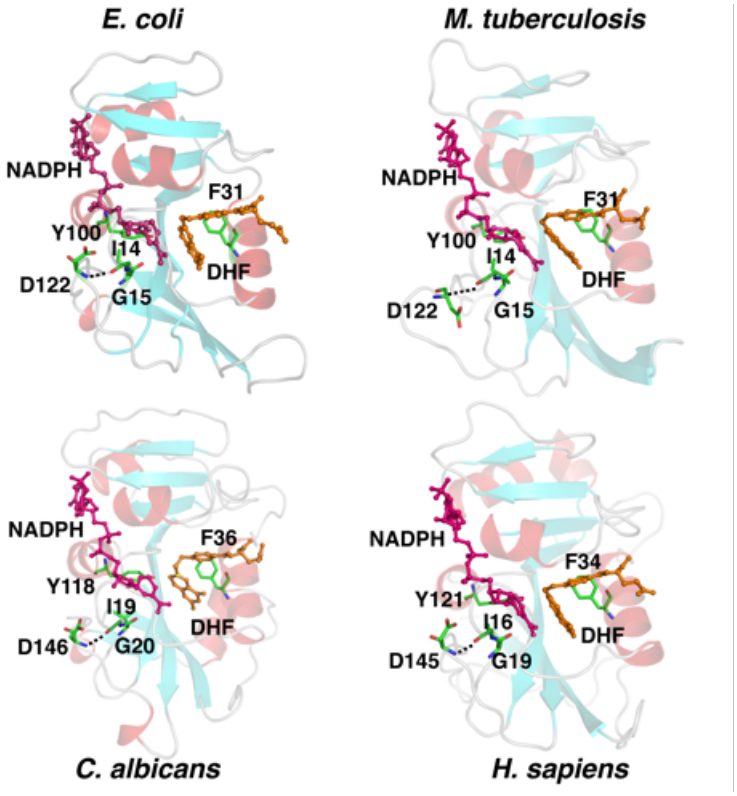

Figure 6.8: **Details of the conserved network of coupled motions in enzyme chromosomal DHFR**. The flexible loops on the surface are connected to the active-site through conserved residues, hydrogen bonds and hydrophobic interactions as listed in the table 6.5.

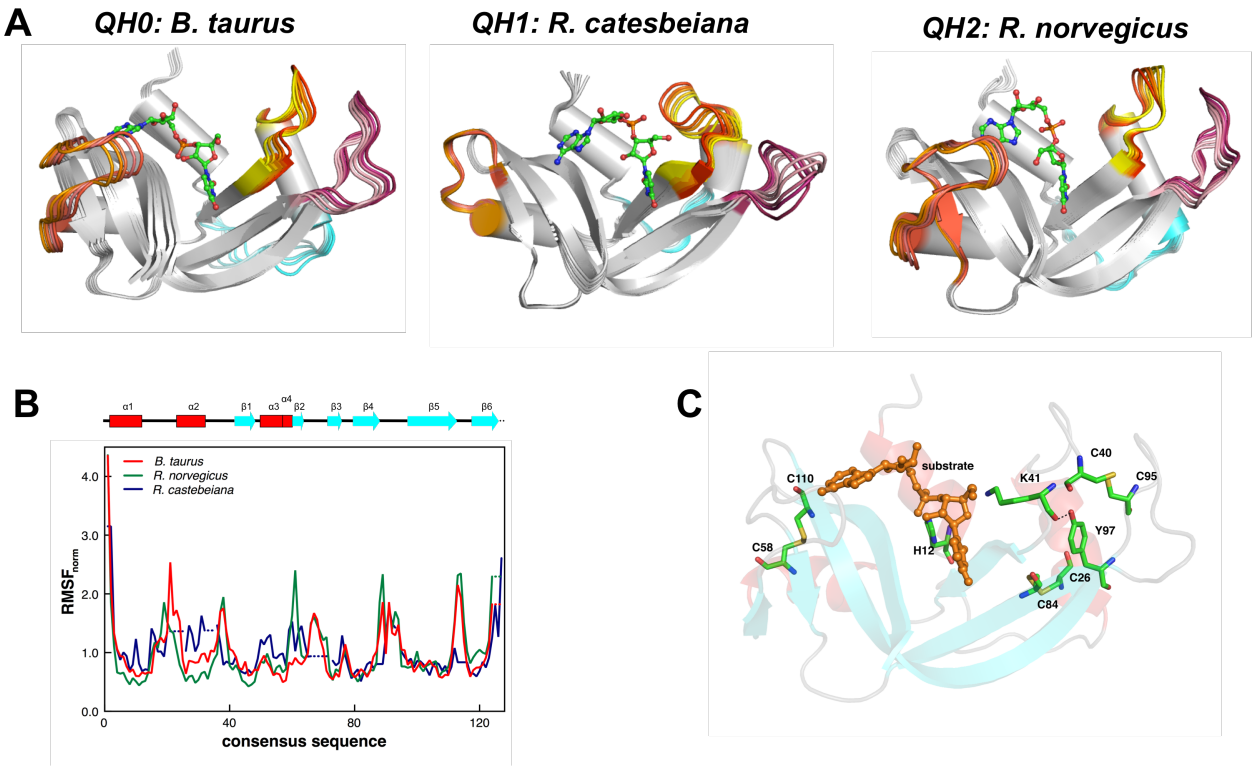

Figure 6.9: **Conservation of reaction coupled flexibility in enzyme RNaseA across 3 species**. (A) Slowest mode coupled to RNA hydrolysis show large fluctuations in same regions (near and away from the active-site) of the enzyme from 3 species. (B) Enzyme back-bone flexibility depicted as normalized root mean square fluctuations (RMSF). (C) Conservation of the network interactions connecting the flexible regions as a part of RNase fold (only B. taurus RNaseA is shown). Other network residues are shown in detail (Fig. 6.12).

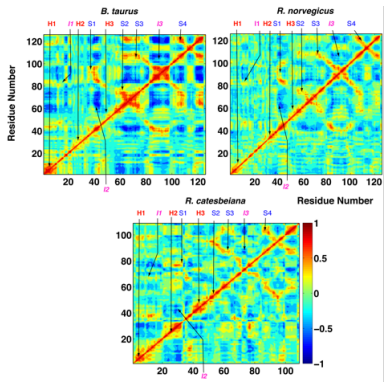

Figure 6.10: **Cross correlations in RNase A**. Regions H1-H3 and S1-S4 correspond to correlations observed from secondary structural elements (α1-α3, β1-β5) respectively. Regions I1-I2 corresponds to distal correlations observed. The distal correlations observed from RNase A are depicted in the table 6.6.

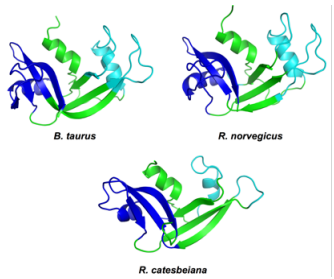

Figure 6.11: **3 identical clusters were identified across the species.**The 3 α-helices are clustered into three regions (blue, green and cyan), indicating that the dynamics of these helices are quite different. The β-sheet is split into 2 distinct clusters (green and blue) depending on how these regions flank the substrate in the active site. The opposed movements of the β-sheet regions (see movies), and the motions of the flexible loop regions (cyan and blue regions) are a conserved dynamical feature of the RNaseA fold.

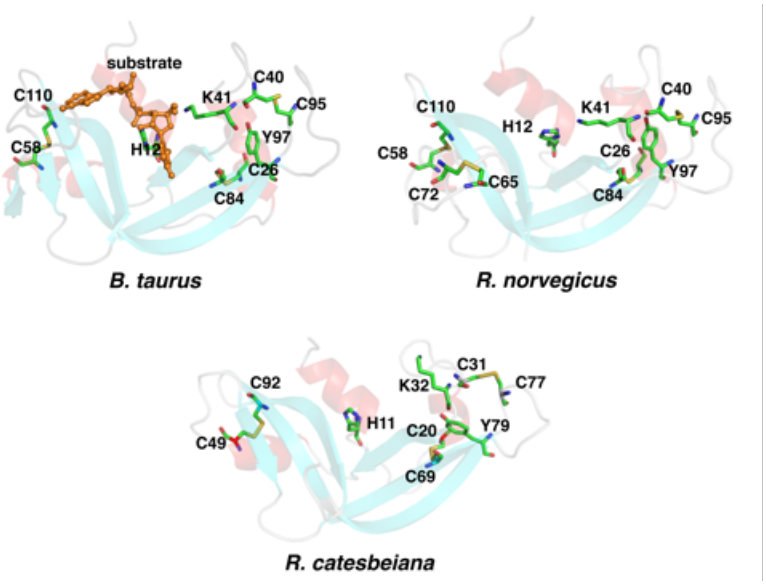

Figure 6.12: **Details of the conserved network of coupled motions in enzyme RNase A**. The flexible loops on the surface are connected to the active-site through conserved residues and disulphide bonds as labeled in the figure.

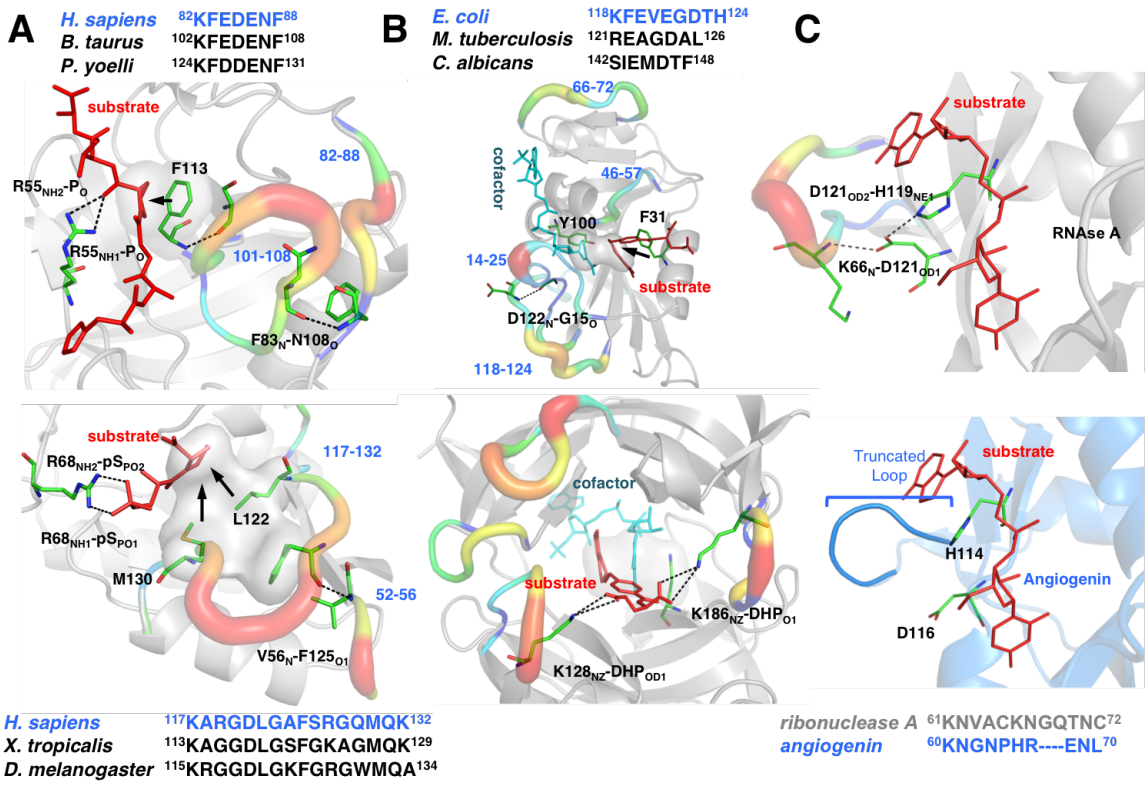

Figure 6.13: **Conservation of Reaction Coupled Dynamics in enzymes catalyzing the same biochemical reaction.** Three reaction mechanisms were considered (A) isomerization (B) hydride transfer and (C) hydrolysis. Observe that in (A) top panel: CypA shows the presence of a highly conserved residue F113 which shows motions coupled to the catalytic step by providing crucial hydrophobic interactions with the bound peptide. In (A) bottom panel, the reaction mechanism for Pin1 PPIase is illustrated; observe the presence of L122 and M130- both residues provide hydrophobic interactions to the substrate. In both enzymes, a conserved hydrogen bond extending from the surface region to the active site is present, illustrating the importance of the network. In (B), we highlight the reaction coupled mechanism for DHFR from chromosomal DHFR (top panel) and a primitive enzyme mitochondrial DHFR (bottom panel). Observe the presence of Y100 behind the substrate in chromosomal DHFR - the corresponding residue in mitochondrial DHFR is Q67 which provides the required hydrophobic interaction for the accurate placement of the substrate in both cases (see Fig. 6.14). In panel (C), we illustrate ribonuclease A (top panel) with angiogenin (bottom), showing the truncated loop. The consequence of losing the loop is that crucial interactions extending from the surface to the active site (shown as hydrogen bonds in the top panel) are missing in angiogenin, which may contribute to its lower catalytic efficiency.

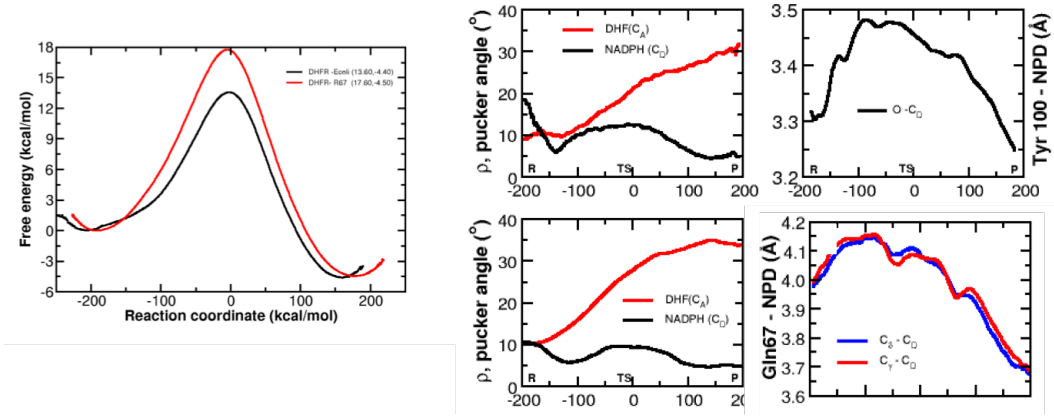

Figure 6.14: **Comparison of R67 DHFR with E.coli DHFR.** (A) Reaction profiles generated from E.coli DHFR (black) versus R67 DHFR (red) show how the primitive enzyme differs in its ability to catalyze the reaction. (B) Equilibrium averages of geometric properties along the reaction coordinate for E. coli DHFR (top panels) and R67 DHFR (bottom panels).

Chapter 7

Enzyme Super-family shows Conservation of Protein Flexibility linked to Catalysis

Enzymes catalyzing diverse chemical reactions have shown the presence of networks of coupled motions promoting catalysis. It is hypothesized that the identification of conserved regions of enzyme flexibility could provide vital information about the link between protein motions and function. As a test, the slow conformational fluctuations associated with the common sub-step of hydride transfer catalyzed by four diverse members of the dinucleotide binding Rossmann fold proteins (DBRPs) enzyme super-family have been characterized. These enzymes, while sharing only about 10% sequence identity and low structural similarity, show remarkably identical patterns in reaction coupled flexibility in three regions. These flexible regions, located in distant areas of enzyme structures, are interconnected by conserved network of interactions. Enzyme motions in these networks impact hydride transfer reaction by decreasing the distance between the donor and acceptor atoms. The preservation of flexibility in distant regions, along with structural interactions in the active-site, is proposed to be important characteristic feature of the DBRP super-family for catalyzing hydride transfer. These findings support the emerging paradigm of understanding enzyme catalysis that states structure encodes dynamics and together structure-dynamics encode function.

7.1 Introduction

Conserved features of proteins provide vital clues about various aspects of their life cycle including the designated function. In enzymes, active-site residues that form crucial interactions with the substrate or structural elements facilitating in positioning of the substrate (or cofactor) are conserved from prokaryotes to eukaryotes. Therefore, identification of conserved features has become an important technique in search for the fundamental understanding of proteins. Recently, an integrated view of enzyme structure, dynamics and catalysis has been proposed [164]. Increasing evidence indicates that in addition to structural interactions between the enzyme and substrate, internal protein motions also play a vital role in the biophysical mechanism of catalysis. In this chapter, extending the concept of conserved elements to flexible enzyme regions, we investigate the interconnection between internal protein motions and catalysis.

Proteins are flexible molecules. Even at ambient conditions proteins are constantly undergoing conformational fluctuations, which enable them to sample the kinetically accessible parts of the conformational landscape [47]. Motions occurring at time- scales of enzyme catalysis have been identified in a number of enzymes, raising the question of whether they are interrelated to mechanism of catalysis or not [6, 40, 156]. The idea of possible connection between protein motions and substrate turnover during the enzyme function is not new and dates back several decades [112, 113]. As experimental techniques are providing more detailed information and computational models are becoming more realistic, new insights into the linkage between protein flexibility and function are becoming available. Experimental techniques including nuclear magnetic resonance (NMR) spin relaxation, single molecule experiments have probed the motion of protein residues in connection with substrate turnover step catalyzed by enzymes [47, 78, 40]. In particular, movements of back-bone as well as single residues in active-site and distal regions at time-scale of the reaction have been identified in enzymes including dihydrofolate reductase (DHFR) catalyzing hydride transfer and cyclophilin A (CypA) catalyzing peptidyl-prolyl cis/trans isomerization [47, 78]. As we described in the previous chapter, computational techniques continue to provide fascinating details ranging from fast motions of single amino-acids in the active-site to slow conformational fluctuations spanning entire protein. Theoretical and computational investigations have lead to the discovery of networks of protein motions/vibrations promoting catalysis in DHFR and CypA. These networks are formed by protein residues and hydrogen bonds extending from surface regions all the way to the active site. The presence of these networks has been confirmed by NMR spin relaxation investigations performed [270].

