July 2017

Modeling of Dynamic Allostery in Proteins Enabled by Machine Learning

Mohsen Botlani-Esfahani
University of South Florida, mohsenb@mail.usf.edu

Modeling of Dynamic Allostery in Proteins Enabled by Machine Learning

by

Mohsen Botlani

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Cell Biology, Microbiology and Molecular Biology
College of Art and Sciences
University of South Florida

Major Professor: Sameer Varma, Ph.D.
Bin Xue, Ph.D.
Sagar Pandit, Ph.D.
Yicheng Tu, Ph.D.

Date of Approval:
June 19, 2017

Keywords: Protein Allostery, Prediction, Machine Learning

DEDICATION

To my Maman and Baba

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Sameer Varma for all of his advises and helpful conversation on scientific and technical aspect of this study. I also thank my committee members Dr. Bin Xue, Dr. Sagar Pandit and Dr. Yicheng Tu who provided constant guidance and have evaluated my progress. I also appreciate my lovely family and friends specially Mr. Hadi Khoshnevis, Dr. Alireza Chakeri, Dr. Reza Mohamadi and Dr. Priyanka Dutta for being wonderful friends and helping me during these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Regulation of protein activity is essential for normal cell functionality. Many proteins are regulated allosterically, that is, with spatial gaps between stimulation and active sites. Biological stimuli that regulate proteins allosterically include, for example, ions and small molecules, post-translational modifications, and intensive state-variables like temperature and pH. These effectors can not only switch activities on-and-off, but also fine-tune activities. Understanding the underpinnings of allostery, that is, how signals are propagated between distant sites, and how transmitted signals manifest themselves into regulation of protein activity, has been one of the central foci of biology for over 50 years. Today, the importance of such studies goes beyond basic pedagogical interests as bioengineers seek design features to control protein function for myriad purposes, including design of nano-biosensors, drug delivery vehicles, synthetic cells and organic-synthetic interfaces. The current phenomenological view of allostery is that signaling and activity control occur via effector-induced changes in protein conformational ensembles. If the structures of two states of a protein differ from each other significantly, then thermal fluctuations can be neglected and an atomically detailed model of regulation can be constructed in terms of how their minimum-energy structures differ between states. However, when the minimum-energy structures of states differ from each other only marginally and the difference is comparable to thermal fluctuations, then a mechanistic model cannot be constructed solely on the basis of differences in protein structure. Understanding the mechanism of dynamic allostery requires not only assessment of high-dimensional conformational ensembles of the various individual states, including inactive, transition and active states, but also relationships between them. This

challenge faces many diverse protein families, including G-protein coupled receptors, immune cell receptors, heat shock proteins, nuclear transcription factors and viral attachment proteins, whose mechanisms, despite numerous studies, remain poorly understood. This dissertation deals with the development of new methods that significantly boost the applicability of molecular simulation techniques to probe dynamic allostery in these proteins. Specifically, it deals with two different methods, one to obtain quantitative estimates for subtle differences between conformational ensembles, and the other to relate conformational ensemble differences to allosteric signal communication. Both methods are enabled by a new application of the mathematical framework of machine learning. These methods are applied to (a) identify specific effects of employed force fields on conformational ensembles, (b) compare multiple ensembles against each other for determination of common signaling pathways induced by different effectors, (c) identify the effects of point mutations on conformational ensemble shifts in proteins, and (d) understand the mechanism of dynamic allostery in a PDZ domain. These diverse applications essentially demonstrate the generality of the developed approaches, and specifically set the foundation for future studies on PDZ domains and viral attachment proteins.

CHAPTER 1

INTRODUCTION

Regulation of protein activity is essential for normal cell functionality. Many proteins are regulated allosterically, that is, with spatial gaps between stimulation and active sites. Biological stimuli that regulate proteins allosterically include, for example, ions and small molecules, post translational modifications, and intensive state-variables like temperature and pH. These effectors can not only switch activities on-and-off, but also fine-tune activities.

Understanding the underpinnings of allostery, that is, how signals are propagated between distant sites, and how transmitted signals manifest themselves into regulation of protein activity, has been one of the central foci of biology for over 50 years [1, 2, 3]. Today, the importance of understanding allosteric mechanisms goes beyond basic pedagogical interests as bioengineers seek design features to control protein function for myriad purposes, including design of nano-biosensors, drug delivery vehicles, synthetic cells and organic-synthetic interfaces [4]. Furthermore, the drug industry seeks solutions to design new therapeutics that can control cellular function from outside the cell, and without need for drugs to penetrate cellular membranes [5]. Drugs are also being designed that can modify protein function without directly interfering with their catalytic sites [6].

For many decades, allostery has been described using two different models: the Monod-Wyman-Changuex(MWC) model [3] and the Koshland-Nemethy-Filmer(KNF) model [7]. The basic idea behind the MWC model is that regardless of the presence of the effector, the protein samples conformations belonging to both active and inactive states, and the effector simply biases the sampling probability toward one state. In contrast, the KNF model

proposes that the protein samples conformations that are uniquely defined by the effector, that is, there is no overlap between conformations sampled in the absence and presence of the effector. Nevertheless, both models are phenomenological in that they don't provide any direct mechanistic and structural insight into how allosteric communication occurs between distant sites [8].

Although the two models above are phenomenologically different, they are both based on the assumption that allosteric signal propagatation and control of the active site occurs through changes in structure. However, there is now mounting evidence in literature, especially in the last decade, that effector-induced changes in entropy or conformational fluctuations also contribute to allosteric control in many protein families [9]. In fact, in 1972 Weber did propose a more general model in which allosteric signals propagated and controlled activity through effector-induced changes in conformational densities [10]. In 1984, Darden and Cooper took this work further and showed theoretically that effector-induced changes in entropy would be more pronounced in cases where the effector induced only minor structural changes in proteins [11]. Indeed, many such proteins have been identified since then, including G protein coupled receptors, nuclear transcription factors, heat shock proteins, immune cell receptors and viral attachment proteins [9]. Effectors induced only minor structural changes in these proteins that are comparable to thermal fluctuations (see Figure 1.1). Despite numerous studies on these proteins, the mechanisms of how "dynamic allostery" regulates their activities remains poorly understood.

The current phenomenological view of allostery is that signaling and activity control occurs via effector-induced changes in protein conformational ensembles [12]. If the structures of two states of a protein differ from each other significantly, then an atomically detailed model of regulation can be constructed in terms of how their minimum-energy structures differ between states. But in such cases, the minimum-energy structures of proteins must differ significantly between states, so that thermal fluctuations are negligible compared to

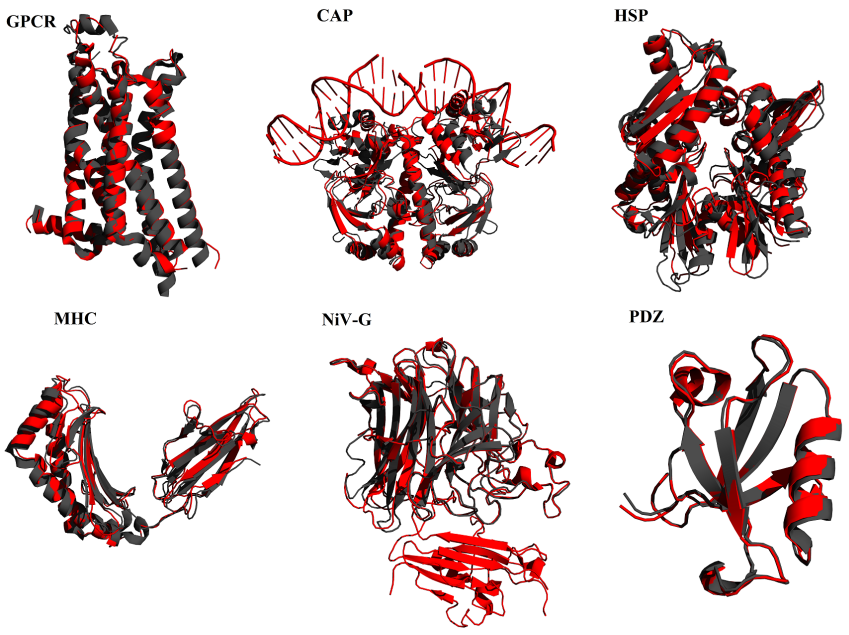

Figure 1.1. Dynamic allostery in different protein families (G protein coupled receptors (GPCR), Catabolite Activator Protein (CAP), Heat shoch protein (HSP), Major histocompatibility comple (MHC), Nipah attachment protein (NiV-G), PDZ).

structural differences. However, when the minimum-energy structures of states differ from each other only marginally and the difference is comparable to thermal fluctuations, then a mechanistic model cannot be constructed solely on the basis of difference in protein structure (Figure 1.2) [11, 13, 8, 14, 15, 12, 16, 17]. Understanding the mechanism of dynamic allostery requires not only assessment of the conformational ensembles of the various individual states, including inactive, transition and active states, but also relationships between them.

Experimental techniques can be used understand dynamic allostery, however, they provide limited information at the molecular level. Some studies used mutagenesis to characterize allostery pathways. Mutagenesis not only deos not provide the whole pathway but also even if a mutation of residue disrupt the pathway there might be multiple other paths in wild-type that would be activated to traduce the allostery signal [18]. On the other hand even thought NMR provide insights to allostery mechanism but it has protein size limita-

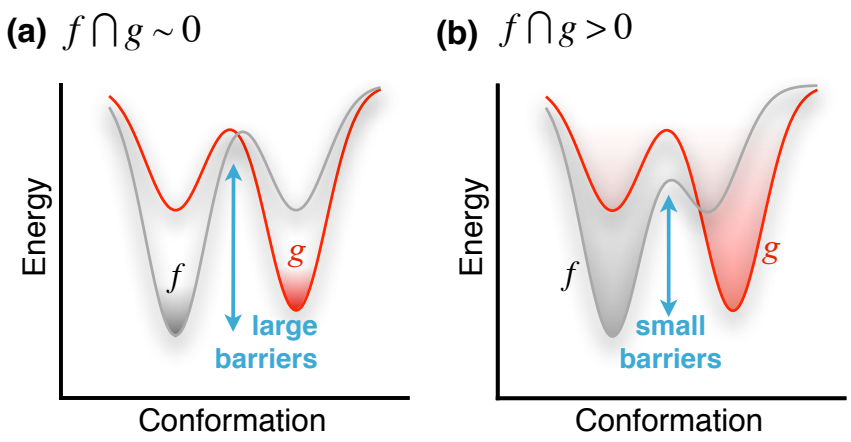

Figure 1.2. For many proteins, the minimum-energy structures of their thermodynamic states differ significantly from each, and the differences in thermal fluctuations are negligible. Part (a) depicts this scenario using a simplified 2-state schematic in which there is negligible overlap between conformational ensembles of two states, that is $f \cap g \sim 0$. Consequently, regulatory models can be constructed in terms of how their structures differ between states. Part (b) depicts an alternative scenario where structural differences between protein states are comparable to thermal fluctuations, and the overlap between conformational ensembles is non-negligible.

tions. Moreover, for small protein it can identify changes in structure and dynamics of a subset of residues, but it cannot link changes to signal propagation [19]

In contrast to experiments, molecular simulation techniques that sample conformational ensembles as a function of energy, can provide direct insight into dynamic allostery. Numerous techniques have been developed to generate conformational ensembles of proteins [20]. Methods have also been developed to relate inter-state differences in structure to allosteric regulation [21, 22, 23]; however, none account for thermal fluctuations. These methods typically rely on average structural differences between states, which renders them unsuitable for studying proteins in which inter-state differences in structure are comparable to thermal fluctuations; but we note that these methods were not intentionally designed to account for thermal fluctuations. Methods have also been developed to connect correlations in thermal fluctuations to signaling [24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. These inter-site fluctuation

correlations can be combined with each other and with the spatial organization of the protein to yield insight into how different spatial regions communicate with each other (intra-state signaling). However, since no information on divergence from a reference state is incorporated, these approaches are not theoretically capable to provide insight into regulatory mechanisms. These methods are discussed in more detail in chapter 2.

This dissertation deals with the development and application of methods that significantly boost the applicability of molecular simulation techniques to probe dynamic allostery. Chapter 3 details a method developed by Leighty and Varma [34] that allows quantification of subtle differences between conformational ensembles of two states in terms of physically meaningful metric, and presents a new indexing scheme that significantly accelerates the machine learning based algorithm. This method overcomes the challenge of finding appropriate feature spaces (or summary statistics) for distinguishing ensembles, and provides a comprehensive difference between ensembles that naturally embodies differences in thermal fluctuations. Chapter 4 presents three new applications that this method enables: (a) identification of specific effects of employed force fields on conformational ensembles [35], (b) comparison of multiple ensembles against each other for determination of common signaling pathways induced by different effectors [36], and (c) identification of the effects of point mutations on conformational ensemble shifts in proteins [37].