Intrinsic protein flexibility of small enzymes has been discovered to be closely linked to

enzyme function, therefore, is an important aspect of the overall shape or the enzyme fold. In the previous chapter (Chapter 6), we have described that the reaction coupled conformational fluctuations are conserved over evolution in a number of enzyme folds catalyzing different types of chemistry: DHFR, CypA and ribonuclease A (RNAseA). Even though the protein sequence is different, active-site and distal regions of the enzyme exhibit same movements across species ranging from bacteria to human. Regions of these enzymes showing high flexibility in the surface residues are connected to active-site through the networks formed by conserved residues and interactions. The reaction coupled motions originate in the flexible surface loop areas of the protein and extend all the way into the active-site connected via the discovered networks. Slow conformational fluctuations in the discovered networks are coupled to the reaction pathway. Mossbauer effect and neutron scattering experiments have indicated that fluctuations in the hydration shell and the bulk solvent enslave protein motions [85]. These solvent driven motions of the protein residues on the surface impact the active-site enzyme-substrate interaction by a series of coupled motions in the discovered networks. The biophysical role of these motions lies in altering the environment in the active-site, near the transition state, impacting reaction trajectories to successfully cross over the activation energy barrier [5]. Furthermore, multiple pathways provide alternate ways for the energy to be propagated to the site of catalysis where the conserved residues through subtle motions manipulate the chemical environment to favor the reaction to proceed from the reactant to the product state.

Enzymes over evolution have developed features that make them especially well suited to catalyze the target chemistry. The ability of enzymes to catalyze the large number of biochemical reactions in diverse substrates begs the questions whether each protein fold has evolved independently. Structural and sequence analysis of a diverse enzyme superfamily with extremely different sequences and structures have provided an example of the existence of common characteristics between functionally diverse members of enzyme super-families share common structural and biochemical features [101]. Babbitt and coworkers have reported that mechanistically diverse enzymes that share a common sub-step have very similar protein folds, based on the exploration of a family of proteins as a sub-set of two dinucleotide binding domains flavin-proteins [209]. It was observed that the conserved structural features are optimal for stereo-specific hydride transfer that is stabilized by specific interactions with amino acids from several motifs distributed among both dinucleotide binding domains. Bahar and co-workers investigated four structurally similar yet functionally diverse proteins and found that the binding dynamics of the proteins were largely similar characterized by a common fold and stretches of residues with distinctive dynamical features [153]. Similarly, protease enzyme super-family as well as the RAS GTPase family shows convergent dynamics associated with substrate binding [55, 224]; where the movements of the active-site residues are similar even though the

structures shown considerable variations. However, the connection between the chemical step during enzyme catalysis and protein flexibility in enzyme super-family has not yet been explored.

Conservation of flexible regions in enzyme super-families could provide vital clues to their role in the catalytic function, similar to how conserved enzyme-substrate interactions provide insights into the mechanism. In this chapter, we present our investigations to characterize reaction coupled flexibility in diverse members of an enzyme super-family. The selected super-family consists of oxido-reducatses that bind and utilize dinucleotide cofactors: nicotinamide adenine dinucleotide (NAD+), its phosphorylated and reduced forms ($NADP^+$, NADH and NADPH). These cofactors play a central role in cellular metabolism and energy production, as hydride-accepting and hydride donating coenzymes, in many essential biochemical processes including glycolysis and the citric acid cycle. We refer to this enzyme super-family as the dinucleotide binding Rossmann fold proteins (DBRPs) due to the presence of the Rossmann fold, $\beta - \alpha - \beta - \alpha - \beta$ motif that binds a nucleotide [230]. Here, protein flexibility of four diverse members coupled to the common sub-step transferring hydride transfer step between substrate and cofactor has been characterized (see Fig. 7.1). The investigated enzymes have very low sequence identity with diverse structural topologies. The reaction coupled flexibility shows existence of promoting motions that are remarkably similar. Detailed analyses indicate that in addition to structural constraints, motions that play a promoting role in catalysis are also a conserved feature of the entire super-family.

7.2 Methods

Complete reaction profiles for the hydride transfer reactions catalyzed by the four enzymes were generated using molecular mechanics potentials in combination with empirical valence bond (EVB) [272] method. Based on previous evidence, protonation of the substrate associated with the reaction was assumed to occur prior to the hydride transfer [8, 13, 273]. Conformations of the enzyme-cofactor-substrate complex sampled along the entire reaction pathway were used to identify the slow conformational fluctuations associated with the hydride transfer reaction.

7.2.1 Modeling the hydride transfer reaction pathway

The four enzyme systems were simulated in explicit water with parm98 force-field and AMBERs associated simulations package. The starting protein conformations were taken

No.	Enzyme	Species	PDB Code
1	dihydrofolate reductase (DHFR)	*Escherichia coli*	1RX2 [234]
2	human biliverdin IX beta reductase (HBBR)	*Homo sapiens*	1HE4 [219]
3	6,7-dihydrobiopterin reductase (DHPR)	*Rattus norvegicus*	1DHR [266]
4	pteridine reductase (PR)	*Leishmania major*	1E92 [106]

Table 7.1: **Rossmann Fold (DBRP) members studied in this chapter**

from the protein data bank. The four proteins simulated are summarized below in table 7.1. For PR, the missing coordinates for following residues 1-5, 74-80 and 123-130 were generated based on homology modeling and simulation in vacuum. For the missing protein regions in enzyme PR, comparison with other known Rossmann fold proteins indicated that the region correspond to extended loops. Therefore, we modeled the C^{α} extended from the protein and AMBERs leap module was used to add the remaining atoms. Slow minimization in vacuum was used to relax to added residues (these coordinates are available on request). For missing substrate in DHPR, the structure was aligned with PR (using DaliLite server: http://www.ebi.ac.uk/Tools/dalilite/) [130, 131, 132] and the coordinates of the substrate in PR was used for modeling the substrate in DHPR. After the model preparation enzyme-substrate in explicit water, the system was equilibrated based on protocol described previously (see Chapter 2). Briefly the model was minimized to remove bad contacts and slowly heated to 300K. All production runs were performed at 300K under NVE conditions. EVB method was used for the reaction pathway sampling, as outlined in Chapter 6.

EVB method in combination with classical molecular mechanics was used for generation of the conformations along the hydride transfer reaction. The EVB method developed by Warshel and coworkers,14 has been used by a wide community of researchers to investigate enzyme reactions[8, 13, 273]. In EVB approach, the Hamiltonian for the system is represented as a matrix.

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \tag{7.1}$$

where H_{11} represents the reactant state and H_{22} represents the product state that can be computed using molecular mechanics method such as AMBER or others. The off- diagonal terms (which can be assumed to be constant for simple implementation) and a correction term (δ_{22}) to the H_{22} can be computed from the experimentally measured forward and reverse reaction rates. Using a mapping potential framework, where the Hamiltonian

is described by the equation:

$$V_\lambda = H_{11} \times (1 - \lambda) + H_{22} \times \lambda \tag{7.2}$$

and varying the mapping parameter λ from 0 to 1, the reaction was gradually mapped along the reaction coordinate, with 7 different values of mapping parameter: λ=0 (reactant state), 0.1, 0.3, 0.5, 0.7, 0.9, 1.0 (product state) were used. The values of these parameters are α=1.785 Å^{-1}, D_e= 103 kcal/mol, and R_e =1.09 for the bond between the donor carbon and the hydride, and α=1.758 Å^{-1}, D_e =103 kcal/mol, and R_e=1.09 Å^{-1} for the bond between the acceptor carbon and the hydride. For each of the 7 bins, 1000 conformations were collected representing the enzyme-substrate conformations sampled along the reaction pathway.

7.2.2 Protein flexibility coupled to the hydride transfer step

The reaction coupled protein flexibility was analyzed using quasi-harmonic analysis (QHA) [150]. The atomic fluctuation matrix for QHA (Eq. 1.6) was constructed from system snapshots traversing the entire reaction path, collected during EVB runs. Reaction path averaged QHA provides information about the protein conformational fluctuations occurring at the time-scale of the enzyme reaction (typically microsecond and longer). The low frequency modes correspond to longer time periods and adequate sampling ensures reproducibility. Coupling of these QHA modes to the hydride transfer step is defined by dot product for the CD-CA vector and the hydride displacement vector in the mode. Note that we have used a similar definition for the computation of the modes coupled to CypA activity indicate large fluctuation in the peptide bond being isomerized.

7.3 Results

As described in Fig. 7.1, these enzymes share the common step of catalyzing hydride transfer from cofactor NAD(P)H to substrate. However, as depicted in Table 1, these enzymes are considerably diverse in sequence as well as in structure, sharing only 4-13% of sequence identity and low structural similarity (RMSD of 3-5 Å). Even though these proteins contain 2 occurrences of the Rossmann fold fused together in the same protein chain, the topologies of the proteins are considerably different. Each Rossmann fold consists of a core β-sheet formed by 4 stands surrounded by α-helices; however, the location β-strands and α-helices in the primary sequence are different between the enzymes. Intrinsic flexibility of the 4 enzymes over the course of hydride transfer indicates

common characteristics (see Fig. 7.2). These enzymes show presence of 4 distinct regions with similar flexibility, as identified by correlated motions between residues and dynamical clustering analysis of enzyme-substrate conformations along the entire reaction pathways. The first distinct region consists of the rigid core formed by the central β-sheet, except for DHFR where it separates into 2 rigid regions.

The central β-sheet in HBBR, DHPR and PR (shown in dark blue) are dynamically coupled and exhibit low conformational flexibility. The β-sheet forms the rigid core of the protein, which is a conserved feature of this enzyme super-family. The flanking helices along the β-sheet cluster into two and are shown in green and yellow respectively. The helices shown in green are dynamically coupled to the co-factor. The parts of the helices shown in yellow couple to the substrate motions and largely involve interactions that stabilize the substrate within the active site environment. The flexible loop regions (orange), layering the surface of each enzyme exhibit large-scale fluctuations. Each of these flexible loops can be mapped on to a corresponding area within DHFR. In DHFR the flexible loops around the active site consist of Met-20, βF-βG, βG-βH loops (cluster A). The equivalent loops in HBBR (150-163, 169-178), DHPR (182-198), and PR (224-254) are located on the opposite side of the protein as marked on the structures. Depending on the location of the active site, these loops show coupled motions in the protein. The substrate-binding loop (cluster B) in DHFR comprises of two flanking helices (shown in green and yellow). Corresponding regions in HBBR (75-88, 115-130), DHPR (82-99, 133-147) and PR (109-147, 181-196) are also marked on the figure. Only in DHFR, the large-scale fluctuations of the adenosine binding loops (cluster C; 64-72) are separated from the rest of the protein (cyan) however, in HBBR (34-49, 52-59), DHPR (37-41, 50-63) and PR (36-42, 63-85), these regions are clustered along with the loops shown in orange. The presence of extra cluster in DHFR (shown in cyan) is attributed to the fact that the two Rossmann folds in DHFR are fused at an angle of 41° [234], and hence the motions between the β-strands may be influenced by the presence of this intrinsic twist in the structure. The presence of similar dynamical coupling indicates that the overall dynamics as measured by the distance fluctuations along the hydride-transfer step is a conserved feature of the enzyme super-family.

Additional clusters are formed by less flexible helices and surrounding loop regions. The most flexible cluster region is formed by surface loops, including the loops in the proximity of the cofactor nicotinamide ring binding pocket in the active-site, showing correlated motions with large displacements. In DHFR this region consists of the Met 20 and βF-βG loop regions, previously implicated in the enzyme activity through dynamical motions Additionally, the surface loops regions near the substrate binding pockets of all the 4 enzymes shows considerable flexibility and correlated motions over the course of

	HBBR (205)	DHPR (236)	PR (288)
DHFR (159[a])	10% / (90[b])1.7 / 3.8 Å	4% / (78) 0.6 / 5.6 Å	5% / (99) 1.8 / 4.6 Å
HBBR		10% / (173) 15.8 / 2.6 Å	12% / (181) 18.0 / 2.7 Å
DHPR			13% / (215) 20.3 / 2.9 Å

Table 7.2: **Sequence and structural comparison the four enzymes investigated in this study.** The alignments were performed using DaliLite pair wise comparison web tool: http://www.ebi.ac.uk/DaliLite/. a = Total number of residues in the enzyme b = Number of resides used in the alignment by DaliLite

reaction. The edge of the β-sheet with surface loop regions (including adjacent helix in HBBR and DHPR), where the adenosine part of the cofactor binds, also indicates large displacement with correlated motions.

Slow conformational fluctuations, coupled to the catalyzed hydride transfer, display motions in structurally equivalent regions of the 4 enzymes. Protein vibrational mode showing the largest coupling to the hydride transfer step for the 4 enzyme systems is depicted in Fig. 7.3. Note that this mode provides information about the enzyme motions at the time-scale of the hydride transfer. As we previously reported, reaction coupled modes for cis/trans isomerization catalyzed by enzyme CypA were obtained using similar methodology [9, 4]. Fig. 7.2 depicts the highest flexibility in the reaction coupled modes is present in structurally equivalent regions of the enzyme fold: residues 9-24 (Met20 loop), 48-54, 64-72 (adenosine binding domain) and 114-126 (βF-βG loop) in DHFR; regions 34-49, 52-59, 75-88, 115-130, 150-163 and 169-178 in HBBR; residues 37-41, 50-63, 82-99, 133-147 and 182-198 in DHPR; and regions 36-42, 63-85, 109-147, 181-196 and 224-254 in PR. These regions of the protein include the surface loop regions where side-chains show large displacements; other flexible regions are located at the edge of the active-site or in its vicinity. The residues in these regions show correlated motions over the course of hydride transfer. Two additional modes that correspond to second and third most reaction coupled modes also show motions in the above mentioned regions. As the animated movies of these modes in the Supporting Information indicates, all regions listed above show displacements in the 3 modes; however, the extent of displacements depends on the mode.

Overall, the detailed characterization of the reaction coupled flexibility show inverse relationship between the enzyme activity and flexibility. The enzyme active-sites are considerably rigid while the loop regions, particularly the solvent exposed external loops, show large displacements in the slow conformational fluctuations coupled to the enzyme

catalysis step. These observations are consistent with other studies [113]. An indication of the reaction coupled flexibility is provided by the aggregated atomic displacements in the 10 modes that show the largest coupling with the hydride transfer, as depicted in Fig. 7.4. This figure depicts that the protein regions with implications in the catalytic mechanism show remarkably identical flexibility, even through the structural organization and the sequence homology is very low, and structures differ in topology as well. The β-strands provide structural motifs, along with conserved residues, for the binding of cofactor and substrate in the rigid environment of the active-site. The dark blue color in Fig. 7.2 indicated this central core is rigid over the course of the reaction, and shows the presence of conserved enzyme-substrate interactions [7]. On the other hand, the three highly flexible surface and associated loop regions are preserved across the diverse members of the enzyme super-family, and have important implications for the catalyzed hydride transfer. These regions are consistent with the findings from correlated motions and dynamic clustering and are marked as Clusters A, B and C in Fig. 7.4.

The biophysical role of these preserved flexible clusters can be understood by detailed structural analysis as well as characterization of motions over the course of hydride transfer (Fig. 7.5). Present within Cluster A and B are networks of motions formed by a chain of interactions that connect the flexible surface loops on the enzyme surface all the way to the residues that form the active-site. Characterization of cross- correlation of motions (see Fig. 7.6) over the course of the reaction and the reaction coupled modes indicate that these networks facilitate in positioning the substrate and cofactor suitable for the hydride transfer. The flexible surface loops of Cluster A impact the cofactor binding pocket, in particular the nicotinamide ring with the donor carbon. The flexible loops as well as the helices in Cluster B play a role in bringing the acceptor atom of the substrate in proximity of the cofactor. Note that in DHFR, the substrate binding pocket is present on the opposite side of the cofactor; therefore, this region is located on the opposite face of the cofactor as compared to the other 3 proteins. The flexible loop regions of Cluster C possibly play a role in the cofactor binding [40]; however, further investigations are required for understanding the role of Cluster C.

In DHFR, a network of coupled motions promoting the hydride step has been previously proposed [7] and verified [40]. Starting from surface hydrogen bond Asp122-Gly15 and mediated through Ile14; a conserved element of Cluster A is active-site residue (Tyr100) positioned behind the donor carbon (CD) within the nicotinamide ring of dinucleotide cofactor. In Cluster B, residue Arg57 and Phe31 facilitate the movement of acceptor carbon (CA) toward the donor carbon (CD). In the active-site, the motions in these networks control the electronic environment; the role of dynamical active-site residues with NADPH/NADH has been implicated in the ring puckering, which has been suggested

as a contributor to the reaction coordinate and an important event in the hydride transfer mechanism [277, 10]. The ring puckering alters the aromatic character (hybridization state of CD changes from sp3 to sp2) thus allowing the hydride to leave. Similarly motions alter the environment around the acceptor carbon (CA changes from sp2 to sp3) making it suitable for the incoming hydride. Our recent studies have also indicated that these interactions, as well as the protein fluctuations in this network, are a conserved part of the DHFR fold (see Chapter 6).

In HBBR, present within Cluster A is the surface hydrogen bond Ile176N- His153O, which relays the motions to Pro151, positioned behind the CD. Similarly within Cluster B, starting from the surface regions consisting of two helices and the associated loop regions that are interconnected by hydrogen-bonds between Ser88 - Asp131, motions are transferred to the active-site through Leu125 positioned behind NA of the substrate. Similar to DHFR, slow conformational fluctuations in these network residues lead to a decrease in the donor-acceptor distance.

In DHPR, motions of Cluster A network hydrogen bond Thr199O-Asp182N and the residue Pro178 positioned behind the nicotinamide ring impact CD. On the other side, the motions of Cluster B network hydrogen bond between Lys150 and Ser107, and the residue Trp86 behind the substrate ring impact CA position; therefore, facilitate the hydride transfer. Similarly in PR, these motions are present in Cluster A network hydrogen bond Ser252N-Leu226O and Pro224 behind CD, and Cluster B network hydrogen bond Asn147-Lys198 and residue Tyr194 behind CA. An important observation is that in HBBR, DHPR, and PR Cluster B consist of edges of two helices and the neighboring surface loop regions, which interact through a number of other hydrophobic and hydrophilic interaction in addition to the ones mentioned above. Additional interactions are expected to be present within the network of coupled motions.

Beyond the active-site, the role of these preserved network interactions possibly lies in the coupling of enzyme motions to those of the surrounding solvent. Previous computational as well as experimental investigations have revealed that the motions in the solvent enslave and drive protein dynamics. Thermodynamical fluctuations of the hydration-shell and the bulk solvent [5] reach the active-site (through the interactions in the networks) where small changes in the enzyme-substrate interaction control the progress of the reaction near the transition state. It has also been shown that the energy, from these fluctuations, alters the recrossing behavior of the dynamical trajectories near the transition state, so that more trajectories become successful.