Chapter 5 presents a new machine learning enabled method that yields relationships between conformational ensemble differences and allosteric signaling pathways. As such, differences between conformational ensembles do not inform us of how signals propagate. An understanding of signal propagation requires a quantitative analysis of correlations in induced ensemble shifts, which the new method allows us to compute and then link to signaling probabilities. This method permits us to directly address fundamental issues in dynamic allostery that remain unresolved. For example, is "dynamic allostery" aptly termed in that regulation occurs due entirely to induced changes in dynamics or do small changes in

energy-minimum structures also contribute? In either case, can we define cutoffs in structural changes, such as in center-of-mass (CoM), below which their contributions to regulation are insignificant? Are there relationships between a residue's propensity to contribute to regulation, and its spatial location or hydrophobicity? If a residue contributes significantly to spatial communication within a state (intra-state signaling), then is it justified to assume that it is also important to propagation of regulatory signals? Do stimulator-binding and unbinding responses occur in the same manner? In general, how different are activating signals from deactivating signals?

Chapter 6 provides an application of this method to understand dynamic allostery in a PDZ domain. PDZ domains are part of many diverse families of proteins where one of their main tasks is to transduce regulatory signals across domains. Our application to the PDZ2 domain of human phosphatase PHPT1E reconciles data from site-directed mutagenesis and NMR experiments. Finally, chapter 7 summarizes the finding and outlines future directions.

CHAPTER 2

BACKGROUND

2.1 Dynamic Allostery

2.2 Classical view

Historic investigations on cooperative oxygen binding of hemoglobin launched a major scientific effort to characterize long-range intra-protein communications [38]. There were two dominant main models to describe allosteric mechanism for decades, the Monod-Wyman-Changuex (MWC) model [3] and the Koshland-Nemethy-Filmer (KNF) model [7]. The MWC model assumes that in the absence of the effector protein samples both active and inactive states with different probabilities (reverse correlation with their free-energy level). The energy level of active state conformation is higher but presence of the effector lowers the free-energy of the active state enough to trap the proteins in it [20]. On the other hand, KNF model proposed that protein only visits inactive state in the absence of the effector but protein undergoes conformational change to active state conformation by interacting with the effector. In other words the MWC model describes the mechanism as the collective motions of many atoms simultaneously comparing to the KNF model that describes it as a sequential change from inactive to active state (Figure 2.1). Both proposed models were based on conformational change between two states with two defined structures [7]. However, since these two models are more phenomenological they dont provide structural details about allosteric communication between sites [8].

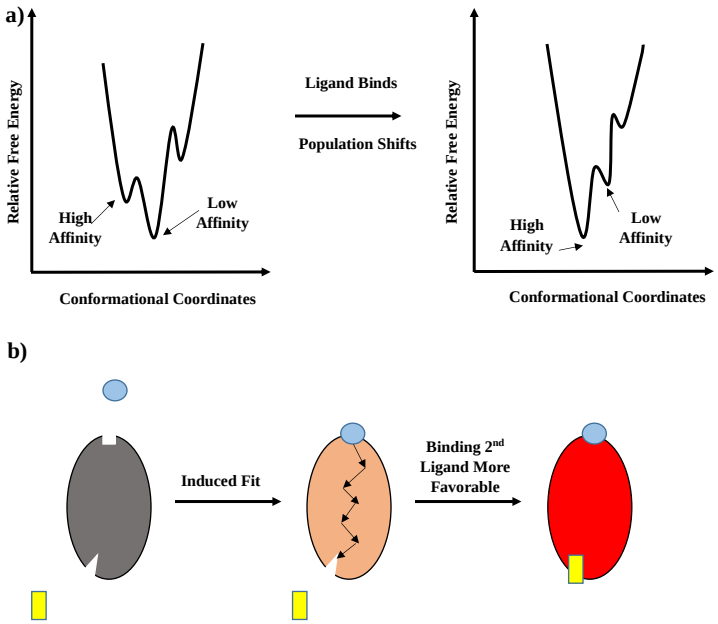

Figure 2.1. a) MWC model b) KNF model (Adapted with permission form [39]).

2.3 Ensemble view

This first study that provided structural insight analyzed the cooperativity of hemoglobin by high-resolution X-ray structures [40, 41]. This study initiated many other similar studies based on the structural view of the allosteric mechanism. Some studies also proposed the existence of conserved allosteric pathways [42]. Evidences such as finding alternating conformations for an active state [43] suggest that understanding allosteric mechanism not only requires structural changes between two states but also the factors that cause structural change. Cooper and Dryden demonstrated the contribution of the entropy in allosteric mechanism their study initiated many other similar studies. They used a statistical thermodynamic formalism to show that just changes in the frequency and amplitude of thermal fluctuations in protein could achieve cooperative binding energies on the order of a few kcal per mol [11, 12]. This new view explains why some allostery mechanisms are not detectable with end-state structure analysis. The term dynamic allostry is coined which describes the

role of entropy in allosteric mechanism [8]. Weber proposed that ligand binding merely shifts the population of conformational states [10]. After decades experimental and theoretical studies revealed that conformational states in a pre-existing equilibrium can influence the function of a protein [44, 45, 46]. Figure 2.2 schematically represent several different ways of the energy landscape remodeling which enables protein to communicate the signal by altering protein dynamics [47]. During folding, a protein moves down the energy funnel from many non-native states to the boxed region that represents an ensemble of conformations that are energetically accessible to the protein. Figure 2.2.b shows one way in which protein regulation occurs by narrowing the width of a single energy well. This reduces protein dynamics resulting in structural rigidification of the same average conformation. Figure 2.2.c demonstrates another way in which protein may be in equilibrium between two distinct conformational states and the effector can alter the relative energies of the wells and, consequently, their relative occupancies. Figure 2.2.d is a variation of c in which the sampling of a higher-energy state in the absence of an effector provides a pathway toward a signal-induced conformational change – the energy landscape is not only narrower but also shifted due to the signal [47].

The concept of re-distribution of conformational states rationalized many allosteric regulations [48, 49, 50, 51, 52, 53, 54, 55, 9]. Since all proteins obey the same physical principles and all proteins exist as a population of conformational states, population shift is their underlying allosteric mechanism. This hypothesis and several observations of allostery signals on protein systems that were considered as non-allosteric proteins spanned the continuum of structure/dynamics classification space [12]. This classification is schematically presented in Figure 2.3. In this figure in one end of the spectrum we can see the human hemoglobin with two different T and R states with multimeric reorientation mechanism [40, 41]. The allostery mechanism of this protein supports the old idea that allostery mechanisms can be explained by analysis of observable changes in the ensemble of average structures. The next

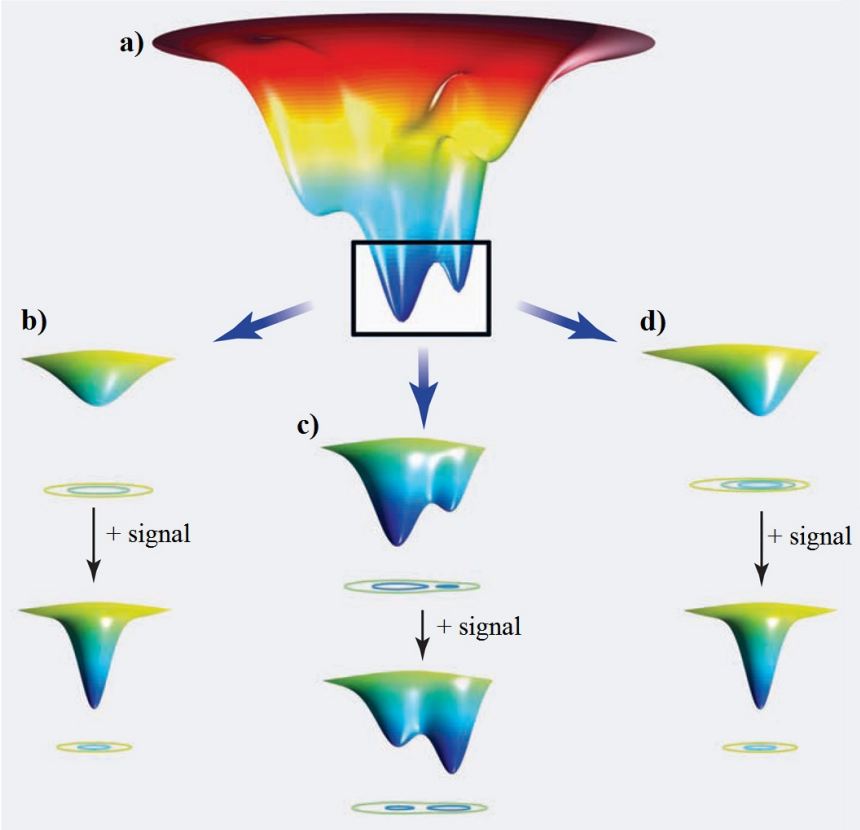

Figure 2.2. Energy landscape remodeling, altering the protein dynamics for signal communication. a) by folding the protein moves down the energy funnel to its native states(higher energy in red and lower energy in blue). b) narrowing the width of a single energy well(structural rigidification). c) altering the relative energies of the wells therefore their relative occupancies. d) is a variation of c in which narrowing and shifting of the well happen simultaneously due to the signal.(adapted with permission from [47])

system is the PDZ domain which NMR experiment showed neither large global structural change nor significant change in backbone dynamics but only detectable change was side-chain dynamics [56, 57]. Catabolic gene activator (CAP) is a DNA binding protein, it is a homodimeric transcription factor. Studies revealed the CAP allosteric response upon cAMP binding is only due to conformational entropy of backbone and side-chain [58, 59, 60]. They also proposed quenching of dynamics upon ligand binding as an entropy penalty of allostery mechanism [8]. The analysis of allostery mechanism of the PDZ and the CAP systems are

very important studies which not only revealed the importance of dynamics in allostery but also demonstrated the limitations of the static view of allostery mechanism.

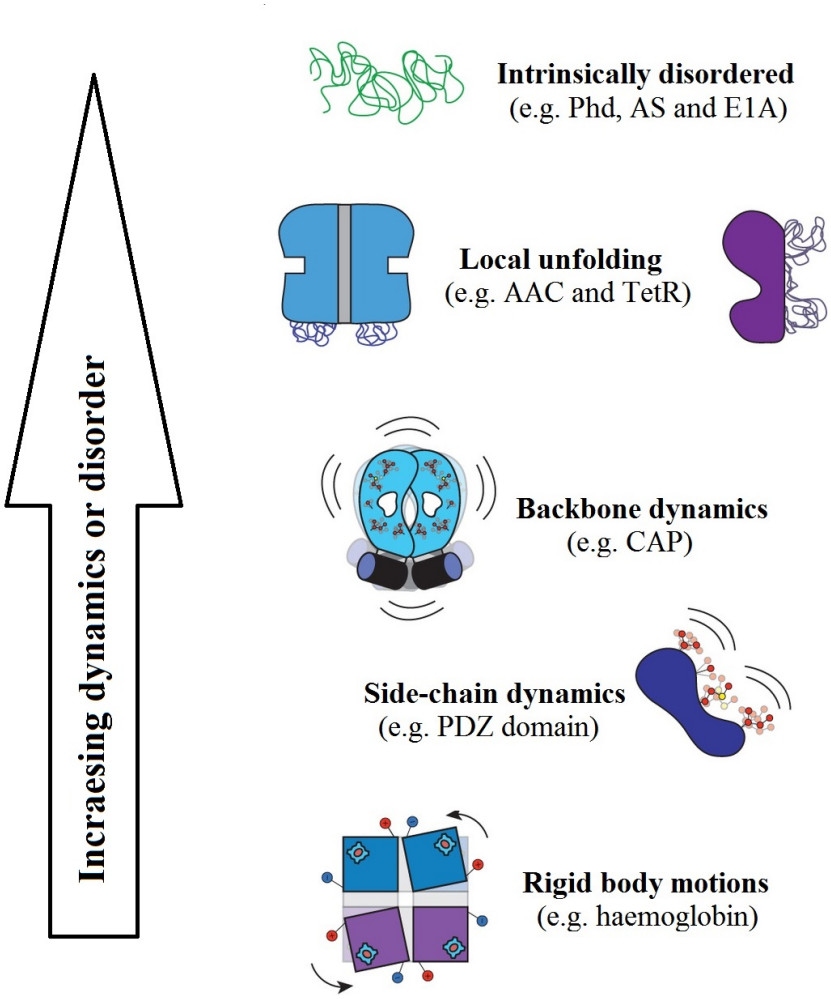

Figure 2.3. The dynamic continuum of allosteric phenomena. Schematic representation of allosteric systems with increasing dynamics, disorder or fluctuations on the vertical axis(adapted with permission form [12]).

The next allostery mechanism with higher contribution of conformational dynamics is local unfolding. As an example for this mechanism a homodimeric enzyme of Enterococcus faecium is called aminoglycoside N-(6)-acetyltransferase II (AAC) which is responsible for antibiotics resistance against aminoglycoside [61]. This enzyme shows positive cooperativity in low temperatures and negative cooperativity in higher temperatures upon acetyl-CoA

binding. This observation agrees with the change of conformational entropy with temperature change, this nonlinear dependency is a signature of local unfolding allostery mechanism [62].

Perhaps the most interesting behavior is observation of allostrery signals in intrinsically disordered proteins (IDPs) with highest conformational fluctuations due to the lack of tertiary structures [63, 64, 65, 66, 67]. Phd/Doc toxin-antitoxin system it is an inhibitor of ribosome A site. This system shows conditional cooperativity with this mechanism and switches between inhibitor and activator of transcription [68]. This spectrum of different contributions of entropy in allostery which leads to many different structural pictures but population shift or re-distribution of conformational states rationalizes these mechanisms [12, 9].

2.4 Experimental approaches

Development of experimental biophysical methods improved our understanding of protein dynamics and helped appreciate its role in biological processes [69, 70, 71, 66]. X-ray crystallography is the major method to resolve 3D structure of proteins. In X-ray crystallography the structure of the protein obtained by finding the position and connectivity of atoms from the map of the electron density. Electron density map itself, inferred from the dispersion of a X-ray beam which is shined through a crystallized protein. This method is one the most widely used methods to study allostery because of the high-resolution structures enables us to find changes in inter-atomic interactions.

Since X-ray crystallography only provide static view it can be used in combination of other methods or provides several structures under different conditions. For example a crystallographic study solved both the unbound as well as leucine-bound structures of a-isopropylmalate synthase (a-IPMS) enzyme of Mycobacterium tuberculosis which is inhibited by binding of leucine. It clarified the location of leucine binding but because the structural

difference was small it could not explain the allostery mechanism [72]. However, later another study by using Hydrogen-deuterium exchange mass spectrometry revealed a large change in dynamics of a network of residues going from the binding site on one domain to the allostery site in another domain of the enzyme [20, 73]. More recently, a time-resolved x-ray diffraction technique was applied to study structural changes of Scapharca dimeric hemoglobin due to a ligand photo-dissociation. An intermediate structure has been seen, with changes at the heme groups their neighboring residues and water molecules at the interface [74].

Nuclear magnetic resonance is also used to determine protein structure. Additionally, NMR can provide information about different motions of a protein such as high-frequency motions of atoms as well as low-frequency motions of entire protein domain, which makes NMR a valuable tool to study dynamic allostery [75]. There are three major NMR experiments for studying dynamic allostery of proteins: chemical shifts, spine-relaxation methods and residual dipolar coupling (RDC).

The chemical shift is the relative change of the resonance frequency of each atomic nucleus, due to its local magnetic environment, to a standard frequency. This method usually is used by comparing the chemical shifts of two states to identify residues that undergo changes in chemical shift [61]. Moreover, chemical shift is capable to differentiate between secondary and tertiary structure transitions [76].

NMR spine relaxation uses this physical property that the rate at which a disturbed molecular system returns to its equilibrium is related to the identity and motion frequency of atoms [20]. This information especially for proteins with small structural changes can be used to detect allostery signal propagation. For example, one can track the residues that undergo changes in motion connecting the binding site to distal allostric site [77]. NMR can also detect the less populated states and their transition rates which is helpful for understanding allosteric mechanisms in proteins [58, 78]. Magnetic dipole interactions between atomic nuclei are averaged out due to the protein free rotation, if the protein is

immersed in a solution. However, if the protein is immersed in a liquid crystal phase partial molecular alignment will lead to incomplete averaging and these interactions (i.e residual dipolar couplings) can be recovered [20]. This analysis provides information about bonds orientations which are sensitive to small structural changes therefore, they can be used to study allostery signals. RDC has been used to determine which allostery model (MWC vs. KNF) is better describes behavior of a system [79].

Even thought NMR provide insights to allostery mechanism but it has protein size limitations. Moreover, for small protein it can identify changes in structure and dynamics of a subset of residues, but it cannot link changes to signal propagation [19]. There are other experimental approaches such as: Fluorescence resonance energy transfer (FRET), Hydrogendeuterium exchange mass spectrometry (HDX) and Atomic force microscopy (AFM) to acquire insights into dynamic allostery [20]. None of these method provide a completer view of dynamics allostery mechanism.

2.5 Computational approaches

The study of allostery is perhaps the best example in structural biology where experimental and computational methods complement and reinforce one another. In this case computational methods are not just a set of tools to complements existing experimental approaches. Particularly molecular simulation and the accompanying analysis can provide answer to questions about the structure and dynamics of the protein that are beyond the capability of modern experimental techniques. On the other hand computational methods need to be validated by experimental approaches [20]. In general the mechanism of allostery at atomic level is mostly based on mechanical operations, changes in dynamics and entropy within a solvated protein. However, experimental methods can resolve a portion structure and minimum information on dynamics, computational methods can provide more details. Exceptionally MD simulation can acquire more details on changes in position, dynamics and

underlying forces in a complex network of atoms compare to any other technique. Despite development of numerous computational methods during last few decades to uncover the allosteric mechanisms within proteins, they have varying degrees of success, but there is no universal technique because of underlying approximations [20]. However, since most of the computational methods are closely related, improvement of one method can potentially cause advances in others.

Next we summarize the various computational methods for studying allostery and discuss their pros-and-cons in providing insights into dynamic allostery. We divide the existing methods into three different major groups based on their underlying assumptions and information they used to develop their methods.

2.5.1 Structure and Evolution based Models

Several methods have been developed that use primarily protein sequence data to detect allostery pathways and binding site. There are two main category of these models, single site and coupled site methods. Single site methods provide a list of individual conserved amino acids in the sequence which are potentially functional but they dont suggest any linkage between them. Coupled methods produce a list of groups of two or more amino acids which appear to have liked effect on function based on their coevolution [5]. There are different metrics for single site sequence-analysis such as Shannon entropy [80], relative Shannon entropy or Kullback-Leibler divergence [81] and von Neumann entropy [82].

On the other hand coupled-site methods look for residue pairs which mutate together with higher frequency than random genetic events. They also calculate amino acids correlation strength. One of the most widely used methods for allostery prediction is the statistical coupling analysis (SCA) developed by Lockless and Ranganthan [42]. This method calculates the change in individual and joint conservation due to different perturbations to establish a coupling energy that indicates evolutionary coupling of two positions in the sequence.

There are challenges that all sequence-analysis methods face such as selection and aggregation of relevant input sequences, interpretation and integration with other type of data such as 3D structure. Moreover, determination of biological relevance of a strong signal is very challenging without further information. For example it is difficult to determine if the conserved residues plays role in allostery or structural integrity of a protein. In most sequence-based analysis to reach statistically significant results we should analyze many sequences and to use many sequences we have to lower sequence similarity criterion[5].

Therefore to acquire better insights for allostery mechanism methods incorporate structural information. The foundation of many allostery prediction tools that were developed for the molecular simulation methods are based on predictive models of the integrated fields of proteomics and bioinformatics [20]. These models were used to create databases that connect the experimental studies to computational approaches for better understanding of protein-protein interactions [83, 84, 85]. Since, structural and energetic characteristics of protein-protein interactions in residue and atomic level overlap most with those of intra-protein allostery mechanisms, development of protein-protein interaction predictive models contributed most to allostery prediction tools [20]. Almost all of the information were used in the development of these models were obtained from protein data bank (PDB) [86]. Where there are tens of thousands protein structures which hundreds of them are known to be allosteric.

An example of this influence is collection of 2300 alanine residue mutants from heterodimeric protein complexes following by an analysis on affinities and structural data. This study attempt to provide a descriptive view on mechanistic details at the resolution of amino acids and their energetic contributions. They also found structural arrangement of amino acids near the interface so called binding hot spots [87]. Another study by incorporating structural alignment to this study find cooperativity among hot spot in binding interactions

in protein-protein interfaces [88]. Figure 2.4 schematically shows analysis for one instance of this dataset.

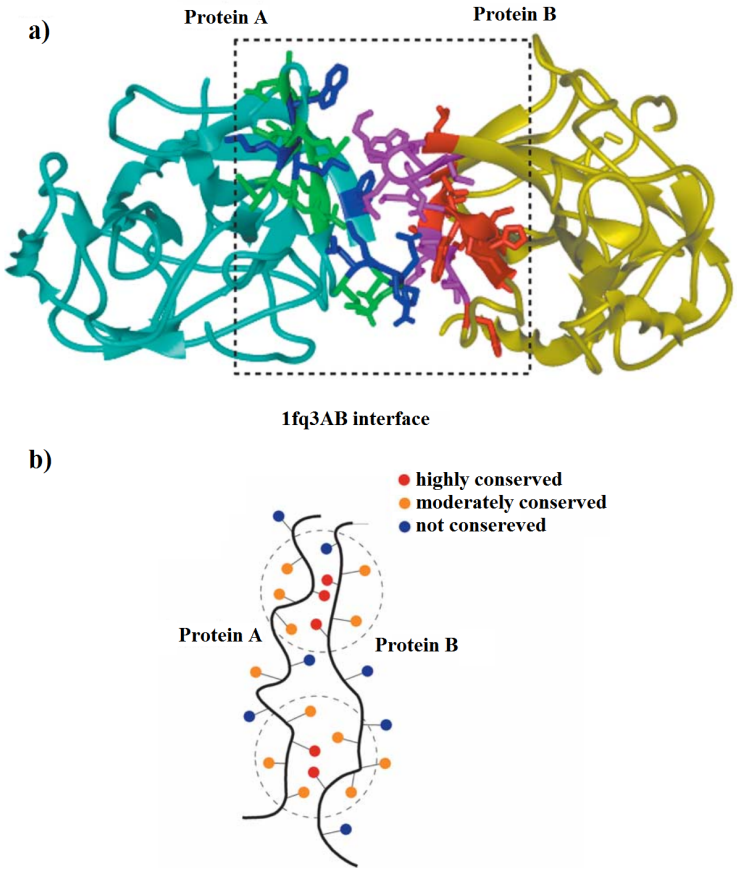

Figure 2.4. a) Graphical representation of a hetrodimeric complex b) incorporation of alignment information(adopted with permission from [88]).

Ofran and Rost tuned an existing surface prediction tool to be able to predict hot spots from sequence data alone without information on an interacting partner [89]. Later, Cho et al suggested considering extra structural features such as atomic density, solvent accessibility, hydrophobicity, noncanonical hydrogen bonds and -interactions for prediction [90]. The improvement in prediction accuracy shows high complexity nature of protein-protein interaction and related function regulations [20].

Another study assumes that by comparing the structure of proteins in different states one can obtain information about the relationship between structure and function of a protein and these structural insights can be hallmarks of allostery pathways. First such investigation compared active and inactive structures of 51 protein by measuring differences in backbone as well as side-chain positions, torsion angles and local contact patterns [22]. Based on this analysis approximately 20 percent of residues of these proteins undergo significant conformational changes. These changes are more pronounced in allosteric proteins rather than non-allosteric proteins especially in more flexible regions such as surface residues and loops.

Wolynes et al. proposed minimal frustration principle states that protein structure is a balance between stability and instability. Where stability leads to successful folding and protein integrity, but instability offers opportunities to the network of structural interactions that in some case take the form of allostery pathway [33, 91, 92]. To test this hypothesis they analyzed the same database of 51 proteins and found the frustrated regions undergo significant rearrangements between alternative states. Frustrated regions or instability regions work as pivot points between surrounding rigid (stable regions). This work demonstrates for allosteric proteins frustrated states by selecting a subset of structural interactions enables competition between similar low energy conformations that can be selected under specific conditions like ligand binding [20].