7.4 Conclusions

In this chapter, we have characterized the protein flexibility in an enzyme super-family of dinucleotide binding proteins. Four diverse members (DHFR, DHPR, PR and HBBR) with very low sequence homology and considerable difference in the structural topology show conserved protein flexibility coupled to the common hydride transfer step. There are three regions of protein that show similar flexibility: cluster A, located near the dinucleotide binding pocket; cluster B, located near the substrate binding pocket; and cluster C, located near the adenosine binding pocket. Each of these clusters show presence of conserved pathways that start from surface regions and lead into the inside of the enzyme. In the active-site, the arrangement of conserved residues subtly controls the electronic environment to favor the product state. Therefore, consistent to with the widely accepted view the structural arrangement of these conserved residues is a critical element of the 3-dimensional make of the enzyme structure or the overall fold. Other diverse members of DBRP super-family confirm the preservation of the three flexible clusters as well as structural features whose dynamical motions impact catalysis (see Fig. 7.7). In addition, the results presented here also indicate that for the system studies the conserved interactions and protein vibrations are also a conserved part of the overall enzyme fold. We therefore hypothesize that the flexibility signature of the enzyme-fold is also an important characteristic that is required for defining an enzyme super-family. The multiple pathways observed in the enzyme fold may possibly provide robustness to the enzyme for performing the designated function.

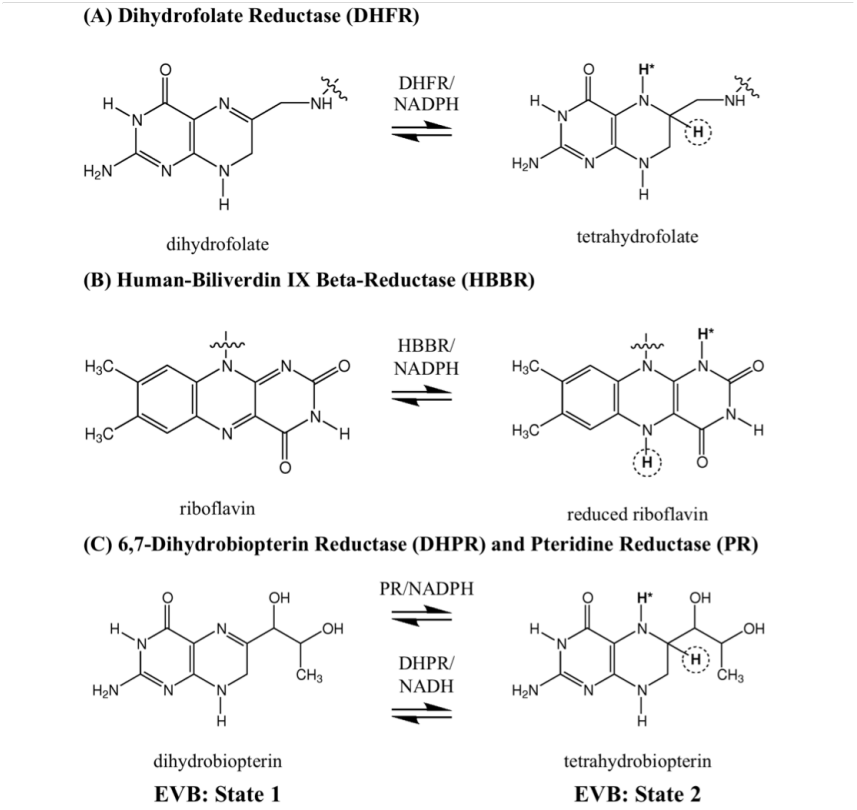

Figure 7.1: **Hydride transfer catalyzed by DBRP enzymes investigated in this study.** Each of the reaction requires binding of the cofactor (NADH or NADPH) to the enzyme. The conversion of substrate requires the transfer of a proton (indicated by H*) and hydride (indicated by circles H). The cofactor NADH/NADPH serves as a hydride donor.

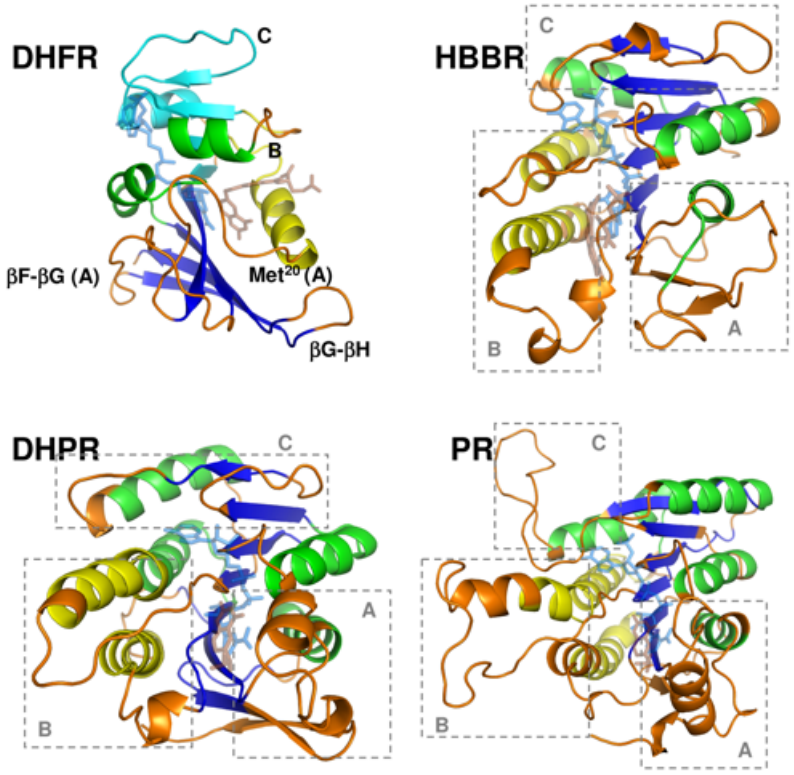

Figure 7.2: **Dynamical clusters in the 4 enzymes reveal common dynamical behavior**. HBBR, DHPR and PR show the presence of four dynamically coupled regions, where as DHFR shows the presence of five clusters. The substrate (brown) and co-factor (light blue) in each protein is shown using a transparent stick representation.

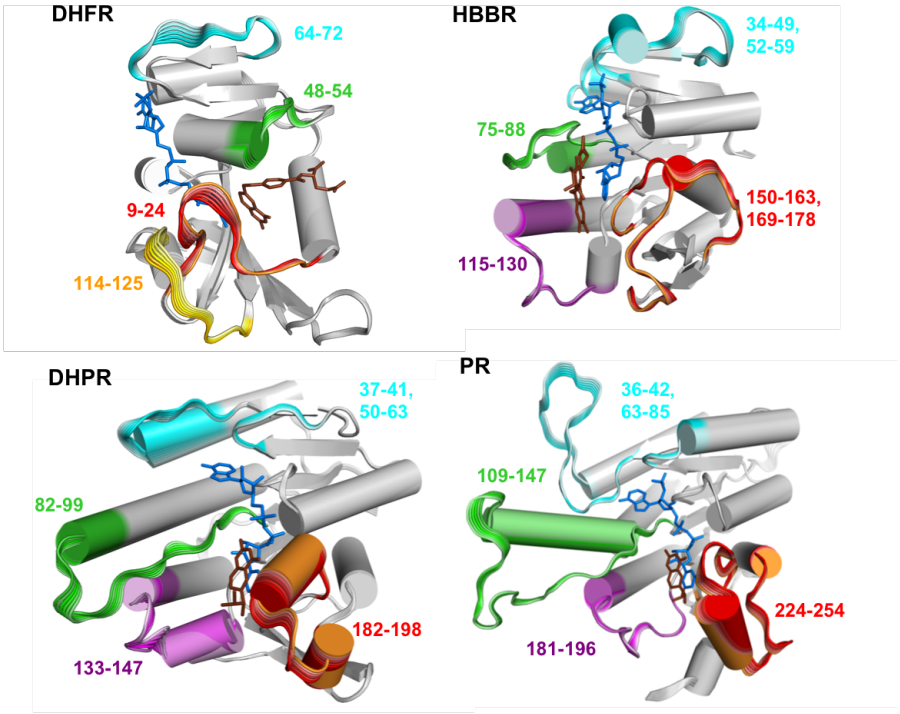

Figure 7.3: **Slow conformational fluctuation coupled to the hydride transfer catalyzed by the four enzyme systems investigated in this study**. The protein vibrational mode is depicted in a movie like fashion with subsequent frames shown in lighter colors. Corresponding regions displaying large motions are colored and labeled. The substrate (brown) and the NAD(P)H cofactor (blue) are shown as sticks; the motions of substrate and cofactor have been omitted for clarity. Movies of these and two additional modes are provided in supplementary information.

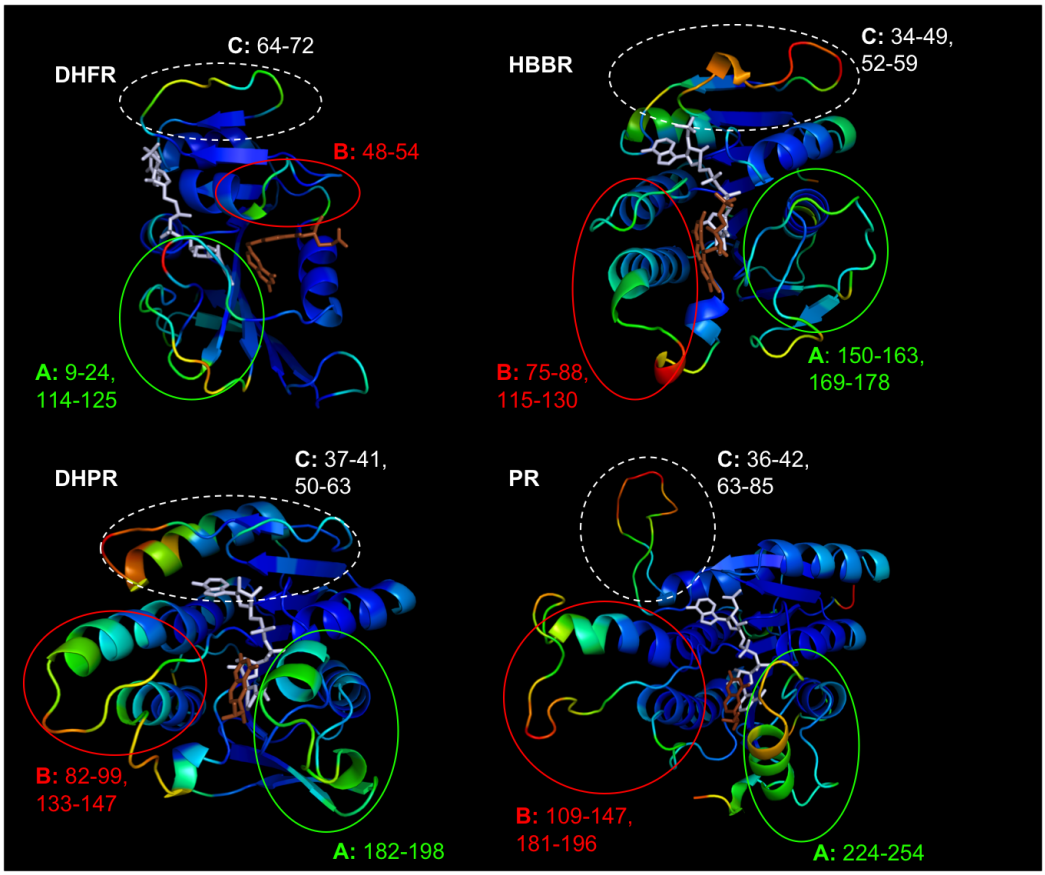

Figure 7.4: **Common dynamical characteristics of dinucleotide binding Rossmann fold proteins (DBRPs) catalyzing hydride transfer**. Aggregate atomic displacements from top 10 reaction coupled modes (obtained by summing the displacement vectors in the modes) are shown with the dark blue regions corresponding to rigid protein, while the red (yellow and green as well) indicates regions displaying large movements coupled with the reaction. Cofactor (orange) and substrate (white) are shown as sticks. The residues corresponding to Cluster A, Cluster B and Cluster C are identified.

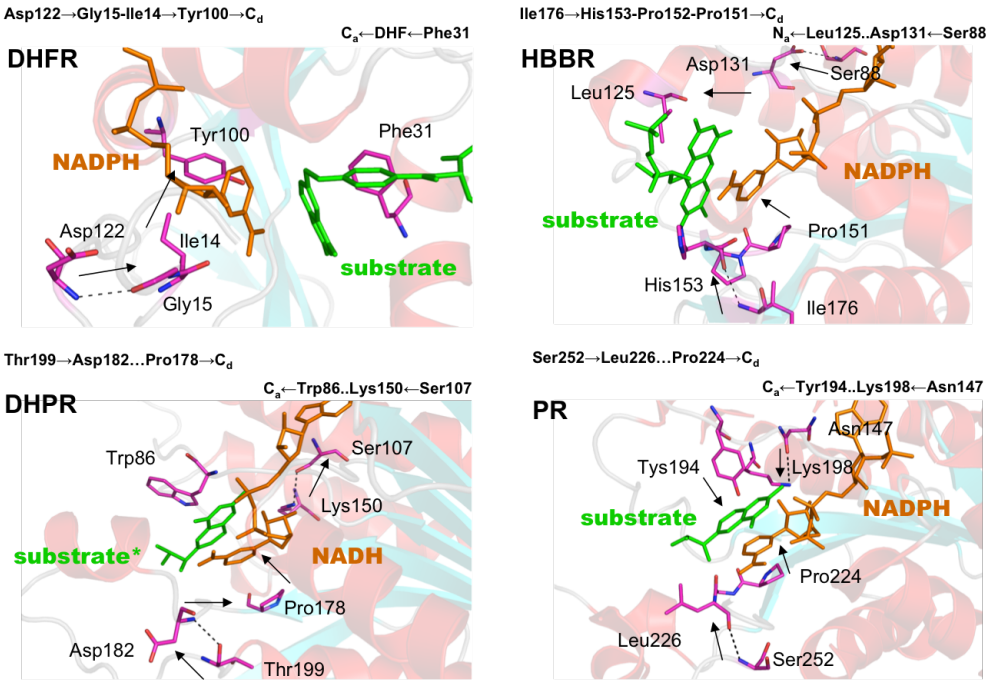

Figure 7.5: **Conservation of network residues in DBRP enzymes**. These networks were identified based on correlated motions and structural analysis. These networks start on the enzyme surface and span all the way to the active-site (indicated by arrows), where they facilitate the movement of the donor (CD) and acceptor (CA or NA) atoms for the hydride transfer.

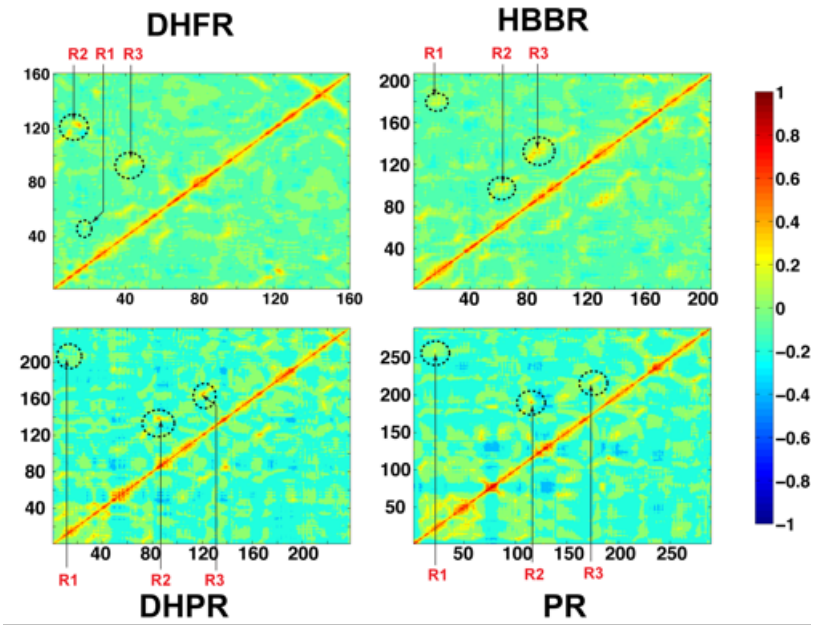

Figure 7.6: **Cross-correlations observed along the reaction profile**. Positively correlated regions are identified by darker shades of red, while negatively correlated regions in the proteins are identified by darker shades of blue. Along the diagonal, positive correlations are observed for secondary structure elements, namely the α-helices and β-sheets. For clarity of presentation, the marking of secondary structural elements is not shown on the correlation plots. Three distally separate regions marked R1-R3 show correlated movements in the protein. These regions are part of a network of coupled motions identified for the super-family of enzymes (see main text). The extent of these four regions is identified in the table below. Note, in DHFR, the regions R1 and R3 are inter-changed depending on the placement of the substrate-binding region, which is located on the opposite side of the enzyme compared to HBBR, DHPR and PR. The positive correlations along R1 are weak in all the four enzymes studied, while R2 and R3 show strong positive correlations.

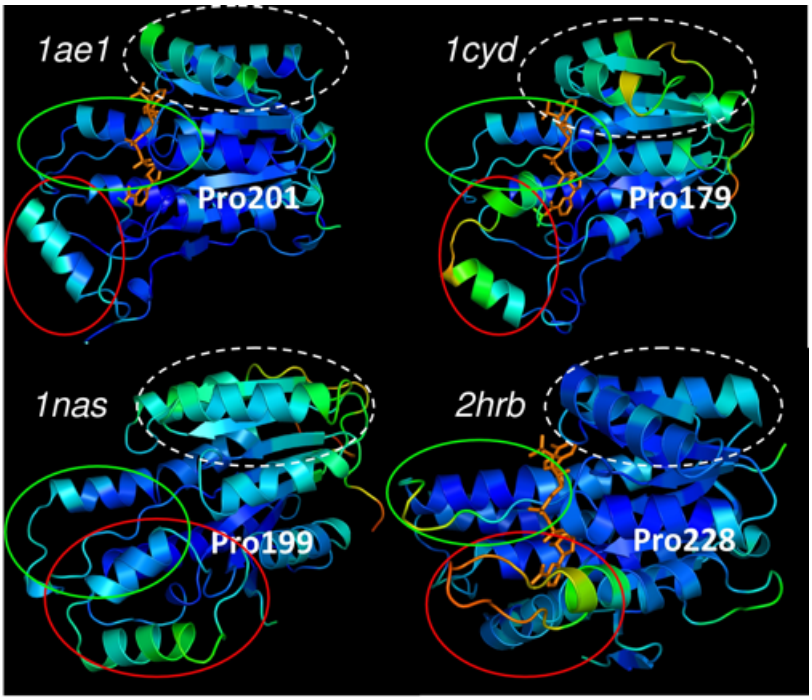

Figure 7.7: **Conserved flexibility in the Clusters A. B, C and conserved network residues in other members of the super-family**. The flexibility based on the X-ray temperature factors is depicted. The conserved residue in Cluster A behind the cofactor nicotinamide ring is listed. 1ae1: Tropinone reductase from Datura stramonium; 1cyd: Carbonyl reductase from Mus musculus; 1nas: Sepiapterin reductase from Mus musculus; and 2hrb: Carbonyl reductase from Homo sapiens.

Chapter 8

Conclusions

In this chapter, we summarize the major contributions of this thesis. We critically examine the outcomes from each of the questions outlined in the introduction (Chapter 1) and analyze them in the context of understanding both protein dynamics and enzyme catalysis. We also outline future work that would arise out of this thesis. In the first part, we present our perspectives on protein dynamics and how we can build better models to describe protein flexibility in the context of its function. In the second part, we discuss our findings in the context of enzyme catalysis and protein design. Finally we conclude with a summary on how our knowledge could be combined with experimental design to better understand how enzymes (and proteins) work.

8.1 Protein Dynamics at Multiple Time-scales

8.1.1 Relation to Previous Work

It is important to set the context of our work in perspective of prior work. Most studies subscribe to the protein's conformational landscape as being anharmonic, involving multiple conformational sub-states that share significant structural and energetic similarities [94]. The means of describing the conformational landscape spanned by a protein has typically relied on highly effective techniques such as normal mode analysis [185] and elastic network models (ENM) [28]. Recent studies on protein dynamics in several important drug-targets have revealed that intrinsic dynamics dominates the binding motions in recognizing the substrate [30]. Analysis using PCA techniques to characterize collective fluctuations in proteins have also revealed that proteins tend to sample substrate bound and

unbound forms within the same ensemble [170].

Our work on characterizing protein dynamics at multiple time-scales have focused on ubiquitin. While previous work using both experimental [170] and theoretical [30] techniques have highlighted the importance of intrinsic dynamics in function, our studies reveal richer detail about how ubiquitin achieves its ability to bind to diverse substrates. As presented in chapter 3, Quasi-Anharmonic Analysis (QAA) revealed that ubiquitin exhibits motions that allow it to adapts its binding region extensively (as presented in Fig. 3.7).

QAA also reveals motions that modulate the binding regions exquisitely to adapt conformations that resemble that of the bound and unbound forms. QAA identifies isoenergetic conformational sub-states (see Fig. 3.8), not revealed by both PCA based techniques (Fig. 2.7) or NMA based techniques. These isoenergetic sub-states can also be further decomposed into a natural hierarchy showing how ubiquitin could potentially modulate its binding regions can adapt to potentially large number of substrates. The natural hierarchy can also be used to interpret motions along every sub-state, indicating how motions that are global in Level 1 of the hierarchy turn into potentially local motions in Level 2 (and subsequent decomposition). All these observations complement both previous work and add to the richness of the description of the conformational landscape spanned by ubiquitin at μs time-scales.

8.1.2 Summary of Contributions

We began by examining, in detail, two questions outlined in the introduction section of the thesis. The first question, asked how one could characterize the nature of protein *dynamics* at multiple time-scales. The current state-of-the-art techniques to investigate protein flexibility were considered. We used quasi-harmonic analysis (QHA) to examine large-scale, collective conformational fluctuations by approximating the fluctuations observed from a molecular dynamics (MD) simulation (and experimental ensembles) as a single harmonic well. We used ubiquitin to describe the equilibrium conformational landscape. By effectively accounting for the structural heterogeneity observed from different crystal structures and sampling the landscape from different starting structures, we were able to obtain an ensemble that could describe the μs-scale root mean-square fluctuations observed via NMR experiments. We also found that the diagonal and off-diagonal elements along the covariance matrices as determined from the NMR ensemble and the MD ensemble shared significant similarities.

QHA approximates the conformational landscape by pursuing variance, which is a second order statistic. Variance, as a statistical measure is sensitive to the outliers in

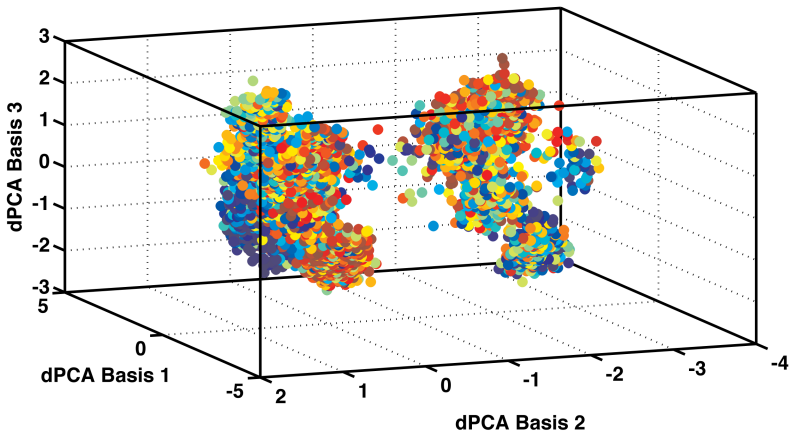

Figure 8.1: **Ubiquitin landscape represented by the first three basis vectors using dihedral PCA [14]**. While the projected conformation show the presence of different clusters, observe that they lack homogeniety when colored by the scaled internal energy, unlike the QAA basis in Fig. 3 of the main text.

the underlying data. Hence, rare occurrences of extreme fluctuations in an experimental ensemble or MD simulation can lead to significant (and potentially adverse) effect on interpreting the conformational landscape. Further, on examining the top-most modes of motion spanned by the QHA vectors in ubiquitin, the conformational landscape, as spanned by these eigenvectors do not show much similarity in terms of their conformations or internal energy distributions (see Fig. 2.7). We note that this heterogeneity is also observed as in the case of performing QHA in internal variables (ϕ-ψ) dihedral angle space; see Fig. 8.1). This heterogeneity in both internal energy values and conformations means that it is difficult to interpret the conformational sub-states spanned by the protein using QHA.

These observations prompted us to characterize the statistics of positional fluctuations in ubiquitin at a much finer detail (Chapter 3). Our observations, in line with previous work, showed that ubiquitin typically exhibits anharmonic (non-Gaussian) fluctuations under ambient conditions. While this is not surprising, we observed that up to 60% of the protein can potentially exhibit anharmonic fluctuations along x, y and z directions. Further, we also noted that the anharmonic motions were dominant along the functionally relevant regions of ubiquitin; hence one would have to account for these anharmonic distributions in these regions to obtain a relevant and useful description of the conformational

landscape spanned by ubiquitin.

In Chapter 3 of this thesis, we built a novel computational model to describe the anharmonic conformational fluctuations in the ubiquitin landscape. We termed this computational model *quasi-anharmonic analysis* (QAA), since it builds a linear model of anharmonic fluctuations. The results from the application of QAA on the ubiquitin landscape illustrated the hierarchical organization of the landscape. Pursuing anharmonic fluctuations leads us to visualize the landscape in terms of sub-states that are not only homogeneous in terms of their conformations, but also in terms of their overall internal energy distributions. The motions described at every level of the landscape further show the inherent ability of ubiquitin to modulate its binding regions to accommodate a wide variety of substrates. The ability to describe anharmonic fluctuations in the context of a protein's conformational and energetic landscape provides a number of opportunities which are described in the next sub-section 8.1.3.

In a complementary effort we outlined an *online* or on-the-fly approach to characterize collective dynamics as MD simulations are progressing in Chapter 4. This was based on a novel representation of MD simulations as *tensors* and tracking how distance maps changed over the course of a MD simulations. We demonstrated our approach on three different proteins: barnase, ubiquitin and cyclophilin A. We showed that our technique can identify constrained and flexible regions and also characterize the dynamical coupling between different regions of the protein. Further, we also showed that it was possible to identify time-points along the simulation where collective behavior differed significantly from the previous time-window(s). The ability to track and monitor MD simulations online offers a powerful means to explore conformational landscape given the context that MD simulations do suffer from sampling inefficiencies. Further work on incorporating online analysis and improving conformational sampling capabilities are explored in subsection 8.1.4.

We also examined the conformational landscape spanned along a reaction pathway for the enzyme cyclophilin A (CypA). When the reactant and product states in the enzyme show relatively small changes with respect to their overall conformations, QHA based description of the enzyme-substrate complex along the reaction pathway and equilibrium simulations of just the reactant and product state share significant overlap in terms of the large-scale fluctuations. However, this overlap in terms of the collective motions provides only a qualitative description of the landscape and must not be interpreted as being directly related to the catalytic activity of the enzyme. A careful evaluation of these collective motions along the reaction co-ordinate is necessary before one can conclude their relevance to the catalytic step of the enzyme (see Chapters 5 and 6).

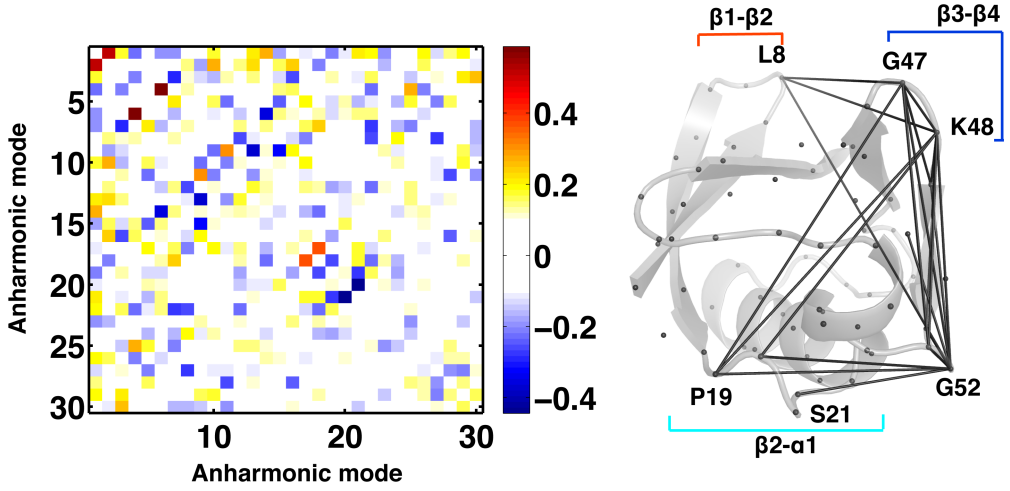

Figure 8.2: **Coupling between anharmonic modes of motion (η).** Most anharmonic modes are weakly coupled as indicated by the coupling co-efficients (left; $|cc| > 0.3$). An example of anharmonic coupling between modes 1 and 2 ($|cc| = 0.41$) showing spatially coupled regions in the protein. Observe the long-range coupling between R1 and $\beta_2 - \alpha_1$. C^{α} atoms are shown as gray spheres and residues commonly activated by modes 1 and 2 are marked and connected by gray lines.

8.1.3 Quasi-anharmonic Analysis

Coupling between Anharmonic Modes of Motion. Unlike QHA, the anharmonic modes of motion from QAA need not be orthogonal. Hence, it is possible for these anharmonic modes to activate each other depending on their intrinsic coupling, given by:

$$\eta = A_i^T A_j / \left(||A_i||\ ||A_j|| \right) \tag{8.1}$$

We illustrate the coupling in the anharmonic modes for the protein ubiquitin (see Chapter 3). As shown in Fig. 8.2, observe that most modes are weakly coupled [198]. Consider modes 1 and 2; mode 1 shows global fluctuations involving R1 and R2 whereas mode 2 activates motions along $\beta_2 - \alpha_1$ and R1. As illustrated in Fig. 8.2 (right), commonly activated residues and their interactions were identified by thresholding the matrix η based on interaction strength. These specific activation patterns along particular anharmonic modes of motion may provide additional insights into how energy transfers from local to global conformational fluctuations [61, 177].

Understanding the Thermodynamical Basis for QAA. The conformational sub-states identified via QAA share significant similarities in terms of their structure and internal energy distributions. A closer and more rigorous understanding of these sub-states will arise when one can interpret them in terms of free-energy changes required to "jump" between the various sub-states. We propose two feasible approaches to examine the thermodynamical basis for QAA.

The first approach uses a novel Markov Random Field (MRF) [148, 147] representation of a protein structure which can then be used to evaluate the free-energy propensity for an ensemble of conformations generated by altering the side-chains. This approach has been successfully used in the past to examine free-energy changes involved in protein-protein interactions [147] and understand the sequence/ structure specificity in proteins [146]. MRF based representation combine the powerful tools of statistical mechanics with efficient algorithms from computer science and thus be used to characterize the free-energey landscape as described by the QAA basis vectors. As illustrated in Fig. 8.3, an MRF based coloring of every conformation from the ubiquitin landscape shows considerable homogeneity in terms of the free-energy values. This observation, in spite of using Rosetta force-field parameters and rotamer discretization for the MRF representation, implies that the QAA interpretation of the landscape is robust (and independent) of issues related to force-field selection and type of MD simulation used to sample the landscape [1].

The second approach is based on the observations from QHA of the configurational entropy S, which is an upper limit in terms of $3N$ independent classical harmonic oscillators. For a protein, large-scale conformational changes involving conformational transitions between multiple states and the high level of anharmonicity in the landscape almost always means that the entropy estimations from QHA are an over-estimate. Recently, Lange and Grubmuller proposed a novel approach, Full Correlation Analysis (FCA) [168], which builds an orthogonal basis description of the landscape spanned from an equilibrium simulation and accounts for linear- and non-linear correlations. Based on FCA, mutual information expansions [17] and adaptive kernel density estimation [122], the authors also estimated configurational entropy that was shown to be better than QHA estimates of the entropy [123]. This approach could potentially be a starting point for estimating entropy using QAA.

QAA in Internal Variable (Dihedral) Space. While analyzing large-scale fluctuations in cartesian coordinates, it is important to align the protein to some reference structure before performing QHA or QAA. This is often challenging, especially in the context of

[1] This work is currently in preparation: Arvind Ramanathan, Xin Gao, Christopher J. Langmead, *Mechanistic Interpretation of the Binding Energy Landscape of Ubiquitin*. 2010.

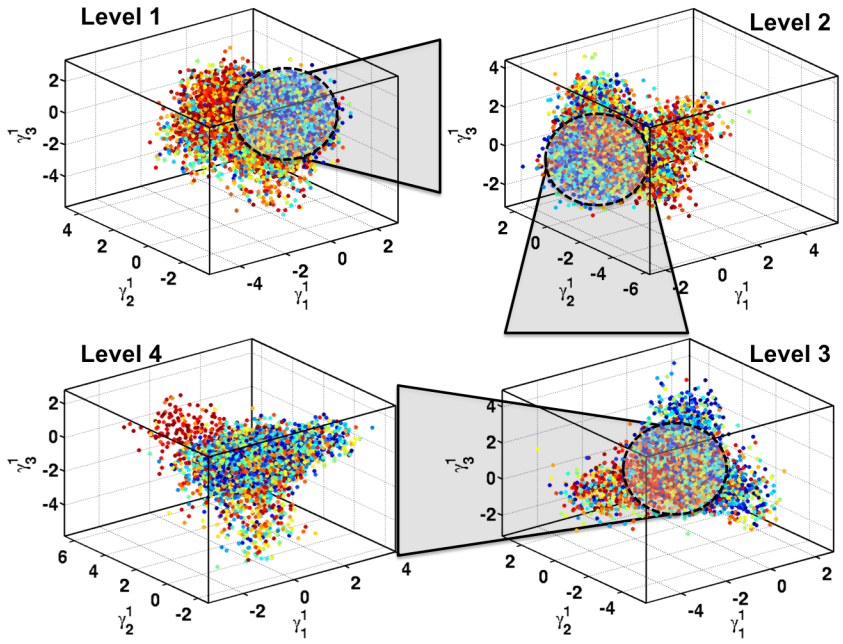

Figure 8.3: **Ubiquitin landscape spanned by the top 3 QAA basis vectors from 1YIW [32] and simulated via Desmond for 150 ns in explicit solvent conditions.** Each conformer represents a local ensemble where by side-chains are flexible sampled through a discrete rotamer library and a free-energy associated with the local ensemble is plotted as a function of the top 3 QAA basis vectors. A clear separation in states with higher free-energy (red colors) and lower free energy (blue colors) can be seen.

protein folding simulations or large conformational changes. Alternatives to this often require the use of internal variable descriptors such as dihedral angles. A simple and straight-forward implementation of dihedral PCA (dPCA) was carried out (see Fig. 8.1) and we observed that this description of the landscape, while revealing several clusters of related conformations, still fails to reveal any energetic homogeneity in the conformations. This indicates that even in the dihedral space, the the dynamics in the $\phi - \psi$ space may have significant anharmonicity, which may be captured by using QAA. There are also indications that the coupling arising in positional fluctuations (and dihedral fluctuations) can be potentially non-linear [168]. Hence it will be of interest to integrate techniques such as QAA with techniques introduced by Schroder and collegaues [237] to see if accounting for non-linear dependencies in the positional fluctuations can further improve our understanding of the conformational landscape.

8.1.4 Dynamic Tensor Analysis: Applications to Improved MD Sampling

Tracking of multiple tensor streams for MD simulations. A fairly straightforward extension to DTA would be the ability to track multiple streams of data simultaneously. A recent implementation of this algorithm has been made available in the data-mining community; however for MD simulations, such a tool needs to incorporate a flexible framework to track multiple features. By tracking multiple streams of co-evolving tensors (such as hydrogen bond/ hydrophobic interaction distances, electrostatic field/ potential tensors, etc.), it would be possible to answer questions such as: (a) how movement of one residue (for example C^{α} displacement) affects the formation of a hydrogen bond?, (b) how do variations in electrostatic potentials affect motions in different parts of the protein? and so on. The answers to these questions can potentially help unravel mechanistic details of how proteins function.

Improved sampling of *rare events* during MD simulations. An advantage of DTA is that it can track changes in collective motions even though these changes may imply minor changes in terms of overall RMSD values. This in turn can be exploited to potentially sample around such events, which represent rare excursions from the observed fluctuations. This provides the biologist with the ability to fork off simulations from each of these excursions to potentially gain insights into the nature of how the landscape has changed. Since most MD simulations do suffer from insufficient sampling, the ability to fork off multiple simulations will provide another means to improve the sampling of the conformational landscape. Further more, changes in collective behavior are of relevance

in biology (in the context of enzyme catalysis) and hence incorporating DTA into standard MD simulation software could be of great advantage.