Many methods used network-based approach to predict allostery signal pathway. This approach assumes that allostery pathway is a subset of residues in the protein that create a sub-network of interaction as new communication pathways for signal transduction. In the extension of their study on local structural rearrangements [22] Daily and Gray investigate coupling among residues [22]. In this study contact network of 15 pairs of active/inactive states of allosteric proteins were created. Results showed in only 5 pairs the changes in local rearrangements were communicated from binding site to allosteric site. For the remaining 10 pairs this rearrangement must accompany large-scale conformational changes in the form of

rigid body motion. These methods were successful in characterizing allostery mechanisms of many protein systems, they also have also contributed to design of new customized proteins [6, 93, 94, 4, 95]

2.5.2 Single state models

More recent methods to uncover the dynamic allostery mechanism and incorporate entropic changes used molecular simulation methods for generating ensembles and following analysis of ensembles. The following analysis consist of constructing a network based on correlated motions [96, 97, 30, 98, 99] and then using graph theory analysis to find potential allostery pathway/s [29] and community structures as subnetworks as well as hubs [97, 100].

Even though molecular simulation provide invaluable information about the dynamics of protein systems at atomic resolution also offers a variety of possible analysis techniques, there are two important limitations that one should consider for using it. First limitation is computational cost which is highly dependent on the system size as well as the level of approximation. For example for atomic simulation of a small protein in a solution the calculation consists of: atomic position, momentum and interactions with all of the neighboring atoms for thousands of atoms. The expense of these calculation exponentially grows with the increase of the system size and number of replicas. The second limitation is related to the level of approximation in potential function (force fields) that define how atoms interact with one another. For example the mathematical complexity of the most accurate potential function (quantum mechanics) makes it intractable for smallest protein [20]. Additionally, the time-scales that this complex calculation is imaginable is far shorter than the time-scales of protein structural changes that associated with protein dynamics.

Therefore, there are two major approximations in most MD force fields treating atoms as point particles and using only Newtonian equations of motion. The force fields' parameters are tuned to reproduce some experimental results. Therefore, force fields can reproduce some

properties successfully (equilibrium properties) while they are not accurately reproduce other properties (dynamic properties) [101, 80].

Molecular simulation of particles can be divided into two main categories, stochastic and deterministic. Monte Carlo is a type of stochastic simulation in which energy landscape get explored by randomly sampling various conformations. At each step the following conformation is selected then associated energy of that conformation is calculated based on a comparison of the transition probability is accepted or rejected. Since, a MC simulation is not deterministic it cannot represent a time evolution in the system. Instead it offers reaching low energy conformation without exploration of deterministic path towards that conformation [20]. Therefore, it is advantageous to use this method to explore long-timescale structural change without a long deterministic simulation. A significant disadvantage of this approach is that highly correlated motions are hard to simulate which are important in some allosteric pathways especially those with large structural changes. As an example of MC molecular simulation implementation to study protein dynamic allostery Dubey et la. were used MC to investigate long-range intra-protein communication in CAP. This long-range signal transduction can happen by correlate side-chain fluctuations alone [102]. On the other hand MD is a deterministic method that simulate time evolution of a system, resulting in a stepwise snapshots of all the particles in the system(with steps in the range of femtoseconds). Due to all of the calculation of position, momentum and many different type of interactions MD is very computationally expensive therefore usually supplemented with different kind of enhanced sampling and other techniques to overcome this limitation.

Another method that designed to interpret the results of experimental techniques that probe residue-residue interaction such as NMR spectroscopy is called pump-probe MD which first was applied on the PDZ domain of the allostery protein calmodulin [103]. In this technique they excite an amino acid by using an oscillating force with specified magnitude, direction and frequency the applied forces are transmitted to other parts of the protein. The

strength of coupling between them indicates the strength of interaction between them. This can be used to detect long-range interactions in allostery mechanism.

Most of the methods that have been developed to provide insight into allostery mechanism and were described above are based on comparing structures of active and inactive states. Many of these methods were successful to relate inter-state differences to allosteric regulations. However, since these methods typically rely on differences on average structures for each state and dont incorporate any thermal fluctuations, they are not suitable for dynamic allostery mechanism.

Normal mode analysis also is used in order to incorporate thermal fluctuation effects in allostery mechanism. In this method structural fluctuations of a protein are decomposed into harmonic orthogonal modes around its minimum energy conformation. Vibrational frequencies of a structure are inversely proportional to the amplitude of the vibration (structural flexibility) therefore, structural transition with higher probability has lower frequency modes. Low frequency modes usually present concerted motion of many atom which offers a dissipation mechanism for external perturbations. Accessible low conformations as well as cooperativity and concerted motions make these modes ideal candidates for allostery signal propagation. However the allostery signal also uses local rearrangements too for this reason is it important to incorporate some of the high-frequency motions and inter-mode coupling. Silvestre-Ryan et al. used a coupled technique of coarse-grain simulation with elastic-network model of NMA to study harmonic and anharmonic structural dynamics contributions of the protease subtilisin. The elastic network model was derived on a sequential sets of conformations were obtained from an all-atom MD simulation in order to capture the temporal variation in the mechanical coupling network of protein dynamics. Results showed that this analysis is able to detect interacting residue pairs based on their strength and variation of mechanical coupling [104]. This approach bridges all-atom and coarse-grained

modeling methods in studying allostery because the force constants used in elastic network were obtain from of MD conformations.

2.5.3 Multi state models

As it is described in previous section most of the methods that incorporate structural dynamics in allostery pathway prediction have used conformational ensemble of only the active state. The advantage of this approach is that parametrization of edges in the interaction is easy to interpret. However, since no information on divergence from a reference state is incorporated, these approaches cannot theoretically provide insight into regulatory mechanisms. In other words understanding allosteric regulation of proteins requires not only assessment of the various individual states, including inactive, transition and active states, but also relationships between states.

There are very few methods that have tried to incorporate information of inactive state in allostery pathway prediction which seems to be more consistent with new view of allostery which is shift in ensemble densities. Kong and Karplus used MD simulation and an interaction correlation analysis to determine residues involve in alloatery signal transduction of hPDZ domain. They defined cumulative energy difference, which is a difference of total energy of each residue with all of its neighbors between two active and inactive states. By using a clustering method they found two different possible allostery pathways [105].

In another study that investigates allostery in PRFAR binding to imidazole glycerol phosphate (IGP) heterodimer [106]. Rivalta et al. simulate the system in bound and apo states and then they construct dynamical networks for both states by mutual information approach. Finally the applied graph theory approaches such as community search and pathway prediction on the average network of two states and applying a frame percentage cutoff for calculating interactions. Since the first approach is based on differences in local energy of the residue between two states it does not provide any mechanistic insight to the dynamic

allostery. Additionally, neither of two methods are based on direct comparison of conformational ensembles, and therefore they are not capable to characterize regulatory nature of dynamic allostery.

2.6 Need for new methods

Characterizing dynamics allostery essentially requires two sets of methods. Firstly, methods are required to compare conformational ensembles of different states in terms of physically meaningful metrics, and secondly methods are required to relate conformational ensemble differences to allostery regulation. At the start of the thesis, Leighty and Varma had developed the very first method to quantify differences in ensembles [34], which unlike existing methods [107] did not require any ad hoc fitting of underlying distributions, and yielded differences in terms of a normalized metric that made cross comparisons of ensembles possible. Chapter 3 describes the development of this method, and the algorithms we developed for fast parallel implementation. At the start of the thesis, there were, as discussed above, no methods that incorporated multi-state information and dynamics simultaneously to construct models of dynamic allostery. We provide the very first method, which is described in chapter 5, and its application to PDZ domains, which is described in chapter 6.

CHAPTER 3

QUANTIFYING DIFFERENCES IN ENSEMBLES

Knowledge about protein structure is very important for understanding of many biological processes especially in molecular level. Structure function relation studies are the bases for many protein engineering as well as drug design studies. However, in addition to average structure intrinsic motion of the proteins around that structure plays role in function of the protein and is affected by many biological stimuli. The extent of the changes are tightly depend to extent of the biological stimuli and can greatly impact the function of proteins. A quantitative characterization of these intrinsic motions is important because it provides a basis for relating the biological stimuli to function of proteins and as a result biological processes. While comparing two protein structures are is tractable with reasonable methods such as RMSD calculation or similar metrics, quantification of conformational ensemble differences of proteins is challenging. The quantification should estimate the differences of two high-dimensional datasets with many degrees of freedom. The number of frames and coordinates of atoms are usually both in the order of thousands.

3.1 Existing methods

Bruschweiler extended the RMSD measure to compare two ensembles of conformations [108]. He defined the inter-ensemble RMSD (eRMSD), as the root of average mean square deviation between all conformations of two ensembles.

$$(eRMSD)^2 = \frac{1}{MN} \sum_{l,k=1}^{M,N} (RMSD(\mathbb{R}^{(l)}, \mathbb{R}'^{(k)}))^2 \tag{3.1}$$

where $\mathbb{R}^{(l)}$ is the lth conformation of $\mathbb{R}$ ensemble and $\mathbb{R}'^{(k)}$ is the kth conformation of $\mathbb{R}'$ ensemble. One of the biggest drawbacks of the eRMSD is that in general it is non-zero even when two ensembles are identical which makes the eRMSD difficult to use quantitatively [107] also it is computationally expensive for ensembles with large number of conformations. There are two major approaches to tackle this problem one focus on global phenomena and the other on local phenomena. The first approach usually uses two dimensionality reduction schemes. One scheme takes into account a subset of degrees of freedom for example using center of mass for amino acids to capture rapidly converting microstates. Another scheme uses some variant of principal component analysis (PCA) which discretized conformational space to the most significant collective variables in order to capture slowly converting macrostates [109]. Second approach which focuses on localized differences typically uses mean position displacements, change in fluctuations, contact maps [110] and correlates motions [111, 112].

3.1.1 Summary statistics based approaches

PCA (and its variations) is the most widely used method to infer protein dynamics from ensemble of conformations. It is a multivariate statistical analysis and a projection method to visualize complex data by reducing the dimensionality of a dataset. In PCA a covariance matrix of positional fluctuations is decomposed into a number of principle components (PCs) in order to maximize the variance of the data on each successive PC with orthogonality constraint of each PC on previous PCs [113]. This is accomplished by diagonalizing covariance matrix to obtain orthogonal eigenvectors and corresponding eigenvalues. The first few PCs or eigenvectors usually correspond to collective modes that approximate the functional motions in the protein also known as quasi-harmonic analysis [113, 114]. There are some limitations for using PCA for analysis of ensembles were sampled by MD. Garcia and colleagues showed the distribution of conformations is multimodal for large systems leading to quasi-harmonic assumption breakdown [115]. Clarage and colleagues demonstrated that

low-frequency correlations are under sampled by nanosecond MD simulations [116]. Not only limited sampling of long-range correlations, but also forced orthogonalization of the modes make the global contribution of individual PCA modes problematic [114]. Balsera et al. showed that even though relaxation time of the fast modes are within the MD sampling window, some of them are not recovered by PCA due to their dependency on the slower, under-sampled modes [117]. The forced orthogonalization also may cause problem for breaking symmetry of large-scale macromolecular assemblies [114]. To address these limitations Zhang and colleagues introduce a modified PCA analysis inspired by local feature analysis (LFA) for analysis of protein dynamics [114].

There are approaches that only focus on the local differences of conformational ensembles. To do this, they first calculate some summary statistics in residue level such as mean position of center of mass of amino acid and root mean square fluctuation of (RMSF) of amino acid and then they compare them against each other.

Figure 3.1 illustrates two conformational ensembles of Ser. The differences between two of their summary statistics, CoMs and RMSFs, are negligible. While such a traditional comparison would suggest they are similar, a visual inspection, however, indicates that they are not, $\mathbb{R}$ contains one rotameric form of the side chain, and $\mathbb{R}'$ contains two rotameric forms, The problem with summary statistics is that enumeration is done prior to identification of the key feature that distinguishes the ensembles. Certainly, this difference would have been evident if the right set of summary statistics were compared. But how does one identify such appropriate feature sets beforehand? This hurdle can be overcome by comparing ensembles directly against each other, and prior to any dimensional reduction.