8.2 Role of Protein Motions in Enzyme Catalysis

In the first part of the thesis, the overarching theme was to characterize the internal motions of proteins at various time-scales. For this purpose, we examined the equilibrium conformational landscape of ubiquitin, a well characterized protein. In the second part, we focussed on the dynamical hypothesis and its effect on enzyme catalysis. Specifically, we were able to gain some insights on how enzymes, over the course of evolution have been constrained to maintain certain "dynamical features" that allow them to carry out their biochemical function. This is significantly different from dynamical features that play a role in substrate binding, since the motions detected are along the progress of the chemical step of the enzyme.

8.2.1 Relation to Previous Work

From the perspective of enzyme catalysis, we were mostly drawn to the question of if and how intrinsic motions in the protein fold affect the biochemical step of an enzyme reaction. Most previous studies have focused on either large-scale studies of sequence alignments of enzyme super-families [19, 101, 102, 196, 209] or analyses of structural profiles along the active site(s) of enzymes to identify molecular function [134]. Dynamical contributions to enzyme-substrate binding have been the focus of several computational and theoretical studies including Bahar and co-workers [24, 30, 287]. While certainly insightful about the nature of intrinsic motions and their impact on substrate binding, the coupling of intrinsic dynamics to the catalytic (chemical) step of the enzyme has only recently attracted attention. Experimental [49, 78, 125, 47] and computational modeling [7, 8, 9, 4, 5] have now revealed the presence of coupled networks of interactions in several enzymes which play an important role in promoting the chemical step of the reaction.

Our studies have now revealed, in a complementary fashion, that these networks are a conserved feature of the enzyme fold across multiple enzymes including cyclophilin A, dihydrofolate reductase and ribonuclease A. We also found a remarkable feature shared by enzymes catalyzing the same biochemical reaction but no structural similarity whatsoever: the dynamics coupled to the reaction step of these enzymes showed similar profiles over the reaction co-ordinate. This remarkable similarity in the dynamics coupled to the reaction step indicates that for the chemical step to progress forward, these motions (at

potentially multiple time-scales [238]) are critical. Our observations also extend to the Rossmann family of enzymes that share a mechanistic sub-step of hydride transfer.

8.2.2 Summary of Contributions

The knowledge of evolutionary constraints on enzyme catalysis could prove to be a powerful tool to design and even enhance the function of novel enzymes. Efforts from Baker and co-workers [141, 171] have already resulted in designing novel enzymes. However, efforts to artificially evolve enzymes in labs have proven to be extremely hard and the enzymes designed through such approaches are several orders of magnitude less efficient than the naturally occurring enzymes [227]. Thus, the most intriguing aspect of enzyme catalysis is whether we can develop tools and technologies that enable (or predict) the design of highly efficient enzymes.

Perhaps, the knowledge of intrinsic dynamics and their role in catalysis could accelerate this field. Our approach to analyze the mechanistic behavior of enzymes is via a rigorous sampling of the conformational landscape along a pre-defined reaction coordinate. While the approach has merits and certain limitations, our observations about the collective conformational fluctuations before, during and after catalysis from the enzyme CypA indicate that the large-scale collective fluctuations overlap consistently. Apart from the large-scale fluctuations observed in multiple states showed remarkable similarity (substrate unbound, reactant, product and over the course of the isomerization reaction), the projections from each of the simulations also showed large overlaps in the space spanned by the lowest frequency QHA based modes.

In Chapters 6 and 7 of this thesis, we analyzed if the motions coupled to catalysis imposed evolutionary constraints. Our studies have revealed complementary insights to several studies that have focused on enzyme-substrate binding. In the first study, we were able to show that enzymes catalyzing the same biochemical reaction, in spite of sharing no apparent structural homology exhibit remarkable similarity when it came to dynamics coupled to the catalytic step. In the case of CypA and Pin1 isomerases, the critical active site residues were surrounded by highly flexible loops connected by a network of hydrogen bonds (and hydrophobic interactions) that influenced the catalytic process. As observed from R67 DHFR and E. coli DHFR, several critical profiles of distance variations along the reaction pathway are very similar. This similarity in dynamics coupled to the catalytic step, in spite of having no apparent structural homology underlines the importance of dynamics in the catalytic process. As a control example, where angiogenin and ribonuclease A shared absolute structural homology, yet angiogenin was at least 5 orders of magnitude less efficient than ribonuclease A illustrates the importance of the network regions; although

residues important for catalysis are conserved as part of the fold, the interactions (which are part of the network) are absent in angiogenin, which may explain why angiogenin is a poor catalyst in comparison to ribonuclease A.

It is remarkable to note that enzymes sharing a mechanistic sub-step common in catalysis do show similar dynamical behavior. A notable observation amongst the enzymes investigated is that these enzymes do share structural homology; however, the topology (or organization) of the secondary structural elements is quite different. It is also important to note that the topology of these enzymes give rise to different orientations of the binding sites. In spite of the variations in orientations of the binding site, there are remarkably similar dynamical features the promote the hydride transfer in these enzymes. While a conserved proline (in case of DHFR, Tyr 100) across multiple members of the family played a role in altering the environment of the active-site, a hydrophobic residue located behind the substrate aided optimal alignment for hydride transfer to take place. Note that these residues are located along flexible regions of the protein and form a connected network that act in concert to promote the reaction progress. Further, from an evolutionary perspective, the identified residues are conserved across multiple enzyme members of the super-family.

On the one side, while these studies provide unique insights into the function of such enzymes, it must be noted that experimental evidence for such mutations along the proposed networks are quite hard to come by. Of the 767 enzymes available, only two or three enzyme systems have kinetic parameters derived from experiments under a variety of conditions that mutate the network regions. The availability of such studies especially for carbonyl reductase (CBR) with a variety of substrates bound at the active site provides ideal positive and negative control strategies to test the proposed dynamical hypothesis. CBR is an interesting test case since the mutations along the network regions provides evidence of both enhancing and diminishing the catalytic efficiencies of the enzyme [79, 222]. Further, as a recent study [144] showed, mutations to the network residues in R67 DHFR showed remarkable alterations to the catalytic power of the enzyme. Hence, such a study of how mutations to the network residues can enhance/ diminish enzyme catalysis would be of great benefit to test our theories.

The principled understanding of how motions in an enzyme is coupled to its catalytic step also brings about knowledge about how proteins may have evolved. For example, statistical coupling analysis (SCA) [184] has been used in a number of application contexts: to study evolution of multiple proteins [250], design proteins [246, 232] and study organization of enzyme structure [109]. Our insights enhance these observations by elucidating the mechanistic roles of how and why certain networks of residues must be constrained by evolution to maintain function by qualifying the specific role they play in catalysis. Fur-

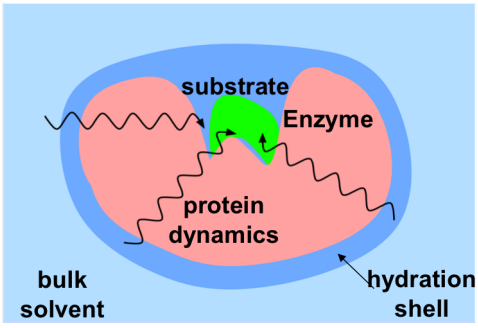

Figure 8.4: **Emerging paradigm of protein dynamics and catalysis.**

ther, knowledge about these networks could enhance one's ability to assess why and how certain residues co-evolve.

8.2.3 Understanding the energetic coupling between solvent fluctuations, enzyme motions and catalysis

An emerging paradigm in enzyme chemistry is that protein motions at all time-scales (from femtoseconds/ picoseconds at the active site to microseconds and beyond collectively throughout the protein) contribute to enzyme catalysis. Although there is much debate, initial studies on the constraining slow conformational motions coupled to the reaction coordinate indicate that they impose a significant barrier on the catalytic process. Furthermore, constraining/ enhancing solvent motions in and around these flexible loop regions of the protein exhibiting slow conformational fluctuations also significantly impacts how successful an enzyme can catalyze its reaction. Based on this, an emerging picture about the relationship between enzyme dynamics and catalysis is arising, as illustrated in Fig. 8.4.

According to this emerging picture, along with structural interactions, internal protein motions at fast time-scales (femto/picoseconds) provide crucial support to alter the chemical environment within the active site, favoring the catalytic step to move forward to the product state [200, 238]. In the absence of co-factors or other promoting molecules (as is the case with CypA), the fluctuations at the flexible surface regions of the protein, enslaved by the solvent [84, 85, 91, 6], provide energy to overcome the activation barrier. The energy from the solvent is first transferred to the flexible surface regions of the protein, which then "trickle" to the active site of the protein via the networks of interactions (as

well as motions in these networks), which then provide the necessary energy to overcome the barriers associated with catalysis [177]

This raises the question if there are inherently "engineered" pathways in the protein through which energy can flow between the flexible regions of the protein to the active site. Topology-based approaches using simple models such as GNM [23], and an integration of information theoretic concepts quantifying the signal transduction processes within an enzyme's three dimensional structure [61] have already indicated that catalytically important residues in an enzyme are better receivers of information than the other residues in the protein. We are actively engaged in evaluating this proposal by combining biophysical techniques involving novel MD simulation techniques with information theoretic approaches to study if an enzyme's catalytic residues are indeed better conductors of energy through them.

8.2.4 Enzyme Catalysis: Allosteric Regulation from distal residues

The presence of energy-flow pathways within the protein raises another question, which is more subtle and difficult to answer. This question relates to our knowledge of allosteric regulation within an enzyme. As recently pointed out by Goodey and Benkovic [105], allostery and catalysis arise via conformational dynamics being the bridge between them. There are several enzymes, known in the literature, where mutations to a distal site in the protein, as far away as 20 Å, close to the surface of the protein, leads to significant effect on catalysis. As ably demonstrated by Benkovic and Ranganathan (and co-workers) [174], modulating these surface residues can bring about novel activation/ inhibition processes that can aid the drug-discovery process.

The approaches used here clearly are applicable in better understanding and characterizing the role of allostery in enzyme catalysis. Particularly, the networks discovered extend from the flexible surface regions all the way to the active site, connected via hydrogen bonds and hydrophobic interactions. Altering site-specific properties may alter allow one to possibly characterize the long-range effects on controlling the catalytic step of the enzyme. Availability of such tools could clearly lead to the design of alternate binding sites or engineering novel drugs that could bind at the allosteric sites to modulate function. Indeed, it is our hope that the discovery of allosteric "hot-spots" on the entire protein could lead to novel insights into the systemic regulation of cells and enhance our understanding of how cells function.

Bibliography

[1] Amadei A., M.A. Ceruso, and A. Di Nola. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics. *Proteins: Struct., Funct. and Bioinformatics*, 36(4):419–424, 1999. 1.1.3, 2.1

[2] E. Acar, C. Aykut-Bingol, H. Bingol, R. Bro, and B. Yener. Multiway analysis of epilepsy tensors. *Bioinformatics*, 23(13):i10–18, 2007. 4.3.4

[3] K R Acharya, R Shapiro, S C Allen, J F Riordan, and B L Vallee. Crystal structure of human angiogenin reveals the structural basis for its functional divergence from ribonuclease. *Proceedings of the National Academy of Sciences of the United States of America*, 91(8):2915–2919, 1994. 6.1

[4] P. K. Agarwal. Cis/trans isomerization in hiv-1 capsid protein catalyzed by cyclophilin a: Insights from computational and theoretical studies. *Proteins: Struct., Funct., Bioinformatics*, 56:449–463, 2004. (document), 1.2.2, 5.1.1, 5.3.2, 5.3.2, 5.3.2, 5.6, 6.1, 6.2.1, 6.3.1, 7.3, 8.2.1

[5] P. K. Agarwal. Role of protein dynamics in reaction rate enhancement by enzymes. *J. Amer. Chem. Soc.*, 127:15248–15256, 2005. 1, 2.1, 4.1, 5.1.1, 5.4, 6.1, 6.5, 7.1, 7.3, 8.2.1

[6] P. K. Agarwal. Enzymes: An integrated view of structure, dynamics and function. *Microbial Cell Factories*, 5, 2006. 1, 1.1, 1.2.2, 2.1, 4.5, 6.1, 7.1, 8.2.3

[7] P. K. Agarwal, S. R. Billeter, and S. Hammes-Schiffer. Nuclear quantum effects and enzyme dynamics in dihydrofolate reductase. *J. Phys. Chem. B*, 106:3283–3293, 2002. 1.2.2, 2.4.2, 6.1, 6.2.1, 7.3, 7.3, 8.2.1

[8] P. K. Agarwal, S. R. Billeter, P. T. R. Rajagopalan, S. Hammes-Schiffer, and S. J. Benkovic. Network of coupled promoting motions in enzyme catalysis. *Proc. Natl. Acad. Sci. USA*, 99:2794–2799, 2002. 1, 1.2.2, 2.1, 4.5, 6.1, 6.2.1, 7.2, 7.2.1, 8.2.1

[9] P. K. Agarwal, A. Geist, and A. Gorin. Protein dynamics and enzymatic catalysis: Investigating the peptidyl-prolyl cis-trans isomerization activity of cyclophilin a. *Biochemistry*, 43(33):10605–10618, 2004. (document), 1, 1.1.2, 1.1.3, 1.2.2, 2.1, 2.2.2, 5.1.1, 5.3.1, 5.3.2, 5.3.2, 5.3.2, 5.4, 5.6, 5.7, 6.1, 6.2, 6.2.1, 6.2.1, 6.3.1, 7.3, 8.2.1

[10] Pratul K. Agarwal, Simon P. Webb, and Sharon Hammes-Schiffer. Computational studies of the mechanism for proton and hydride transfer in liver alcohol dehydrogenase. *Journal of the American Chemical Society*, 122(19):4803–4812, 04 2000. 7.3

[11] Sadaf R. Alam, Pratul K. Agarwal, Melissa C. Smith, Jeffrey S. Vetter, and David Caliga. Using fpga devices to accelerate biomolecular simulations. *Computer*, 40(3):66–73, 2007. 1.1.2, 4.1

[12] W. John Albery and Jeremy R. Knowles. Evolution of enzyme function and the development of catalytic efficiency. *Biochemistry*, 15(25):5631–5640, 12 1976. 1.2.1

[13] Cristobal Alhambra, Jose C. Corchado, Maria Luz Sanchez, Jiali Gao, and Donald G. Truhlar. Quantum dynamics of hydride transfer in enzyme catalysis. *Journal of the American Chemical Society*, 122(34):8197–8203, 08 2000. 7.2, 7.2.1

[14] Alexandros Altis, Phuong H. Nguyen, Rainer Hegger, and Gerhard Stock. Dihedral angle principal component analysis of molecular dynamics simulations. *The Journal of Chemical Physics*, 126(24):244111, 2007. (document), 1.1.3, 1.1.3, 8.1

[15] A. Amadei, A. B. M. Lissen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins: Struct., Funct., Genet.*, 17:412–425, 1993. 1.1.3, 2.1, 2.4.2, 4.1

[16] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *NIPS*, pages 757–763. MIT Press, 1996. 3.3.1

[17] Phil Attard, Owen G. Jepps, and Stjepan Marčelja. Information content of signals using correlation function expansions of the entropy. *Phys. Rev. E*, 56(4):4052–4067, Oct 1997. 8.1.3

[18] Gary S. Ayton, Will G. Noid, and Gregory A. Voth. Multiscale modeling of biomolecular systems: in serial and in parallel. *Current Opinion in Structural Biology*, 17(2):192–198, 2007. 4.1

[19] Patricia C Babbitt. Definitions of enzyme function for the structural genomics era. *Current Opinion in Chemical Biology*, 7(2):230 – 237, 2003. 1.2.1, 8.2.1

[20] B.W Bader and T.G Kolda. Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4):635–653, December 2006. 4.4

[21] B.W. Bader and T.G. Kolda. Efficient MATLAB computations with sparse and factored tensors. *SIAM Journal on Scientific Computing*, 30(1):205–231, December 2007. 4.4

[22] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman. Vibrational dynamics of folded proteins. significance of slow and fast modes in relation to function and stability. *Phys. Rev. Lett.*, 80:2733–2736, 1998. (document), 1.1.1, 4.6

[23] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in protein using a single parameter harmonic potential. *Folding and Design*, 2:173–181, 1997. 1.1.1, 8.2.3

[24] I. Bahar, C. Chennubhotla, and D. Tobi. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Cur. Op. Struct. Biol.*, 17:633–640, 2007. 1, 2.4.2, 5.1, 8.2.1

[25] I. Bahar and Q. Cui. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. Mathematical and Computational Biology Series. Chapman and Hall/ CRC, New York, 2003. 1.1.1, 1.1.3, 2.1, 2.3.4, 2.4.2, 4.1

[26] I. Bahar, B. Erman, R. L. Jernigan, A. R. Atilgan, and D. G. Covell. Collective motions in hiv-1 reverse transcriptase: examination of flexibility and enzyme function. *J. Mol. Biol.*, 285(3):1023–37, 1999. 1.1.1

[27] I. Bahar and A. J. Rader. Coarse grained normal mode analysis in structural biology. *Cur. Op. Struct. Biol.*, 15:1–7, 2005. 2.4.2

[28] Ivet Bahar and AJ Rader. Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology*, 15(5):586 – 592, 2005. Carbohydrates and glycoconjugates/Biophysical methods. 1.1.1, 2.4.2, 8.1.1

[29] A. Bairoch. The enzyme database in 2000. *Nucl. Acids Res.*, 28(1):304–305, 2000. 1.2.1

[30] Ahmet Bakan and Ivet Bahar. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proceedings of the National Academy of Sciences*, 106(34):14349–14354, 2009. 1.2.2, 2.4.2, 8.1.1, 8.2.1

[31] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten. Principal component analysis and long time protein dynamics. *The Journal of Physical Chemistry*, 100(7):2567–2572, 02 1996/02/15/. 2.1, 3.1

[32] D. Bang, G.I. Makhatadze, V. Tereshko, A. A. Kossiakoff, and S. B. Kent. Total chemical synthesis and x-ray crystal structure of a protein diastereomer: [d-gln 35]ubiquitin. *Angewandte Chemie International Edition*, 44(25):3852–3856, 2005. (document), 2.2.1, 8.3

[33] Duhee Bang, Alexey V Gribenko, Valentina Tereshko, Anthony A Kossiakoff, Stephen B Kent, and George I Makhatadze. Dissecting the energetics of protein [alpha]-helix c-cap termination through chemical protein synthesis. *Nat Chem Biol*, 2(3):139–143, 03 2006/03//print. 2.2.1

[34] N. P. Barton, C. S. Verma, and L. S. D. Caves. Inherent flexibility of calmodulin domains: A normal-mode analysis study. *The Journal of Physical Chemistry B*, 106(42):11036–11040, 09 2002/09/27/. 2.1