3.1.2 Direct comparison of ensembles

Even though the methods with different schemes of dimensionality reductions have shown applicability in many studies, they are prone to biases. The main reason is because of

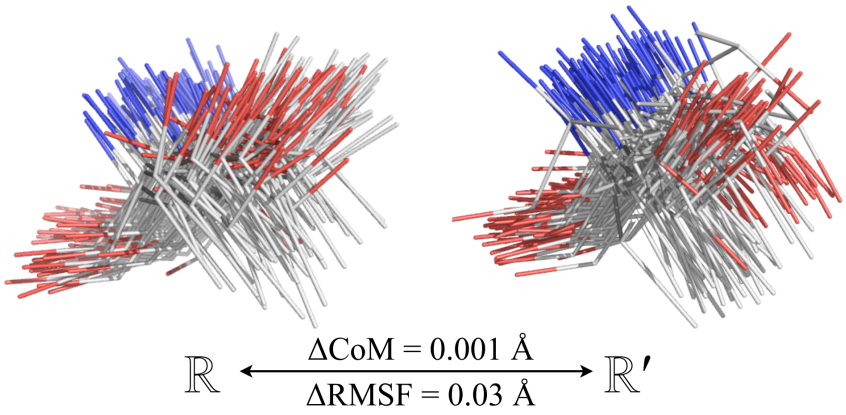

Figure 3.1. Comparison of two conformational ensemble of Ser amino acid by comparing their CoMs and RMSFs

comparing the ensembles after dimensionality reduction, therefore some information is left out or filtered out before comparison. There are other issues with these methods such as need for prior knowledge for each problem in order to use appropriate dimensionality reduction scheme based on defining features of protein intrinsic motion. For example in cases such as protein folding in which protein undergo large structural change, it can be assumed that changes in fluctuations has minor effect and one safely disregard these changes [34]. Even the more recent method based on asymmetric Kullback-Leibler divergence of information theory on internal coordinates or dihedral angles suffer from similar issues [109]. Using dihedral angles which not only reduces the dimensionality of the conformational space but also prefers some degrees of freedom and modes of motions over the others. Lindorff-Larson and Ferkinghoff-Borg used symmetrized version of Kullback-Leibler divergence which is Jenson-Shannon divergence [107]. In both of these approaches each ensemble first is estimated by a probability density function (PDF) then the differences between these PDFs were estimated by an information theory measure. As we explained before using Gaussian distributions for PDF estimation is not always an accurate estimation.

The direct comparison of two ensemble is possible by combining two ensembles into one ensemble and using the appropriate form of PCA. This method can quantify the variations

between two ensembles, but extending this approach to cross-comparison of multiple ensembles is not straightforward.

3.2 Development of a new method

A proper quantification of changes in molecular motions requires simultaneous consideration of all modes of motions. To achieve this goal a method has been developed by using a well-known classifier of the machine learning field called support vector machine (SVM) [34]. This method defines a true metric upon a capability of SVM to separate two overlapping ensembles, instead of using SVM as a classifier. This metric quantifies the physical overlap of two distributions.

3.2.1 Support vector machine

SVMs are traditionally used for predicting binary classification of data [118, 119, 120, 121]. A SVM is first trained on a set of instances $(x_1, x_2, ...)$ with known group identities $(y_1, y_2, ... = \{-1, +1\})$ and then the trained SVM is used for predicting the group identity of an unclassified instance. A SVM can also be constructed when the instances are $3N$-dimensional molecular conformations $(\mathbf{r})$ and belong to two ensembles, $f = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_m\}$ and $g = \{\mathbf{r}_{m+1}, \mathbf{r}_{m+2}, ..., \mathbf{r}_{2m}\}$ [122]. This method utilize the properties of the classification function generated during SVM training to obtain physically meaningful estimate for differences between the conformational ensembles f and g. Our next method which is explained in chapter 5 uses this mathematical framework for repartitioning the ensembles f and g to obtain the subsets f^* and g^*.

The training of the SVM is setup as a Lagrange optimization problem, where the goal is to maximize the linear separation between the two groups in some Hilbert space (see Figure 3.2).

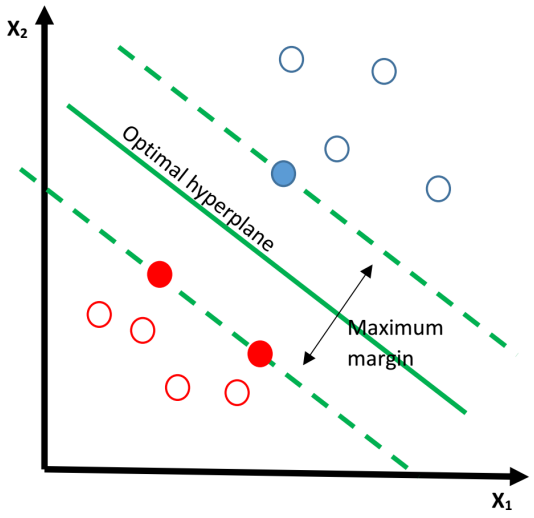

Figure 3.2. The svm algorithm results a hyperplane which maximizes the margin

Essentially, two hyperplanes

$$y_i(\mathbf{w} \cdot \mathbf{r} - b) = 1, \tag{3.2}$$

with $y_i = \pm 1$ are sought that are constructed from a subset of the instances such that the distance $2/\|\mathbf{w}\|$ between the hyperplanes is maximized. This distance is maximized by minimizing

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{k=1}^{2m} \alpha_k [y_k(\mathbf{w} \cdot \mathbf{r}_k - b) - 1] \tag{3.3}$$

with respect to $\|\mathbf{w}\|$ and b, and maximizing it with respect to the Lagrange multipliers α_k. Note that the square on $\|\mathbf{w}\|$ permits quadratic optimization and the $1/2$ coefficient is introduced for mathematical convenience. Substituting the conditions

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum^{2m} \alpha_k y_k \mathbf{r}_k \tag{3.4}$$

and

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum^{2m} \alpha_k y_k = 0. \tag{3.5}$$

into 3.3 rearranges the auxiliary function to

$$L = \sum^{2m} \alpha_k - \frac{1}{2} \sum_{k,l}^{2m} \alpha_k \alpha_l y_k y_l K(\mathbf{r}_k, \mathbf{r}_l), \tag{3.6}$$

where $K(\mathbf{r}_k, \mathbf{r}_l) = \mathbf{r}_k \cdot \mathbf{r}_l$, which is then maximized under the constraint $\sum \alpha_k y_k = 0 \; \forall \; k$. An additional box constraint is introduced, $0 \leq \alpha_k \leq C$, in which C serves an upper limit on the magnitudes of the Lagrange multipliers.

Note that in the optimization of 3.6, the feature used for classifying $\mathbf{r}$ is its linear projection on other $\mathbf{r}$. The vectors $\mathbf{r}$, however, are generally not linearly separable in the Euclidean space when they represent molecular conformations. Such issues are dealt with by choosing alternative kernels that are, by themselves, inner products in the transformed Hilbert space, [119, 120, 121] that is, $K(\mathbf{r}_k, \mathbf{r}_l) = \langle \phi(\mathbf{r}_k), \phi(\mathbf{r}_l) \rangle$. The primary advantage of such "kernel-tricks" are that they bypass the need to determine the explicit form of the function $\phi(\mathbf{r})$ that transforms the data from the original space to the desired Hilbert space to make the data linearly separable [34].

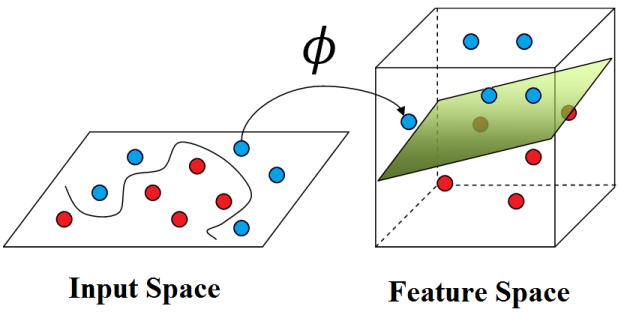

Figure 3.3. Kernel trick or mapping the data into a Hilbert space where the data are linearly separable.

The optimization of 3.6 produces Lagrange multipliers that are ultimately plugged into the binary classifier, $F(\mathbf{r}) = \sum_{i=1}^{2m} \alpha_k y_k K(\mathbf{r}_k, \mathbf{r})$, to predict the group identities of previously unseen data $\mathbf{r}$. More importantly, we note that the optimization of 3.6 produces two different

sets of Lagrange multipliers, $\{\alpha_k\} = 0$ and $\{\alpha_k\} > 0$. The $\mathbf{r}_k$ whose corresponding $\alpha_k > 0$ essentially define the maximum margin hyperplanes that are sought in 3.2, and these $\mathbf{r}_k$ are referred to as support vectors. By choosing appropriate kernel function and tuning hyper-parameters on variety of distributions which is explained in the following, total number of generated support vectors s, where $2 \leq s < m$, can be used as a quantitative estimate of the normalized overlap between the two distributions, that is,

$$Overlap = \|f \cap g\| = s/2m. \tag{3.7}$$

The difference between the two ensembles can then be quantified in terms of a normalized metric $\eta \in [0, 1)$,

$$\eta^{f \leftrightarrow g} = 1 - \|f \cap g\|, \tag{3.8}$$

which takes up a value closer to unity as the difference between ensembles increases.

3.2.2 Tunning hyper-parameters

We choose a Gaussian radial distribution function as the kernel due to its stationary and performance in classification comparing to linear, polynomial, or sigmoidal kernel [34, 119], that is,

$$K(\mathbf{r}_k, \mathbf{r}_l) = \exp(-\gamma \|\mathbf{r}_k - \mathbf{r}_l\|^2), \tag{3.9}$$

then the parameter γ controls the width of the kernel and thereby the smoothness of the underling nonlinear classifier. The interpretation of its effect on molecular conformation is the influence a given conformation has on its neighboring conformation. Smaller γ corresponds to larger Gaussian widths or larger contribution of molecular rearrangements on classification [34]. The box constraint C controls the complexity of the whole model. These two parameters together define the Hilbert space and, can be optimized to yield overlaps with high accuracy.

For tuning these hyper parameters we need to have an estimate on the typical atomic fluctuations on MD simulations. To acquire this estimate we consider several MD simulations of the test case model system which is explained in next chapter. Based on the results a single-particle Gaussian ensemble is generated with $\mu = 0$ and $\sigma = 0.5$ then, two other sets of distributions were generated by changing the mean and standard deviation of the Gaussian function. In one set the mean of the Gaussian varied in unit increment of $\Delta\mu/\sigma_0 = \{1, 2, ..., 20\}$ and in the second set standard deviation varied in unit increment of ratio $\sigma/\sigma_0 = \{2, 3, ..., 15\}$. In the context of protein motion these two sets correspond to changes in mean position displacements and fluctuations, we tried to go beyond typical distributions in molecular simulations. To find the best combination of C and γ which minimizes the mean absolute error (MAE) between overlaps (analytical and SVM estimated) we used a grid search scheme for $C \in [1, 10^8]$ and $\gamma \in [10^{-3}, 10^8]$ We found $C = 100$ and $\gamma = 0.4$ minimized the MAE ≤ 2.5 and results of this comparison for two sets of distribution is depicted in Figure 3.4.

3.2.3 Testing the method and comparison with similar approaches

To test the generalization power and robustness of the method we generated another 300 Gaussian distributions with changing the mean and standard deviation simultaneously were not used in parameter tuning step. The comparison of the estimated overlap and analytical overlaps for two different widely used SVM implementations, svmLight [123] and LIBSVM [124] is shown in Figure 3.5. LIBSVM not only show better more accurate results comparing to svmLight with MAE of 3.2 for all 300 distributions.

In order to compare the performance of the method with other similar approaches for quantifying the differences between two distributions we used 5 different widely-used class separability measures. These measure are:

Absolute value two-sample t-test with pooled variance estimate (ttest) [125].

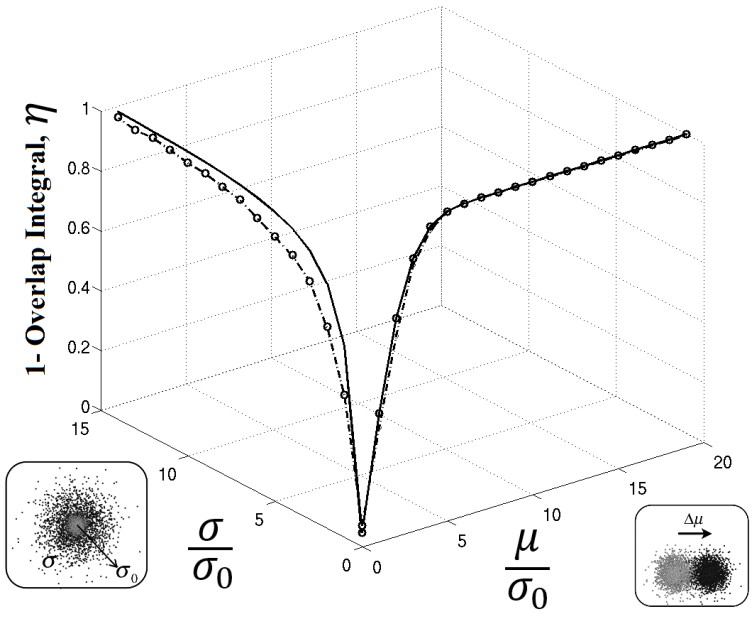

Figure 3.4. The results of optimized svm to estate the discriminability (overlap complement) of two distributions solid line is the analytical estimates and data points are svm outputs. Tow distributions have different fluctuation width in left side of the image and inset illustrates two ensembles with ratio of $\frac{\sigma}{\sigma_0} = 3$. Where distributions on the right hand side are different in mean position. The inset on right shows two distribution with difference of $\frac{\Delta\mu}{\sigma_0} = 4$

.

Relative entropy, also known as Kullback-Leibler distance or divergence (entropy) [81].

Minimum attainable classification error or Chernoff bound (bhattacharyya) [126].

Area between the empirical receiver operating characteristic (roc) [127].

Absolute value of the standardized u-statistic of a two-sample unpaired Wilcoxon test, also known as Mann-Whitney (Wilcoxon) [128].

In this approach we used combinations of these class separabilities in two steps total of 25 combinations. In the first step we calculate the separability measure for distributions of each coordinate and weight the distributions based on that and in the second step we calculate separability measure for normalized weighted distributions. Figure 3.6 illustrates the performance of five of these combinations. Only by visual inspection one can realize

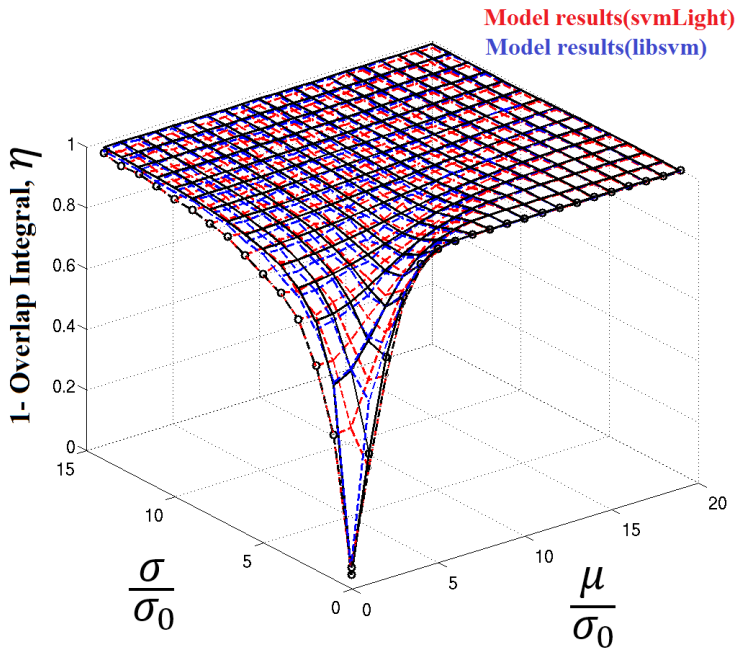

Figure 3.5. The results of testing svm versus analytically estimated discriminability on distributions that were not part of training svm. 300 distributions with simultaneous change in mean position and fluctuation width. The figure also shows relative accuracy of two widely used svm codes. Continuous lines are analytical results svmLight results in red and LIBSVM in blue.

that the most accurate estimate of separability comparing to analytical results amont all 25 combinations belong to using Wilcoxon in the first step and bhattacharyya in the second step. By comparing its accuracy with the accuracy of the new method in figure 3.5 it is evident that even this combination is much less accurate in estimating separability of two distributions. Therefore, even though using SVM for quantifying differences between two distributions is computationally much more expensive but it is necessary for the desired accuracy.

3.2.4 Multi-modal distributions

Assumption of Gaussianity for distributions of particles as a result of central limit theorem is not always valid especially for systems such as proteins with numerous many-body

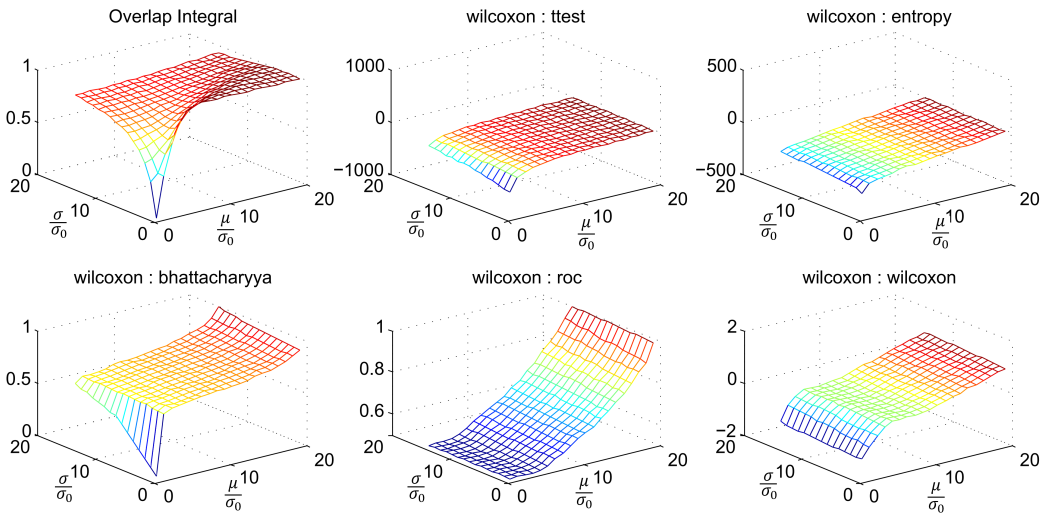

Figure 3.6. Testing results of using conventional class separability measure versus analytically estimated discriminability (top left corner). Among all 25 combinations of ttest, entropy, bhattacharyya, Wilcoxon, Wilcoxon:bhattacharyya showed closest agreement to analytical separabilities.

interactions [37]. However, from theoretical standpoint the overlap of two multi-Gaussian distributions $\mathbb{R} = \sum c_i f_i$ and $\mathbb{R}' = \sum c'_i f'_i$, where f_i are Gaussian and c_i are weights, is essentially a of overlaps between Gaussian distributions, that is,

$$\eta = 1 - ||\sum_{i=1}^{n} c_i f_i \bigcap \sum_{j=1}^{n} c'_j f'_j|| = 1 - ||\sum_{i,j=1}^{n} c_i f_i \bigcap c'_j f'_j|| \tag{3.10}$$

Therefore, the method should work for multi-Gaussian distributions. Figure 3.7 shows the performance of the method for computing the overlap between 400 for each bimodal, trimodal and quadrimodal distributions with arbitrary selection of parameters. In each case, MAE is less than 6% and Pearson correlation coefficient is larger than 0.97.

3.2.5 Testing on different coordinates

For using this method to quantify the differences between conformational ensembles of proteins one can use different methods for generating conformational ensembles. Some of

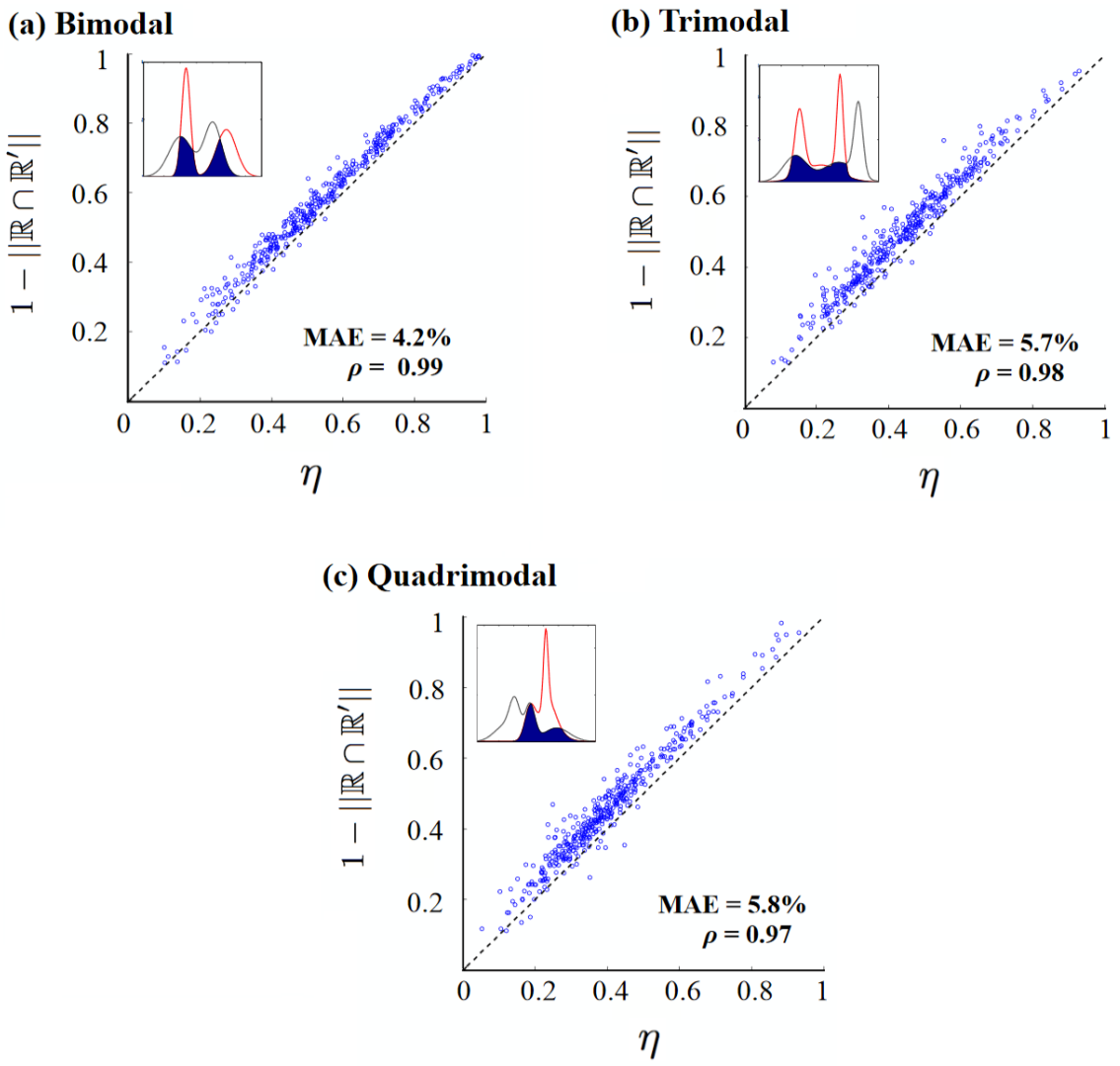

Figure 3.7. The correlation between analytically calculated disciminabilities and svm estimates of 400 arbitrary multi-modal distributions. a)bimodal b) trimodal c) quadrimodal (Reprinted with permission form [37])

these methods are Molecular dynamics simulation (MD), Monte Carlo (MC) methods, Simulated annealing (SA), Essential dynamics PCA-ED methods and Hybrid quantum mechanical/molecular mechanical (QM/MM) methods [129]. Cartesian coordinates of each atom were used as the inputs of the SVM classifier. Figure 3.8 shows the efficient indexing scheme (3DArray: Coordinates, Atoms, Frames) that we used for the algorithm.