[35] H. Beach, R. Cole, M. L. Gill, and J. P. Loria. Conservation of mu s-ms enzume motions in the apo- and substrate-mimicked state. *J. Amer. Chem. Soc.*, 127:9167–9176, 2005. 6.2.1, 6.4.3

[36] C.D. Behrsin, M.L. Bailey, K.S. Bateman, K.S. Hamilton, L.M. Wahl, C.J. Brandl, B.H. Shilton, and D.W. Litchfield. Functionally important residues in the peptidyl-prolyl isomerase pin1 revealed by unigenic evolution. *Journal of Molecular Biology*, 365(4):1143 – 1162, 2007. 6.4.1

[37] Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995. 3.3.1

[38] M.-C. Bellisent-Funel, J.-M. Zanotti, and S. H. Chen. Slow dynamics of water molecules on the surface of globular proteins. *Faraday Discuss.*, 103:281–294, 1996. 1.1

[39] Marie-Claire Bellissent-Funel, Roy Daniel, Dominique Durand, Michel Ferrand, John L. Finney, Stephanie Pouget, Valerie Reat, and Jeremy C. Smith. Nanosecond protein dynamics: First detection of a neutron incoherent spinàíecho signal. *Journal of the American Chemical Society*, 120(29):7347–7348, 07 1998/07/11/. 1.1

[40] S. J. Benkovic and S. Hammes-Schiffer. A perspective on enzyme catalysis. *Science*, 267(90-93), 2003. 1.2.2, 2.1, 3.1, 5.4, 6.1, 7.1, 7.3, 7.3

[41] Stephen J. Benkovic, Gordon G. Hammes, and Sharon Hammes-Schiffer. Free-energy landscape of enzyme catalysis. *Biochemistry*, 47(11):3317–3321, 2008. 1.2.2, 6.1

[42] Brad Bennett, Paul Langan, Leighton Coates, Marat Mustyakimov, Benno Schoenborn, Elizabeth E. Howell, and Chris Dealwis. Neutron diffraction studies of Escherichia coli dihydrofolate reductase complexed with methotrexate. *Proceedings of the National Academy of Sciences*, 103(49):18493–18498, 2006. 1.1

[43] H. J. C Berendsen and S. Hayward. Collective protein dynamics in relation to function. *Current Opinion in Structural Biology*, 10(2):165–169, 2000. 4.1, 4.3.4

[44] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucl. Acids Res.*, 28(1):235–242, 2000. 2.2.1

[45] Robert B. Best, Kresten Lindorff-Larsen, Mark A. DePristo, and Michele Vendruscolo. Relation between native ensembles and experimental structures of proteins. *Proceedings of the National Academy of Sciences*, 103(29):10901–10906, 2006. 2.3.1

[46] A. R. Bizzarri and S. Cannistraro. Flickering noise in the potential energy fluctuations of proteins as investigated by md simulations. *Phys. Lett. A*, 236:596–601, 1997. 4.1

[47] D. D. Boehr, H. J. Dyson, and P. E. Wright. The dynamic energy landscape of dihydrofolate reductase. *Science*, 313:1638–1642, 2006. 1.1.2, 1.2.2, 1.2.2, 3.1, 5, 5.4, 6.1, 6.3.2, 7.1, 8.2.1

[48] David D. Boehr, H. Jane Dyson, and Peter E. Wright. An nmr perspective on enzyme dynamics. *Chemical Reviews*, 106(8):3055–3079, 07 2006. 6.1

[49] Daryl A. Bosco, Elan Z. Eisenmesser, Susan Pochapsky, Wesley I. Sundquist, and Dorothee Kern. Catalysis of cis/trans isomerization in native hiv-1 capsid by human cyclophilin a. *Proc. Natl. Acad. Sci. USA*, 99(8):5247–5252, 2002. 1.1, 1.1.2, 1.2.2, 5, 8.2.1

[50] K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. *SC Conference*, 0:43, 2006. 1.1.2, 4.1

[51] Bernard Brooks and Martin Karplus. Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Sciences*, 80(21):6571–6575, 1983. 1.1.1, 2.1, 4.1

[52] William R. Cannon and Stephen J. Benkovic. Solvation, reorganization energy, and biological catalysis. *Journal of Biological Chemistry*, 273(41):26257–26260, 1998. 1.1, 2.1, 6.1

[53] Stavros Caratzoulas, Joshua S. Mincer, and Steven D. Schwartz. Identification of a protein-promoting vibration in the reaction catalyzed by horse liver alcohol dehydrogenase. *Journal of the American Chemical Society*, 124(13):3270–3276, 03 2002/03/08/. 2.1

[54] Jean-Francois Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999. 3.3.1, 3.3.1

[55] V. Carnevale, S. Raugei, C. Micheletti, and P. Carloni. Convergent dynamics in the protease enzymatic superfamily. *J. Amer. Chem. Soc.*, 128:9766–9772, 2006. 1.2.2, 7.1

[56] D. A. Case, T. A. Darden, T. E. III Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, B. Wang, D. A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. W. Caldwell, W. S. Ross, and P. A. Kollman. Amber 7. 2003. 1.1.2, 2.2.2

[57] L. S. Caves, J.D. Evanseck, and M. Karplus. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.*, 7(3):649–666, 1998. 1.1.3, 2.1, 2.2, 2.4.1

[58] Xiaolin Cheng, Ivaylo Ivanov, Hailong Wang, Steven M. Sine, and J. Andrew McCammon. Nanosecond-timescale conformational dynamics of the human [alpha]7

nicotinic acetylcholine receptor. *Biophysical Journal*, 93(8):2622 – 2634, 2007. 2.4.1

[59] C. Chennubhotla and I. Bahar. Markov methods for hierarchical coarse-graining of large protein dynamics. In *LNBI*, pages 379–393, 2006. 1.1.1

[60] C. Chennubhotla and I. Bahar. Markov propagation of allosteric effects in biomolecular systems: application to groel–groes. *Mol. Sys. Biol.*, 2(36), 2006. 1.1.1

[61] C. Chennubhotla and I. Bahar. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput Biol*, 3(9):e172, 2007. 1.1.1, 8.1.3, 8.2.3

[62] C. Chennubhotla, A. J. Rader, L.-W. Yang, and I. Bahar. Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. *Physical Biology*, 2(4):S173–S180, 2005. 1.1.1

[63] J. D Chodera, N. Singhal, V. S. Pander, K. A. Dill, and W. C Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126:155101, 2007. 4.3.4

[64] J B Clarage, T Romo, B K Andrews, B M Pettitt, and G N Phillips. A sampling problem in molecular dynamics simulations of macromolecules. *Proceedings of the National Academy of Sciences of the United States of America*, 92(8):3288–3292, 1995. 1.1.2, 2.1, 4.1

[65] R. Cole and J. P. Loria. Evidence of flexibility in function of ribonuclease a. *Biochemistry*, 41:6072–6081, 2002. 1.1.2, 6.3.3, 6.4.3, 6.5

[66] William J. Cook, Leigh C. Jeffrey, Eileen Kasperek, and Cecile M. Pickart. Structure of tetraubiquitin shows how multiubiquitin chains can be formed. *Journal of Molecular Biology*, 236(2):601 – 609, 1994. 2.2.1

[67] T. M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2006. 3.3.2

[68] Valerie Daggett. Long timescale simulations. *Current Opinion in Structural Biology*, 10(2):160 – 164, 2000. 2.1

[69] Valerie Daggett and Alan Fersht. The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol*, 4(6):497–502, 2003. 10.1038/nrm1126. 4.4.2

[70] Isabella Daidone, Andrea Amadei, Danilo Roccatano, and Alfredo Di Nola. Molecular dynamics simulation of protein folding by essential dynamics sampling: Folding landscape of horse heart cytochrome c. *Biophysical Journal*, 85(5):2865 – 2871, 2003. 1.1.3

[71] R. M. Daniel, R. V. Dunn, J. L. Finney, and J. C. Smith. The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.*, 32:69–92, 2003. 1.1, 2.1

[72] R. M. Daniel, J. L. Finney, V. Reat, R. Dunn, M. Ferrano, and J. C. Smith. Enzyme dynamics and activity: Time-scale dependence of dynamical transitions in glutamate dehydrogenase solution. *Biophysical Journal*, 77(4):2184 – 2190, 1999. 2.1

[73] Warren L. DeLano. The pymol molecular graphics system, 2003. 2.2.2, 4.4.1, 4.4.2, 6.2.1

[74] Deena L. Di Stefano and A. Joshua Wand. Two-dimensional proton nmr study of human ubiquitin: a main chain directed assignment and structure analysis. *Biochemistry*, 26(23):7272–7281, 11 1987/11/01/. 2.2.1

[75] Ivan Dikic, Soichi Wakatsuki, and Kylie J. Walters. Ubiquitin binding domains - from structures to functions. *Nature Reviews Molecular Cell Biology*, 10:659–671, 2009. 2.2.1

[76] P. Durand, G. Trinquier, and Y.H. Sanejouand. A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, 34:759–771, 1994. 1.1.1

[77] E. Z. Eisenmesser, D. A. Bosco, M. Akke, and D. Kern. Enzyme dynamics during catalysis. *Science*, 295(5559):1520–1523, 2002. 1.1.2, 1.2.2, 2.1, 5, 5.1.1, 5.4, 6.1, 6.3.1

[78] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D.M. Korzhnev, Wolf-Watz. M., D.A. Bosco, J.J. Skalicky, L.E. Kay, and D. Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438:117–121, 2005. 1.2.2, 1.2.2, 2.1, 5, 5.1.1, 5.3.1, 5.3.1, 5.3.2, 5.4, 6.1, 6.3.1, 7.1, 8.2.1

[79] Yasser El-Hawari, Angelo D. Favia, Ewa S. Pilka, Michael Kisiela, Udo Oppermann, Hans-Jrg Martin, and Edmund Maser. Analysis of the substrate-binding site of human carbonyl reductases cbr1 and cbr3 by site-directed mutagenesis. *Chemico-Biological Interactions*, 178(1-3):234 – 241, 2009. Enzymology and Molecular Biology of Carbonyl Metabolism. 8.2.2

[80] R. Elber and M. Karplus. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. *Science*, 235(4786):318–321, 1987. 3.1

[81] R. Elber and M. Karplus. Enhanced sampling in molecular dynamics: use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J. Am. Chem. Soc.*, 112(25):9161–9175, 1990. 4.1

[82] Erich Elsen, Mike Houston, V. Vishal, Eric Darve, Pat Hanrahan, and Vijay Pande. N-body simulation on gpus. In *SC '06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, page 188, New York, NY, USA, 2006. ACM. 1.1.2

[83] Daniel L. Ensign, Peter M. Kasson, and Vijay S. Pande. Heterogeneity even at the speed limit of folding: Large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *Journal of Molecular Biology*, 374(3):806 – 816, 2007. 4.1

[84] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and F. G Parak. Slaving: Solvent fluctuations dominate protein dynamics and functions. *Proc. Natl. Acad. Sci. USA*, 99:16407–16051, 2002. 1, 1.1, 1.1, 6.5, 8.2.3

[85] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and R. D Young. Bulk-solvent and hydration-shell fluctuations, similar to - and -fluctuations in glasses, control protein motions and functions. *Proc. Natl. Acad. Sci. USA*, 101:14408–14413, 2004. 1, 1.1, 1.1, 6.1, 7.1, 8.2.3

[86] A. R. Fersht. Protein folding and stability: the pathway of folding of barnase. *FEBS Letters*, 325(1-2):5–16, 1993. 4.4.2, 4.4.2

[87] A. R. Fersht and Valerie Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108(4):573–582, 2002. 1.1.2

[88] Alan R. Fersht. On the simulation of protein folding by short time scale molecular dynamics and distributed computing. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14122–14125, 2002. 2.2

[89] E. Fischer. Enzyme catalysis. *Ber. Dtsch. Chem. Ges.*, 27:3189, 1894. 1

[90] J.S. Fraser, M.W. Clarkson, S.C. Degnan, R. Erion, D. Kern, and T. Alber. Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, 462(7273):669–673, 2009. 1.1, 1.1.2, 1.2.2, 1.2.2, 3.1

[91] H. Frauenfelder, P. W. Fenimore, G. Chen, and B. H. McMahon. Protein folding is enslaved by solvent motions. *Proc. Natl. Acad. Sci. USA*, 103:15469–15472, 2006. 1, 1.1, 8.2.3

[92] H. Frauenfelder, F. Parak, and R. D. Young. Conformational substates in proteins. *Ann. Rev. Biophys. Biophys. Chem.*, 17:451–479, 1988. 1.1.2, 3.1

[93] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou. Temperature-dependent x-ray diffraction as a probe of protein structural dynamics. *Nature*, 280(5723):558–563, 1979. 3.1

[94] H Frauenfelder, SG Sligar, and PG Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991. 1.1.2, 8.1.1

[95] Peter L. Freddolino, Feng Liu, Martin H. Gruebele, and Klaus Schulten. Ten-microsecond MD simulation of a fast-folding WW domain. *Biophys. J.*, page biophysj.108.131565, 2008. 4.1

[96] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Computational Science Series. Academic Press, San Diego, 2002. 1.1.2, 1.1.2, 5.1.2

[97] H Fujisaki and JE Straub. Vibrational energy relaxation in proteins. *Proc. Nat. Acad. Sci. USA*, 102(19):6726–6731, 2005. 3.1

[98] Nicholas Furnham, Tom L Blundell, Mark A DePristo, and Thomas C Terwilliger. Is one solution good enough? *Nat Struct Mol Biol*, 13(3):184–185, 03 2006. 6.1

[99] Angel E. García. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, 68(17):2696–2699, Apr 1992. 2.1

[100] M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar. How enzymes work: Analysis by modern rate theory and computer simulations. *Science*, 303:186–195, 2004. 2.1

[101] J. A. Gerlt and P. C. Babbitt. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct superfamilies. *Annu. Rev. Biochem.*, 70:209–246, 2001. 1.2.1, 7.1, 8.2.1

[102] Margaret E Glasner, John A Gerlt, and Patricia C Babbitt. Evolution of enzyme superfamilies. *Current Opinion in Chemical Biology*, 10(5):492 – 497, 2006. Analytical techniques / Mechanisms. 1.2.1, 8.2.1

[103] N Go, T Noguti, and T Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences of the United States of America*, 80(12):3696–3700, 1983. 1.1.1

[104] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996. 3.3.1

[105] Nina M Goodey and Stephen J Benkovic. Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol*, 4(8):474–482, 08 2008. 8.2.4

[106] David G. Gourley, Alexander W. Schuttelkopf, Gordon A. Leonard, James Luba, Larry W. Hardy, Stephen M. Beverley, and William N. Hunter. Pteridine reductase mechanism correlates pterin metabolism with drug resistance in trypanosomatid parasites. *Nat. Struct. Mol. Biol*, 8(6):521–525, 2001. 7.2.1

[107] V. Gupta, S. Muyldermans, L. Wyns, and D. M. Salunke. The crystal structure of recombinant rat pancreatic rnase a. *Proteins: Struct. Funct. Bioinform.*, 35(1):1–12, 1999. 6.2.1

[108] R. Gussio, N. Pattabiraman, G. E. Kellogg, and D. W. Zaharevitz. Use of 3d qsar methodology for data mining the national cancer institute repository of small molecules: Application to hiv-1 reverse transcriptase inhibition. *Methods*, 14:255–263, 1998. 4.3.4

[109] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: Evolutionary units of three-dimensional structure. *Cell*, 138(4):774 – 786, 2009. 8.2.2

[110] Thomas A. Halgren and Wolfgang Damm. Polarizable force fields. *Current Opinion in Structural Biology*, 11(2):236 – 242, 2001. 1.1.2

[111] T. Haliloglu, I. Bahar, and B. Erman. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, 79:3090–3093, 1997. 1.1.1

[112] G. G. Hammes. Mechanism of enzyme catalysis. *Nature*, 204:342–343, 1964. 7.1

[113] G. G. Hammes. Multiple conformational changes in enzyme catalysis. *Biochemistry*, 41(8221-8228), 2002. 2.1, 6.1, 7.1, 7.3

[114] Sharon Hammes-Schiffer. Impact of enzyme motion on activity. *Biochemistry*, 41(45):13335–13343, 10 2002/10/11/. 1.2.2, 2.1

[115] Sharon Hammes-Schiffer and Stephen J. Benkovic. Relating protein motion to catalysis. *Ann. Rev. Biochem.*, 75:519–541, 2006. 1, 6.1

[116] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *App. Stat.*, 28(1):100–108, 1979. 4.3.3

[117] Timothy Havel, Irwin Kuntz, and Gordon Crippen. The theory and practice of distance geometry. *Bulletin of Mathematical Biology*, 45(5):665–720, 09 1983. 4.2

[118] S. Hayward and N. Go. Collective variable description of native protein dynamics. *Annual Review of Physical Chemistry*, 46(1):223–250, 1995. 1.1.3, 3.1, 4.1

[119] S. Hayward, A. Kitao, and N. Go. Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. *Proteins: Struct., Funct., & Genetics*, 23:177–186, 1995. 1.1.2, 3.1, 4.1

[120] X. He, D. Cai, and P. Niyogi. Tensor subspsace analysis. In *Nineth Annual Conference on Neural Information Processing Systems 2005*, 2005. 4.3.2

[121] William Heller. Influence of multiple well defined conformations on small-angle scattering of proteins in solution. *Acta Crystallographica Section D*, 61(1):33–44, 2005. 1.1

[122] Ulf Hensen, Helmut Grubmüller, and Oliver F. Lange. Adaptive anisotropic kernels for nonparametric estimation of absolute configurational entropies in high-dimensional configuration spaces. *Phys. Rev. E*, 80(1):011913, Jul 2009. 8.1.3

[123] Ulf Hensen, Oliver F. Lange, and Helmut Grubmüller. Estimating absolute configurational entropies of macromolecules: The minimally coupled subspace approach. *PLoS ONE*, 5(2):e9179, 02 2010. 8.1.3

[124] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007. 1, 1.1, 5.1

[125] Katherine A. Henzler-Wildman, Ming Lei, Vu Thai, S. Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, 12 2007/12/06/print. 1.1, 1.1.2, 2.1, 2.4.2, 5, 8.2.1

[126] Berk Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E*, 62(6):8438–8448, Dec 2000. 2.3.2

[127] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 02 2008. 1.1.2

[128] Konrad Hinsen. *Normal Mode Analysis: Theory and Applications to biological and chemical systems*, pages 1–16, 2003. 3.1