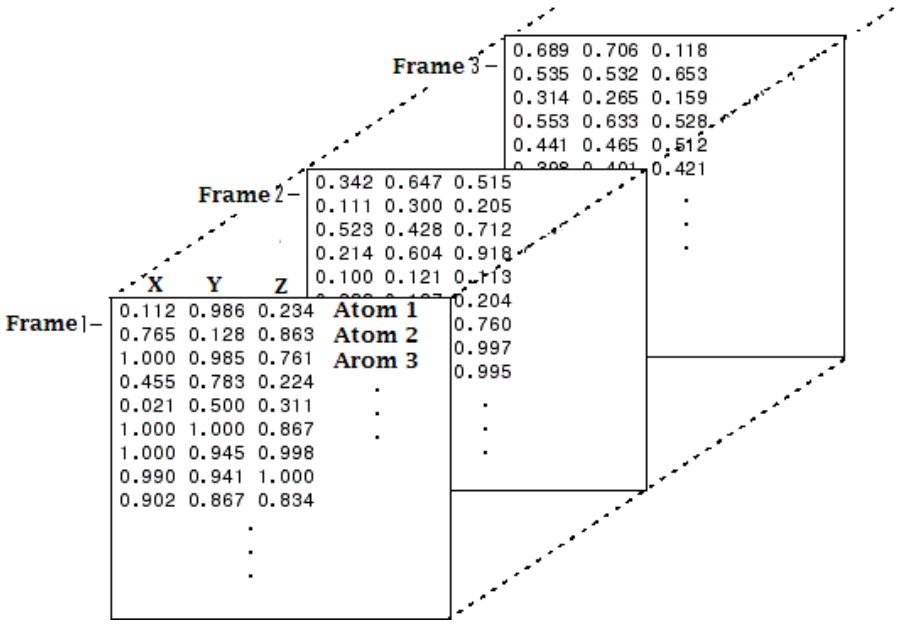

Figure 3.8. Indexing scheme that is used for the method. Cartesian coordinates of each atom in different frames

Since method showed the robustness on quantifying differences between multi-modal and more complex distributions it can be used for using other features as inputs also. These features are collective variables or internal coordinates. Dihedral angles are one type of internal coordinates and using them as features for comparisons have some benefits. Dihedral angles can provide information about certain degrees of freedom explicitly. They are responsible for most low-frequency motions. These low-frequency motions are related to bond rotations and correlated changes in side-chain rotamers. These motions are highly anharmonic type of correlation which are tightly correlated to function of proteins and specially play a key role in allosteric transitions [112]. Figure. 3.9 depicts these changes.

One the other hand since dihedral angles are internal coordinates, which are independent of actual position of atoms therefore are not sensitive to protein displacement and rotations. Using internal coordinates can remove potentially spurious correlations that can rise due to standard structural alignments. In standard structural alignments minimization of the RMSD error in structural alignments in Cartesian space can yield correlated displacements in

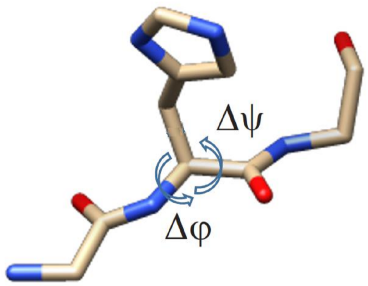

Figure 3.9. Using dihedral angles or internal coordinates as the input which are independent of actual positions of atoms. They are responsible for many low-frequency motions such as bond rotations.

many atoms' positions as some atoms are fit better that others [22, 112, 130]. However these errors are not large for systems with dynamic allostery which undergoes small conformational changes. For example we used dihedral angles in similar manner that we used Cartesian coordinates as input features for a case study that is explained in next chapter. The title of the study is determination of intersecting pathways the results showed very close agreement with the one with Cartesian coordinates. We expect that this is because of small structural changes of the system by ligand binding.

3.2.6 Source code and dissemination

After testing the method on different sets of distributions and real test cases which are explained in the following chapters. To improve the accuracy and performance of the method we used LIBSVM package which showed more accurate estimation with less computation time. We integrated multiple code for this analysis into a single standalone code with a new hashing algorithm. The new code is about 100 times faster and uses Gromacs APIs [131]. Moreover, it has the parallel processing capability and is available at $https://simtk.org/home/conf\_ensembles$ figure 3.10 shows it web page.

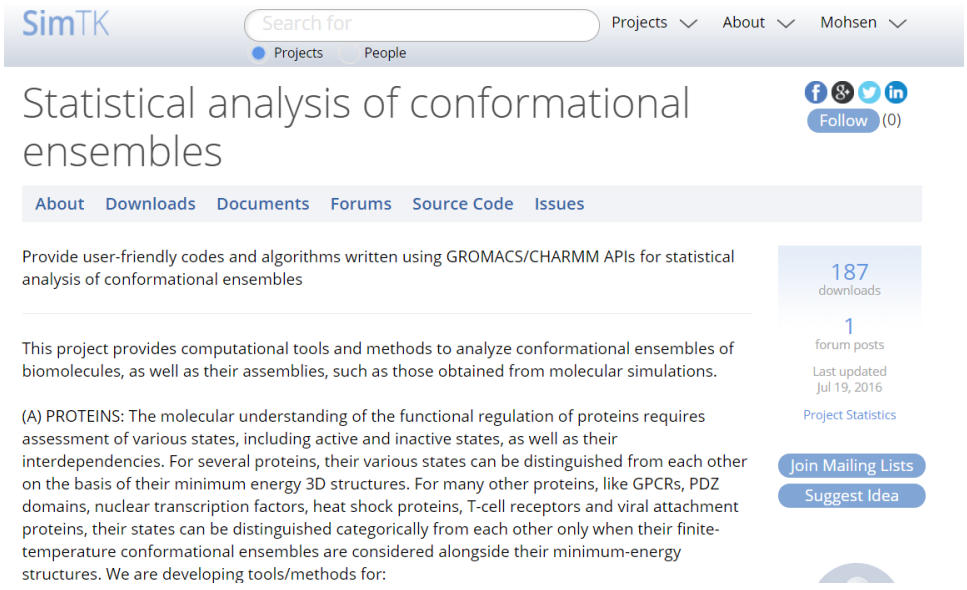

Figure 3.10. The dissemination of code for quantification of ensemble differences at SimTK website.

CHAPTER 4

APPLICATIONS OF NEW METHOD FOR ENSEMBLE COMPARISON

Intrinsic motion of the proteins around its native structure plays role in function of the protein and is affected by many biological stimuli. A quantitative characterization of these intrinsic motions is important because it provides a basis for relating the effects of biological stimuli to function of proteins and as a result biological processes. A new method has been developed [34] by using a machine method that is explained in chapter 3 and it showed robustness in quantifying differences for different distributions. To test the performance of the new method on protein conformational ensembles especially when the effects of the stimuli lead to negligible structural changes we employed the new method for four different applications. First two applications compare two ensembles where the latter two are cross-comparison of multiple conformational ensembles.

4.1 Two-ensemble comparison I: Ranking residues based on their extent of changes

Nipah virus belogs to paramyxoviruses family that are zoonotic pathogens and can cause illness and fatality in domestic animals and human [132, 133, 134, 135, 136]. The binding of G protein of this virus (NiV-G) to the host cell triggers changes in it that ultimately activate the viral fusion protein. Crystal structures of the NiV-G protein shows minor changes in backbone due to the binding to Ephrin B2 receptor (NiV-G preferred host cell protein) [137]. The root-mean-square deviation (RMSD) between apo G and bound G is 1.9 A and most of the backbone rearrangements occur on certain loops near the binding site [137]. Microsec-

ond MD simulation model also suggest similar minor rearrangement of between backbone NiV-G by binding to Ephrin B2. These suggest that other modes of motion including changes in backbone and side-chain orientation, as well as amino acid fluctuation contribute in signal transduction. The understanding of signal transduction mechanism depend on a proper quantitative assessment of all modes of motion simultaneously. This requires direct comparison, without dimensionality reduction of ensembles representing NiV-G motions in both bound and unbound states [34]. Figure 7.1 shows the schematic representation of implementation of new developed method which is capable to do the desired assessment. The results are normalized quantitative estimates of differences between two ensembles at residue level. Figure also shows the mapping of rank-ordered of these estimate to 3D structure of protein. The amino acids that undergo high changes in motion, top 25% not only include amino acids that are directly involved in NiV-G, Ephrin interface but also include contiguously region from interface to residues over 2 nm away from interface. These residues could be part of allosteric pathway of NiV-G binding signal to the viral fusion protein NiV-F [132, 134, 135, 136, 138, 139]

Recent mutagenesis study investigate the effect of two adjacent stretches of amino acids I203-G211 and N195-L202 that belong to the same loop and are showed in Figure 4.2. While the first one showed crucial effect the second showed minor role in fusion [140]. The estimates of the method showed similar results high changes in motion for the former stretch and small changes of the latter stretch. The figure also shows intrinsic changes not just comprise of backbone displacement but also change in side-chain orientation and fluctuations [34].

4.2 Corss-comparison of multiple ensembles I: Force field comparison

The crystal structure of the ephrin-bound of NiV-G shows one of the highest number of water molecules among protein-protein interfaces [35]. MD simulation also indicate that this extensive interstitial region accommodate large number of water molecules. Moreover, while

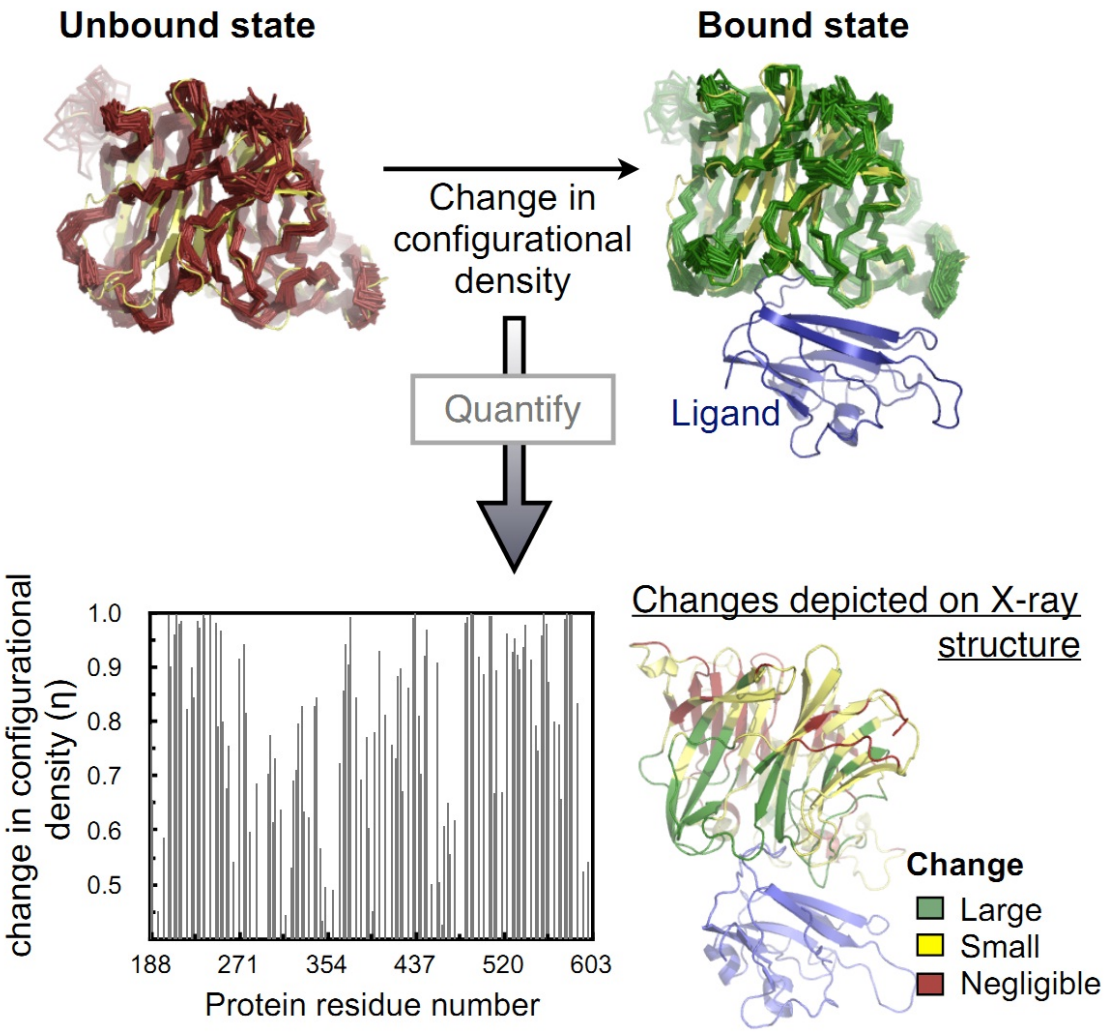

Figure 4.1. On top twenty representative structures of NiV-G superimposed on X-ray structure in yellow. On bottom quantification of ensemble changes on residue level due to binding of a ligand (Adapted with permission form [34]).

water molecules in MD simulation tend to occupy crystallographic sites, most of them have residence time of hundred picosecond(see Figure 4.3). But do they play a physiological role in viral fusion? The Nipah fusion protein (NiV-F) plays the major role in viral fusion. NiV-G binds to ephrin and sends the signal allosterically to NiV-F by changes in conformational density [34].