[129] Satoshi Hirano, Masato Kawasaki, Hideaki Ura, Ryuichi Kato, Camilla Raiborg, Harald Stenmark, and Soichi Wakatsuki. Double-sided ubiquitin binding of hrs-uim in endosomal protein sorting. *Nat Struct Mol Biol*, 13(3):272–277, 03 2006/03//print. 2.2.1

[130] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233(1):123–138, 1993. 7.2.1

[131] L. Holm and C. (1996) Sander. Mapping the protein universe. *Science*, 273:595–603, 1996. 7.2.1

[132] Liisa Holm and Jong Park. DaliLite workbench for protein structure comparison. *Bioinformatics*, 16(6):566–567, 2000. (document), 6.1, 7.2.1

[133] Chun-Hua Hsu, You-Di Liao, Yun-Ru Pan, Lih-Woan Chen, Shih-Hsiung Wu, Ying-Jen Leu, and Chinpan Chen. Solution structure of the cytotoxic rnase 4 from oocytes of bullfrog rana catesbeiana. *Journal of Molecular Biology*, 326(4):1189 – 1201, 2003. 6.2.1

[134] Ryan G. Huff, Ersin Bayram, Huan Tan, Stacy T. Knutson, Michael H. Knaggs, Allen B. Richon, Peter Santago, and J. Fetrow. Chemical and structural diversity in cyclooxygenase protein active sites. *Chemistry and Biodiversity*, 2(11):1533–1552, 2005. 8.2.1

[135] T. Ichiye and M. Karplus. Anisotropy and anharmonicity of atomic fluctuations in proteins: analysis of a molecular dynamics simulation. *Proteins*, 2(3):236–259, 1987. 3.1

[136] T. Ichiye and M. Karplus. Anisotropy and anharmonicity of atomic fluctuations in proteins: implications for x-ray analysis. *Biochemistry*, 27(9):3487–3497, 1988. 3.1, 3.3.2

[137] Basak Isin, Klaus Schulten, Emad Tajkhorshid, and Ivet Bahar. Mechanism of signal propagation upon retinal isomerization: Insights from molecular dynamics simulations of rhodopsin restrained by normal modes. *Biophysical Journal*, 95(2):789 – 803, 2008. 1.1.2

[138] Lakshminarayan Iyer, A Maxwell Burroughs, and L Aravind. The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. *Genome Biology*, 7(7):R60, 2006. 2.2.1

[139] Sergei Izvekov and Gregory A. Voth. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B*, 109(7):2469–2473, 2005. 1.1.2

[140] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins: Struct., Funct., Genet.*, 44(2):150–65, 2001. (document), 4.3.4, 4.6

[141] Lin Jiang, Eric A. Althoff, Fernando R. Clemente, Lindsey Doyle, Daniela Rothlisberger, Alexandre Zanghellini, Jasmine L. Gallaher, Jamie L. Betker, Fujie Tanaka, III Barbas, Carlos F., Donald Hilvert, Kendall N. Houk, Barry L. Stoddard, and David Baker. De Novo Computational Design of Retro-Aldol Enzymes. *Science*, 319(5868):1387–1391, 2008. 8.2.2

[142] Ivan T. Jolliffe. *Principal Component Analysis*. Springer, 2002. 1.1.3, 2.4.2

[143] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, 32:922–923, 1976. 1.1.3

[144] G. Kamath, Elizabeth E. Howell, and P. K. Agarwal. The tail wagging the dog: Insights into catalysis in r67 dihydrofolate reductase. *Proc. Nat. Acad. Sci. U. S. A.*, under revision, 2010. 8.2.2

[145] George A. Kaminski, Richard A. Friesner, Julian Tirado-Rives, and William L. Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B*, 105(28):6474–6487, 04 2001. 1.1.2

[146] H. Kamisetty, B. Ghosh, C. Bailey-Kellogg, and C.J. Langmead. Modeling and Inference of Sequence-Structure Specificity. In *Proc. of the 8th International Conference on Computational Systems Bioinformatics (CSB)*, pages 91–101, 2009. 8.1.3

[147] H. Kamisetty, A. Ramanathan, C. Bailey-Kellogg, and C. J. Langmead. Accounting for conformational entropy in predicting binding free energies of protein-protein interactions. *PLoS Comput Biol*, (under review), 2010. 8.1.3

[148] H. Kamisetty, E.P. Xing, and C.J. Langmead. Free Energy Estimates of All-atom Protein Structures Using Generalized Belief Propagation. *J. Comp. Bio.*, 15(7):755–766, 2008. 8.1.3

[149] M. Karplus, T. Ichiye, and B. M. Pettitt. Configurational entropy of native proteins. *Biophys. J.*, 52(6):1083–1085, 1987. 1.1.3

[150] M. Karplus and J. N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981. 1.1.3, 2.1, 2.4.2, 4.1, 5.2.4, 6.2.1, 7.2.2

[151] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9:646–652, 2002. 1.1.2

[152] Dorothee Kern and Erik RP Zuiderweg. The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, 13(6):748 – 757, 2003. 1

[153] O. Keskin, R. L. Jernigan, and I. Bahar. Proteins with similar architecture exhibit similar large-scale dynamic behavior. specificity resides in local differences. *Biophys. J.*, 78:2093–2106, 2000. 1.2.2, 7.1

[154] Ryo Kitahara, Shigeyuki Yokoyama, and Kazuyuki Akasaka. Nmr snapshots of a fluctuating protein structure: Ubiquitin at 30bar-3kbar. *Journal of Molecular Biology*, 347(2):277 – 285, 2005. 2.3.2

[155] Akio Kitao and Nobuhiro Go. Investigating protein dynamics in collective coordinate space. *Curr. Opi. Struct. Biol.*, 9(2):164 – 169, 1999. 1.1.3

[156] Judith P. Klinman. An integrated model for enzyme catalysis emerges from studies of hydrogen tunneling. *Chemical Physics Letters*, 471(4-6):179 – 193, 2009. 5.4, 6.1, 7.1

[157] A.E. Klon, A. Heroux, L.J. Ross, V. Pathak, C.A. Johnson, J.R. Piper, and D.W. Borhani. Atomic structures of human dihydrofolate reductase complexed with nadph and two lipophilic antifolates at 1.09 a and 1.05 a. *J. Mol. Biol.*, 320:677–693, 2002. 6.2.1

[158] S.C.L. Kmerlin and A. Warshel. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Prot.: Struct. Funct. Bioinform.*, 78(6):1339–1375, 2009. 1.2.2

[159] Peter A. Kollman, Irina Massova, Carolina Reyes, Bernd Kuhn, Shuanghong Huo, Lillian Chong, Matthew Lee, Taisung Lee, Yong Duan, Wei Wang, Oreola Donini, Piotr Cieplak, Jaysharee Srinivasan, David A. Case, and Thomas E. Cheatham. Calculating structures and free energies of complex molecules: Combining molecular

mechanics and continuum models. *Accounts of Chemical Research*, 33(12):889–897, 10 2000. 1.1.2

[160] Dmitry A. Kondrashov, Adam W. Van Wynsberghe, Ryan M. Bannen, Qiang Cui, and GeorgeN. Phillips Jr. Protein structural variation in computational models and crystallographic data. *Structure*, 15(2):169 – 177, 2007. 2.4.2

[161] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Nat. Acad. Sci. U. S. A.*, 44:98–104, 1958. 1

[162] E. L. Kovrigin and J. P. Loria. Characterization of the transition state of functional enzyme dynamics. *J. Amer. Chem. Soc.*, 128:7724–7725, 2006. 6.3.3

[163] E. L. Kovrigin and J. P. Loria. Enzyme dynamics along the reaction coordinate: critical role of a conserved residue. *Biochemistry*, 45:2636–2647, 2006. 1.1.2, 1.2.2, 6.3.3, 6.5

[164] J Kraut. How do enzymes work? *Science*, 242(4878):533–540, 1988. 7.1

[165] P. Kroonenberg and J. D. Leeuw. Principal component analysis of three-mode data by means of alternating least-squares algorithm. *Psychometrika*, 45(1):69–97, 1980. 4.3.2

[166] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, and P.A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comp. Chem.*, 13(8):1011–1021, 1992. 5.1.2

[167] Wladimir Labeikovsky, Elan Z. Eisenmesser, Daryl A. Bosco, and Dorothee Kern. Structure and dynamics of pin1 during catalysis by nmr. *Journal of Molecular Biology*, 367(5):1370 – 1381, 2007. 6.4.1

[168] Oliver Lange and Helmut Grubmuller. Full correlation analysis of conformational protein dynamics. *Prot.: Struct. Funct. Bioinform.*, 70(4):1294–1312, 2007. 3.3.2, 4.3.4, 8.1.3, 8.1.3

[169] Oliver F. Lange and Helmut Grubmuller. Can principal components yield a dimension reduced description of protein dynamics on long time scales? *The Journal of Physical Chemistry B*, 110(45):22842–22852, 10 2006/10/12/. 2.1, 3.1

[170] Oliver F. Lange, Nils-Alexander Lakomek, Christophe Fares, Gunnar F. Schroder, Korvin F. A. Walter, Stefan Becker, Jens Meiler, Helmut Grubmuller, Christian Griesinger, and Bert L. de Groot. Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution. *Science*,

320(5882):1471–1475, 2008. 1.1.3, 2.1, 2.2.1, 2.2.1, 2.3.1, 3.2.1, 3.2.1, 3.4.2, 3.4.3, 8.1.1

[171] Jonathan K. Lassila, David Baker, and Daniel Herschlag. Origins of catalysis by computationally designed retroaldolase enzymes. *Proceedings of the National Academy of Sciences*, 107(11):4937–4942, 2010. 8.2.2

[172] A. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall PTR, 2 edition, 2001. 1.1.2, 1.1.2

[173] Eric H. Lee, Jen Hsin, Marcos Sotomayor, Gemma Comellas, and Klaus Schulten. Discovery through the computational microscope. *Structure*, 17(10):1295 – 1306, 2009. 1.1.2

[174] Jeeyeon Lee, Madhusudan Natarajan, Vishal C. Nashine, Michael Socolich, Tina Vo, William P. Russ, Stephen J. Benkovic, and Rama Ranganathan. Surface Sites for Engineering Allosteric Control in Proteins. *Science*, 322(5900):438–442, 2008. 8.2.4

[175] Hongxing Lei and Yong Duan. Improved sampling methods for molecular simulation. *Current Opinion in Structural Biology*, 17(2):187 – 191, 2007. Theory and simulation / Macromolecular assemblages. 1.1.2

[176] D. Leitner. Energy flow in proteins. *Ann. Rev. Phys. Chem.*, 59, 2008. 3.1

[177] D.M. Leitner and J. E. Straub, editors. *Proteins Energy, Heat and Signal Flow*. CRC Press (Taylor and Francis Group), 2010. 8.1.3, 8.2.3

[178] T. Lenaerts, J. Ferkinghoff-Borg, F. Stricher, L. Serrano, J. W. H. Schymkowitz, and F. Rousseau. Quantifying information transfer by protein domains: Analysis of the fyn sh2 domain structure. *BMC Struct. Biol.*, 8:43, 2008. 4.3.4

[179] Michael Levitt, Christian Sander, and Peter S. Stern. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of Molecular Biology*, 181(3):423 – 447, 1985. 1.1.1, 2.1

[180] R. M. Levy, A. R. Srinivasan, W. K. Olson, and J. A. McCammon. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers*, 23:1099–1112, 1984. 3.1

[181] T.R. Lezon, A. Sali, and I. Bahar. Global motions of the nuclear pore complex: Insights from elastic network models. *PLoS Comput. Biol.*, 9:e1000496, 2009. 1.1.1

[182] Da-Wei Li, Dan Meng, and Rafael Brüschweiler. Short-range coherence of internal protein dynamics revealed by high-precision in silico study. *Journal of the American Chemical Society*, 131(41):14610–14611, 09 2009. 2.2.1

[183] R. Li, R. Sirawaraporn, P. Chitnumsub, W. Sirawaraporn, J. Wooden, F. Athappilly, S. Turley, and W.G. Hol. Three-dimensional structure of m. tuberculosis dihydrofolate reductase reveals opportunities for the design of novel tuberculosis drugs. *J. Mol. Biol.*, 295:307–323, 2000. 6.2.1

[184] Steve W. Lockless and Rama Ranganathan. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science*, 286(5438):295–299, 1999. 6.5, 8.2.2

[185] Jianpeng Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13(3):373 – 380, 2005. 1.1.2, 1.2.2, 2.1, 2.4.2, 8.1.1

[186] Gia G. Maisuradze and D. Leitner. Free energy landscape of a biomolecule in dihedral principal component space: Sampling convergence and correspondence between structures and minima. *Proteins: Structure, Function, and Bioinformatics*, 67(3):569–578, 2007. 1.1.3

[187] D. Malmodin and M. Billeter. Multiway decomposition of nmr spectra with coupled evolution periods. *J. Am. Chem. Soc.*, 127(39):13486–13487, 2005. 4.3.4

[188] T. Mamonova, B. Hespenheide, R. Straub, M. F. Thorpe, and M. Kurnikova. Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.*, 2(4):S137–47, 2005. 4.3.4, 4.4.2, 4.4.2

[189] Paul Maragakis, Kresten Lindorff-Larsen, Michael P. Eastwood, Ron O. Dror, John L. Klepeis, Isaiah T. Arkin, Morten O. Jensen, Huafeng Xu, Nikola Trbovic, Richard A. Friesner, Arthur G. Palmer, and David E. Shaw. Microsecond molecular dynamics simulation shows effect of slow loop dynamics on backbone amide order parameters of proteins. *The Journal of Physical Chemistry B*, 112(19):6155–6158, 03 2008. 2.2.1

[190] Neelan J. Marianayagam and Sophie E. Jackson. The folding pathway of ubiquitin from all-atom molecular dynamics simulations. *Biophysical Chemistry*, 111(2):159 – 171, 2004. 2.2.1

[191] Phineus R. L. Markwick, Guillaume Bouvignies, Loic Salmon, J. Andrew McCammon, Michael Nilges, and Martin Blackledge. Toward a unified representation of protein structural dynamics in solution. *Journal of the American Chemical Society*, 131(46):16968–16975, 11 2009. 2.2.1

[192] Christopher Kroboth Materese, Christa Charisse Goldmon, and Garegin A. Papoian. Hierarchical organization of eglin c native state dynamics is shaped by competing direct and water-mediated interactions. *Proc. Nat. Acad. Sci. USA*, 105(31):10659–10664, 2008. 2.1, 3.1

[193] J. A. McCammon, J.B. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977. 1.1

[194] Sean R. McGuffee and Adrian H. Elcock. Diffusion, crowding amp; protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput Biol*, 6(3):e1000694, 03 2010. 1.1.2

[195] G.J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, N.Y., 1988. 3.4.2

[196] Elaine C Meng, B.J. Polacco, and Patricia C. Babbitt. Superfamily active site templates. *Prot.: Struct. Funct. Bioinform.*, 55(4):962–976, 2004. 1.2.1, 8.2.1

[197] O. Miyashita, J. N. Onuchic, and P. G. Wolynes. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22):12570–12575, 2003. 2.1

[198] K. Moritsugu, O. Miyashita, and A. Kidera. Vibrational energy transfer in a protein molecule. *Phys. Rev. Lett.*, 85(18):3970–3973, 2000. 8.1.3

[199] Yuguang Mu, Phuong H. Nguyen, and Gerhard Stock. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Structure, Function, and Bioinformatics*, 58(1):45–52, 2005. 1.1.3

[200] Zachary D Nagel and Judith P Klinman. A 21st century revisionist's view at a turning point in enzymology. *Nat Chem Biol*, 5(8):543–550, 08 2009. 8.2.3

[201] V. M. Naik and S. Krimm. Vibrational analysis of peptides, polypeptides and proteins. *Int. J. Pept. Protein Res.*, 23(1):1–24, 1984. 1.1.1

[202] Geeta J. Narlikar and Daniel Herschlag. Mechanistic aspects of enzymatic catalysis:lessons from comparison of rna and protein enzymes0. *Annual Review of Biochemistry*, 66(1):19–59, 1997. 1.2.1

[203] D. L. Nelson and M. M. Cox. *Principles of Biochemistry*. W. H. Freeman and Company, New York, fourth edition edition, 2005. 1, 1.1.2

[204] Dzung T. Nguyen and David A. Case. On finding stationary states on large-molecule potential energy surfaces. *The Journal of Physical Chemistry*, 89(19):4020–4026, 09 1985/09/01/. 1.1.1, 2.2.2

[205] D. W. Noid, K. Fukui, B. G. Sumpter, C. Yang, and R. E. Tuzun. Time-averaged normal coordinate analysis of polymer particles and crystals. *Chemical Physics Letters*, 316(3-4):285 – 296, 2000. 2.1

[206] S. B. Nolde, A. S. Arseniev, V. Yu, and M. Billeter. Essential domain motions in barnase revealed by md simulations. *Proteins: Struct., Funct. and Bioinformatics*, 46(3):250–258, 2003. 4.4.2, 4.4.2

[207] S. H. Northrup, M. R. Pear, C. Y. Lee, J. A. McCammon, and M. Karplus. Dynamical theory of activated processes in globular proteins. *Proc. Natl. Acad. Sci. USA*, 79:4035–4039, 1982. 3.1

[208] S. H. Northrup, M. R. Pearl, J. D. Morgan, J. A. McCammon, and M. Karplus. Molecular dynamics of ferrocytochrome c: magnitude and anisotropy of atomic displacements. *J. Mol. Biol.*, 153:1087–1111, 1981. 3.1

[209] Sunil Ojha, Elaine C Meng, and Patricia C Babbitt. Evolution of function in the "two dinucleotide binding domains" flavoproteins. *PLoS Comput Biol*, 3(7):e121, 07 2007. 1.2.1, 7.1, 8.2.1

[210] M. H. M. Olsson, W. W. Parson, and A. Warshel. Dynamical contributions to enzyme catalysis: Critical tests of a popular hypothesis. *Chem. Rev.*, 106:1737–1756, 2006. 1.2.2, 5.4, 6.1

[211] Marcel Ottiger, Oliver Zerbe, Peter Gntert, and Kurt Wthrich. The nmr solution conformation of unligated human cyclophilin a,. *Journal of Molecular Biology*, 272(1):64 – 81, 1997. (document), 5.2.1, 5.2.4, 5.3.1, 5.4

[212] A. Di Pace, A. Cupane, M. Leone, E. Vitrano, and L. Cordone. Protein dynamics. vibrational coupling, spectral broadening mechanisms, and anharmonicity effects in

carbonmonoxy heme proteins studied by the temperature dependence of the soret band lineshape. *Biophysical Journal*, 63(2):475 – 484, 1992. 2.1

[213] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C.D. Snow, E. J. Sorin, and B. Zagrovic. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68(1):91–109, 2003. 2.2, 4.1

[214] A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw Hill, 2002. 3.1

[215] F. Parak and H. Formanek. Untersuchung des Schwingungsanteils und des Kristallgitterfehleranteils des Temperaturfaktors in Myoglobin durch Vergleich von Mössbauer-absorptionsmessungen mit Röntgenstrukturdaten. *Acta Crystallographica Section A*, 27(6):573–578, Nov 1971. 3.1

[216] F Parak and E W Knapp. A consistent picture of protein dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 81(22):7088–7092, 1984. 3.1

[217] L.C Pauling. Molecular architecture and biological reactions. *Chem. Engg. News*, 24:1375–1377, 1946. 6.1

[218] David A. Pearlman, David A. Case, James W. Caldwell, Wilson S. Ross, Thomas E. Cheatham, Steve DeBolt, David Ferguson, George Seibel, and Peter Kollman. 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, 1995. 1.1.2, 2.2.1, 2.2.2

[219] Pedro Jose Barbosa Pereira, Sandra Macedo-Ribeiro, Antonio Parraga, Rosa Perez-Luque, Orla Cunningham, Kevin Darcy, Timothy J. Mantle, and Miquel Coll. Structure of human biliverdin ix[β] reductase, an early fetal bilirubin ix[beta] producing enzyme. *Nat. Struct. Mol. Biol*, 8(3):215–220, 2001. 7.2.1

[220] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with namd. *J. Comp. Chem.*, 26:1781–1802, 2005. 1.1.2, 3.4.2, 4.1

[221] Cecile M. Pickart. Mechanisms underlying ubiquitination. *Annual Review of Biochemistry*, 70(1):503–533, 2001. 2.2.1

[222] Ewa S. Pilka, Frank H. Niesen, Wen Hwa Lee, Yasser El-Hawari, James E. Dunford, Grazyna Kochan, Vladimir Wsol, Hans-Joerg Martin, Edmund Maser, and Udo Oppermann. Structural basis for substrate specificity in human monomeric carbonyl reductases. *PLoS ONE*, 4(10):e7113, 10 2009. 8.2.2

[223] Steve Plimpton, Roy Pollock, and Mark Stevens. Particle-mesh ewald and rrespa for parallel molecular dynamics simulations. In *In Proceedings of the Eighth SIAM Conference on Parallel Processing for Scientific Computing*, 1997. 2.2.2

[224] Francesco Raimondi, Modesto Orozco, and Francesca Fanelli. Deciphering the deformation modes associated with function retention and specialization in members of the ras superfamily. *Structure*, 18(3):402 – 414, 2010. 1.2.2, 7.1

[225] A. Ramanathan and P. K. Agarwal. Computational identification of slow conformational fluctuations in proteins. *J. Phys. Chem. B*, 2009 (in press). 3.1, 3.2.1, 3.3.2, 3.4.2, 4.4.1, 6.2.1

[226] A. Ramanathan, P. K. Agarwal, and C. J. Langmead. Using tensor analysis to characterize contact-map dynamics in proteins. Technical Report CMU-CS-08-109, Carnegie Mellon University, January 2008. 4.3.2

[227] P. A. Ramero and F. H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews*, 10:866–876, 2009. 8.2.2

[228] Francisca E. Reyes-Turcu, John R. Horton, James E. Mullally, Annie Heroux, Xiaodong Cheng, and Keith D. Wilkinson. The ubiquitin binding domain znf ubp recognizes the c-terminal diglycine motif of unanchored ubiquitin. *Cell*, 124(6):1197 – 1208, 2006. 2.2.1

[229] Thomas H. Rod, Jennifer L. Radkiewicz, and Charles L. Brooks. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12):6980–6985, 2003. 6.1

[230] M. G. Rossmann, M.A. Moras, and K. W. Olson. Chemical and biological evolution of a nucleotide binding protein. *Nature*, 250:194–199, 1974. 7.1

[231] Manuel Rueda, Pablo ChacÛn, and Modesto Orozco. Thorough validation of protein normal mode analysis: A comparative study with essential dynamics. *Structure*, 15(5):565 – 575, 2007. 2.4.2

[232] William P. Russ, Drew M. Lowery, Prashant Mishra, Michael B. Yaffe, and Rama Ranganathan. Natural-like function in artificial ww domains. *Nature*, 437(7058):579–583, 09 2005. 8.2.2

[233] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J. C. Berendsen. 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, 1977. 2.2.2

[234] M.R. Sawaya and J. Kraut. Loop and subdomain movements in the mechanism of escherichia coli dihydrofolate reductase: Crystallographic evidence. *Biochemistry*, 36:586–603, 1997. (document), 6.2.1, 6.7, 7.2.1, 7.3

[235] Tamar Schlick, R. D. Skeel, A. T. Brunger, L. V. Kale, J. Hermans, K. Schulten, and J. A. Board, Jr. Algorithmic challenges in computational molecular biophysics. *J. Comp. Phys.*, 151:9–48, 1999. 4.1

[236] J. Schnell, H. J. Dyson, and P. E. Wright. Effect of cofactor binding and loop conformation on side-chain dynamics in dihydrofolate reductase. *Biochemistry*, 43:374–383, 2004. 6.1

[237] Gunnar F. Schroder. *Simulation of Fluorescence Spectroscopy Experiments*. PhD thesis, Universitat Gottingen, 2004. 8.1.3

[238] Steven D Schwartz and Vern L Schramm. Enzymatic transition states and dynamic motion in barrier crossing. *Nat Chem Biol*, 5(8):551–558, 08 2009. 8.2.1, 8.2.3

[239] J. Shao, S.W. Tanner, N. Thompson, and T.E. Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation*, 3(6):2312–2334, 2007. 4.1

[240] Robert Shapiro, James F. Riordan, and Bert L. Vallee. Characteristic ribonucleolytic activity of human angiogenin. *Biochemistry*, 25(12):3527–3532, 06 1986. 6.1, 6.4.3

[241] Robert Shapiro and Bert L. Vallee. Site-directed mutagenesis of histidine-13 and histidine-114 of human angiogenin. alanine derivatives inhibit angiogenin-induced angiogenesis. *Biochemistry*, 28(18):7401–7408, 09 1989. 6.1, 6.4.3

[242] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo,

J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. In *ISCA '07: Proceedings of the 34th annual international symposium on Computer architecture*, pages 1–12, New York, NY, USA, 2007. ACM. 1.1.2, 4.1

[243] Michael R. Shirts and Vijay S. Pande. Mathematical analysis of coupled parallel simulations. *Phys. Rev. Lett.*, 86(22):4983–4987, May 2001. 2.2

[244] Scott A. Showalter and Rafael Brüschweiler. Validation of molecular dynamics simulations of biomolecules using nmr spin relaxation as benchmarks: Application to the amber99sb force field. *Journal of Chemical Theory and Computation*, 3(3):961–975, 03 2007. 2.2.1

[245] A. Smilde, R. Bro, and P. Geladi. *Multi-way Analysis: Applications in the Chemical Sciences*. J. Wiley and Sons, Ltd., 2004. 4.3.4

[246] Michael Socolich, Steve W. Lockless, William P. Russ, Heather Lee, Kevin H. Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, 09 2005. 8.2.2

[247] G. Song and R. L. Jernigan. An enhanced elastic network model to represent the motions of domain-swapped proteins. *Prot.: Struct. Funct. Bioinform.*, 63(1):197–209, 2006. 1.1.2

[248] D. Staykova, J. Fredriksson, W. Bermel, and M. Billeter. ssignment of protein nmr spectra based on projections, multi-way decomposition and a fast correlation approach. *Journal of Biomolecular NMR*, 2008. 4.3.4

[249] J. E. Stone, J.C. Phillips, Peter L. Freddolino, D. J. Hardy, L.G. Trabuco, and K. Schulten. Accelerating molecular modeling applications with graphics processors. *Computational Chemistry*, 28(16):2618–2640, 2007. 1.1.2

[250] G. M. Suel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10:59–69, 2003. 4.5, 6.5, 8.2.2

[251] J. Sun, S. Papadimitrou, and P. S. Yu. Window-based tensor analysis on high-dimensional and multi-aspect streams, 2006. 4.3.2

[252] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 374–383, 2006. 4.3.2, 4.4

[253] Wesley I. Sundquist, Heidi L. Schubert, Brian N. Kelly, Gina C. Hill, James M. Holton, and Christopher P. Hill. Ubiquitin recognition by the human tsg101 protein. *Molecular Cell*, 13(6):783 – 789, 2004. 2.2.1

[254] F. Tama, F.X. Gadea, O. Marques, and Y.H. Sanejouand. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Struct., Funct., & Genetics*, 41(1):1–7, 2000. 1.1.1

[255] F. Tama and Y.H. Sanejouand. Conformational change of proteins arising from normal modes calculations. *Prot. Engg.*, 14:1–6, 2001. 1.1.1

[256] P. Taylor, J. Dornan, A. Carrello, R. F. Minchin, T. Ratajczak, and M. D. Walkinshaw. Two structures of cyclophilin 40: folding and fidelity in the tpr domains. *Structure*, 9(5):431–438, 2001. 6.2.1

[257] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22(22):4673–4680, 1994. 6.2.1

[258] Li Tian and Richard A. Friesner. Qm/mm simulation on p450 bm3 enzyme catalysis mechanism. *Journal of Chemical Theory and Computation*, 5(5):1421–1431, 04 2009. 5.1.2

[259] M.M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996. 1.1.1

[260] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977. 5.1.2

[261] Alexander L. Tournier and Jeremy C. Smith. Principal components of the protein dynamical transition. *Phys. Rev. Lett.*, 91(20):208106, Nov 2003. 2.1

[262] F. E. Vajdos, S. H Yoo, M. Houseweart, W. I. Sundquist, and C. P. Hill. Crystal structure of cyclophilin a complexed with a binding site peptide from the hiv-1 capsid protein. *Protein Sci.*, 6:2297–2307, 1997. 5.2.1, 5.2.2, 6.2.1

[263] D. M. van Aalten, J. B. Findlay, A. Amadei, and H. J. Berendsen. Essential dynamics of the cellular retinol-binding protein–evidence for ligand-induced conformational changes. *Protein Eng.*, 8(11):1129–1135, 1995. 1.1.3

[264] D. M. F. Van Aalten, B. L. De Groot, J. B. C. Findlay, H. J. C. Berendsen, and Amadei A. A comparison of techniques for calculating protein essential dynamics. *J. Comp. Chem.*, 18(2):169–181, 1997. 1.1.3

[265] G. van Rossum. Python reference manual. Technical Report CS-R9525, CWI, 1995. 2.2.2

[266] K I Varughese, M M Skinner, J M Whiteley, D A Matthews, and N H Xuong. Crystal structure of rat liver dihydropteridine reductase. *Proceedings of the National Academy of Sciences of the United States of America*, 89(13):6080–6084, 1992. 7.2.1

[267] M. Vedadi, J. Lew, J. Artz, M. Amani, Y. Zhao, A. Dong, G.A. Wasney, M. Gao, T. Hills, S. Brokx, W. Qiu, S. Sharma, A. Diassiti, Z. Alam, M. Melone, A. Mulichak, A. Wernimont, J. Bray, P. Loppnau, O. Plotnikova, K. Newberry, E. Sundararajan, S. Houston, J. Walker, W. Tempel, A. Bochkarev, I. Kozieradzki, A. Edwards, C. Arrowsmith, D. Roos, K. Kain, and R. Hui. Genome-scale protein expression and structural biology of plasmodium falciparum and related apicomplexan organisms. *Mol.Biochem.Parasitol.*, 151:100–110, 2007. 6.2.1

[268] Senadhi Vijay-kumar, Charles E. Bugg, and William J. Cook. Structure of ubiquitin refined at 1.8åresolution. *Journal of Molecular Biology*, 194(3):531 – 544, 1987. 2.2.1, 2.2.1

[269] A. Joshua Wand. Dynamic activation of protein function: A view emerging from nmr spectroscopy. *Nat Struct Mol Biol*, 8(11):926–931, 2001. 1.1

[270] Lin Wang, Nina M. Goodey, Stephen J. Benkovic, and Amnon Kohen. Coordinated effects of distal mutations on environmentally coupled tunneling in dihydrofolate reductase. *Proceedings of the National Academy of Sciences*, 103(43):15753–15758, 2006. 7.1

[271] Y. Wang, A. J. Rader, I. Bahar, and R. L. Jernigan. Global ribosome motions revealed with elastic network models. *J. Struct. Biol.*, 147:302–314, 2004. 1.1.1

[272] A. Warshel. *Computer modeling of chemical reactions in enzymes and solutions*. John Wiley and Sons, 1991. 6.2.1, 7.2

[273] A. Warshel and J. Villa-Frexia. Comment on: effect of active site mutation of phe93trp in the horse liver alcohol dehydrogenase enzyme on catalysis: A molecular dynamics study. *J. Phys. Chem. B*, 107:12370–12371, 2003. 7.2, 7.2.1

[274] Arieh Warshel. Molecular dynamics simulations of biological reactions. *Accounts of Chemical Research*, 35(6):385–395, 04 2002. 1.2.2

[275] Arieh Warshel, Pankaz K. Sharma, Mitsunori Kato, Yun Xiang, Hanbin Liu, and Mats H. M. Olsson. Electrostatic basis for enzyme catalysis. *Chemical Reviews*, 106(8):3210–3235, 06 2006. 1, 1.2.2

[276] Arieh Warshel and Robert M. Weiss. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *Journal of the American Chemical Society*, 102(20):6218–6226, 09 1980. 1.2.2, 5.1.2, 6.2.1

[277] James B. Watney, Pratul K. Agarwal, and Sharon Hammes-Schiffer. Effect of mutation on enzyme motion in dihydrofolate reductase. *Journal of the American Chemical Society*, 125(13):3745–3750, 03 2003. 7.3

[278] Paul L. Weber, Stephen C. Brown, and Luciano Mueller. Sequential proton nmr assignments and secondary structure identification of human ubiquitin. *Biochemistry*, 26(23):7282–7290, 11 1987/11/01/. 2.2.1

[279] W. Whiteley. *Rigidity of Molecular structures: generic and geometric analysis*. Rigidity Theory and Applications. Kluwer Academic/ Plenum, New York, 1999. 4.3.4

[280] M. Whitlow, A. J. Howard, D. Stewart, K. D. Hardman, L. F. Kuyper, D. P. Baccanari, M.E. Fling, and R.L. Tansik. X-ray crystallographic studies of candida albicans dihydrofolate reductase. high resolution structures of the holoenzyme and an inhibited ternary complex. *J. Biol. Chem.*, 272:30289–30298, 1997. 6.2.1

[281] R. Wolfenden. Transition state analogues for enzyme catalysis. *Nature*, 223:704–705, 1969. 6.1

[282] Adam W. Van Wynsberghe and Qiang Cui. Interpreting correlated motions using normal mode analysis. *Structure*, 14(11):1647 – 1653, 2006. 2.1

[283] AH Xie, AFG van der Meer, and RH Austin. Excited-state lifetimes of far-infrared collective modes in proteins. *Phys. Rev. Lett.*, 88(1), 2002. 3.1

[284] D. Xu, S. Yan, L. Zhang, H.-J. Zhang, Z. Liu, and H.-Y. Shum. Concurrent subspaces analysis. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 203–208, 2005. 4.3.2

[285] H. Yanagawa, K. Yoshida, C. Torigoe, J. S. Park, K. Sato, T. Shirai, and M. Go. Protein anatomy: functional roles of barnase module. *J. Biol. Chem.*, 268(8):5861–5865, 1993. 4.4.2

[286] L. Yang, E. Eyal, C. Chennubhotla, J. Lee, A. M. Gronenborn, and I. Bahar. Insights into equilibrium dynamics of proteins from comparison of nmr and x-ray data with computational predictions. *Structure*, 15:1–9, 2007. 2.1

[287] Lee-Wei Yang and Ivet Bahar. Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. *Structure*, 13(6):893 – 904, 2005. 1.2.2, 2.4.2, 8.2.1

[288] Lei Yang, Guang Song, and Robert L. Jernigan. How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophysical Journal*, 93(3):920 – 929, 2007. 1.1.2

[289] Z. Yang, I. Bahar, and W. Widom. Vibrational dynamics of icosahedrally symmetric biomolecular assemblies compared with predictions based on continuum elasticity. *Biophys. J.*, 96:4438–4448, 2009. 1.1.1

[290] B. Yener, E. Acar, P. Aguis, K. Bennett, S. Vandenberg, and G. Plopper. Multiway modeling and analysis in stem cell systems biology. *BMC Systems Biology*, 2(1):63, 2008. 4.3.4

[291] Ji Oh Yoo, A. Ramanathan, and C. J. Langmead. Pytensor: A python based tensor library. Technical Report CMU-CS-10-102, Carnegie Mellon University, 2010. 4.4

[292] X Yu and DM Leitner. Vibrational energy transfer and heat conduction in a protein. *J. Phys. Chem. B*, 107(7):1698–1707, 2003. 3.1

[293] G. Zaccai. How soft is a protein? a protein dynamics force constant measured by neutron scattering. *Science*, 288:1604–1607, 2000. 1.1

[294] M. I. Zavodszky, M. Lei, M. F. Thorpe, A. R. Day, and L. A. Kuhn. Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins: Struct., Funct. and Bioinformatics*, 57(2):243–261, 2004. 1.1

[295] Wenjun Zheng, Bernard R. Brooks, and D. Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Nat. Acad. Sci. U. S. A.*, 103(20):7664–7669, 2006. 1.1.3

[296] Pavel I. Zhuravlev, Christopher Kroboth Materese, and Garegin A. Papoian. Deconstructing the native state: Energy landscapes, function, and dynamics of globular proteins. *The Journal of Physical Chemistry B*, 113(26):8800–8812, 05 2009. 2.1

[297] Pavel I Zhuravlev and Garegin A Papoian. Functional versus folding landscapes: the same yet different. *Current Opinion in Structural Biology*, 20(1):16 – 22, 2010. Folding and binding / Protein-nucleic acid interactions. 2.1

[298] A. Zhuravleva, D. M. Korzhnev, S. B. Nolde, L. E. Kay, A. S. Arseniev, M. Billeter, and V. Y. Orekhov. Propagation of dynamic changes in barnase upon binding of barstar: An nmr and computational study. *J. Mol. Biol.*, 367(4):1079–1092, 2007. 4.4.2, 4.5