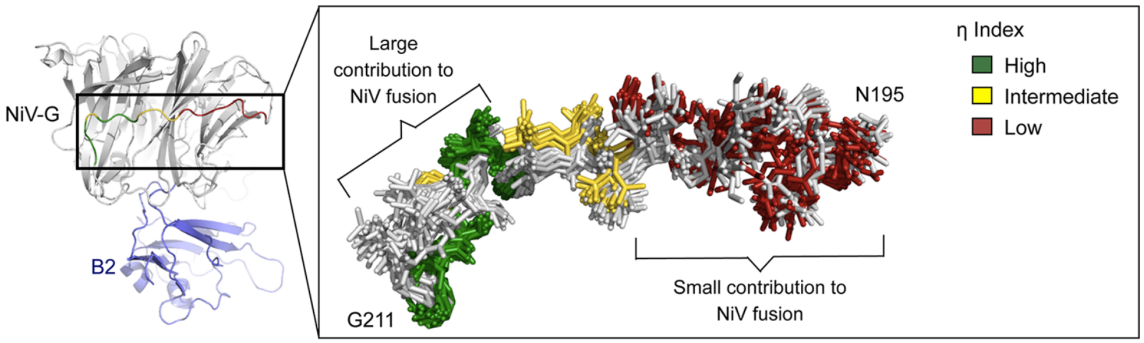

Figure 4.2. Effect of Ephrin-B2 binding on the intrinsic motion of a specific loop of NiV-G, NQILKPKLISYTLPVVG, and its relationship with alanine-scanning mutagenesis experiments.29 Twenty representative configurations of the segment, ten each from the MD simulation of NiV-G in its phrin-bound and unbound states, are shown superimposed on each other. While the ten configurations from the simulation of NiV-G in its unbound states are colored gray, the ten configurations of NiV-G in its Ephrin-bound state are color-coded according to their discriminability index. We find an exact correspondence between the portions of the loop that have a high discriminability index, that is, those that undergo a high change in intrinsic motion, and those that were shown from experiments to contribute significantly to viral fusion(reprinted with permission from [34]).

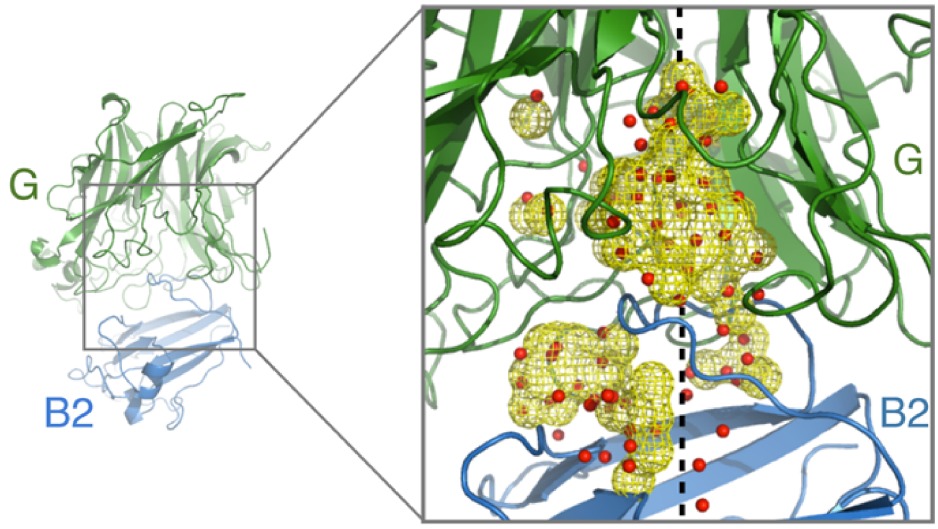

Figure 4.3. Illustration of correspondence of 65 highly occupied regions by water molecules during MD simulation (yellow mesh) and interstitial waters resolved in X-ray structure.(adapted with permission from [35].

If $\mathbb{R}_{apo}$ and $\mathbb{R}_{bound}$ represent conformational densities of G protein in apo and ephrin-bound states respectively, then the changes in conformational densities is $\Delta\mathbb{R} := \mathbb{R}_{apo} \rightarrow \mathbb{R}_{bound}$. Consequently, if water molecules at interface of G-ephrin plays a role in allostery F-activating signal they must contribute to $\Delta\mathbb{R}$. To answer this question we determined $\Delta\mathbb{R}$ in two different condition when we model water molecule effect in MD explicitly and when use implicit water models [35]. Implicit solvent models do not consider discrete nature of water molecules. This not only changes the protein-protein interface volume due to the absence of water molecules. Also since 10% of the water molecules make a network of hydrogen-bond interactions with two proteins and each other the absence of discrete water molecules can alter $\Delta\mathbb{R}$. Figure 4.4.a shows the comparison of $\Delta\mathbb{R}$ of implicit and explicit solvent simulations. In this figure dots show the discriminability or $\eta \in [0, 1)$ for each residue where higher number indicates higher difference between ensembles. The estimated $\Delta\mathbb{R}$ from using implicit and explicit solvent models are statistically different with Pearson correlation of 0.28. This divergence is even larger for amino acids that are part of allosteric pathway [36] dots colored red. Even though this shows the effect of absence of interstitial water with treating the solvent with implicit models, it does not delineate specific role of them. For further investigation on specific contribution a subset of residue in the G protein is identified that their conformational densities in the apo are unaffected by the treatment of the bulk solvent. To do this we compute $\eta_{imp\leftrightarrow exp}$ which is the ensemble difference for residues of apo state when they have simulated in implicit versus explicit solvent. The residues with $\eta_{imp\leftrightarrow exp}$ smaller than a specific tolerance are considered unaffected by treatment of the bulk solvent. We choose $d^2 = BT/8\pi^2 T_{xray}$ as the tolerance, the mean square deviation of a residue obtained from crystallographic B factor [141]. The T/T_{xray} ration is for rescaling the B factor from X-ray temperature to T_{xray}=100 K to physiological temperature T=310 K [35]. Therefore, for a given residue if $\eta_{imp\leftrightarrow exp} < erf(d/\sqrt{2})$ then the estimated difference between ensembles generated in implicit and explicit solvent simulation is smaller than the spread of the residues

electron density observed in X-ray diffraction. The error function is used to transform the tolerance to the appropriate Hilbert space where eta is estimated. The subset of G residues that meet this condition comprise 114 out of 416 residues. Figure 4.4.b shows even for these residues estimated conformational density shift of binding is statistically different. Since the dynamics of these residues were not affected by implicit solvent treatment, this difference reflects the specific effect of this treatment on G-B2 interactions.

4.3 Corss-comparison of multiple ensembles II: Determination of intersecting allosteric pathways

As mentioned earlier NiV-G by binding to ephrin of the host cell sends a signal to activate NiV-F. Moreover, structural difference between apo and ephein-bound states is minor. Therefore, for further analysis of the allosteric signal we require to quantitatively compare the conformational ensemble of these states since as a result of ensemble/thermodynamic view $\Delta\mathbb{R}_{signal} \subset \Delta\mathbb{R}$. In this study we quantify the $\Delta\mathbb{R}$ induced by ephrins, B2, B3 and a well characterize mutant of B2 [142]. The sequence similarity of B2 and b3 is about 50% and B2 mutant differs in two residue identities, L281Y and W282M. This mutant which we refer to as B2m binds to G protein weakly compared to B2 and B3 but still triggers viral fusion [142]. In previous study we quantified the ensemble difference or $\Delta\mathbb{R}$ induced by binding of ephrin B2 [34]. However, since there is no formal relationship between allostery pathway and extent of $\Delta\mathbb{R}$, quantitative analysis of $\Delta\mathbb{R}$ does not provide basis to label a subset of changes as allostery signal. To further investigate the allostety pathway we first generate 2 other ensembles when G bound to B3 and B2m, then analyzed the changes in ensemble due to binding of G to three ephrins collectively. In other words if there is a common allostery pathway for transducing the binding signal it should be a subset of $\Delta\mathbb{R}_{int}$ which is defined as follow

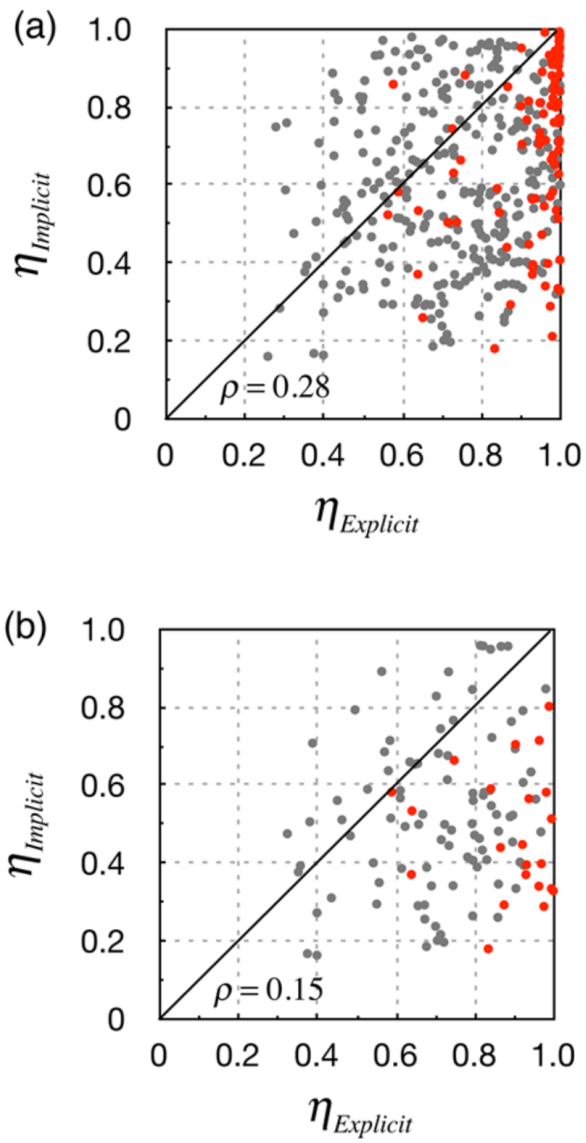

Figure 4.4. Correlation between B2-induced conformational density shifts simulated in explicit versus implicit solvent. a) The 416 dots represent the estimated value for G residues and those are in red are those that are part of allosteric signaling pathway [36]. b) The 114 dots represent residues that their conformational density shits are negligible considering their X-ray B factors.(Reprinted with permission from [35])

$$\Delta\mathbb{R}_{int} := \Delta\mathbb{R}_{B2} \cap \Delta\mathbb{R}_{B3} \cap \Delta\mathbb{R}_{B2m} \tag{4.1}$$

where $\Delta\mathbb{R}_{B2}$, $\Delta\mathbb{R}_{B3}$ and $\Delta\mathbb{R}_{B2m}$ are the changes in G conformational density induced by binding to B2, B3 and B2m respectively. Figure x schematically shows all of the ensemble comparisons we refer the apo state of G protein as G() and ephrinX-bound state as G(X) consequently η_{x1} refers to discriminability of G() and G(X1) and $\eta_{x1/x2}$ discriminability of G(X1) and G(X2).

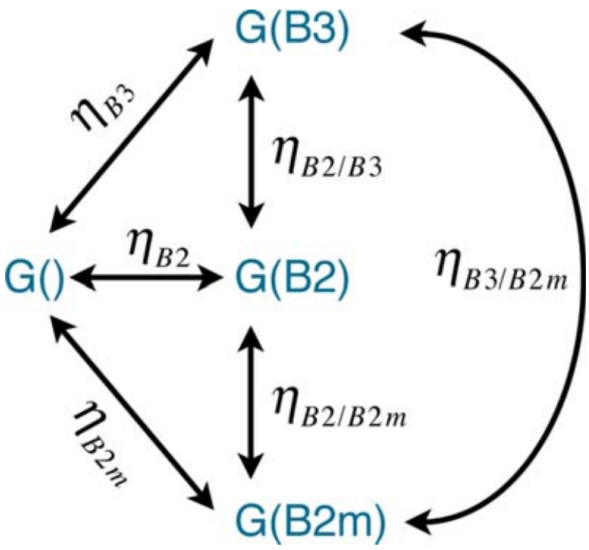

Figure 4.5. Schematic representation of different conformational density shift analysis. G() represents free ensemble where G(X) represents X-bound G ensemble. Therefore, η_{X1} is shift between bound and X1 bound where $\eta_{X1/X2}$ is shift between G(X1) and G(X2) (Reprinted with permission from [36])

To determine whether two ephrins X1 and X2 induce similar changes in conformational density of a residue in G we apply the following two criteria,

$$min\left\{\eta_{X1}, \eta_{X2}\right\} > \eta_{X1/X2} \tag{4.2}$$

and

$$|\eta_{X1} - \eta_{X2}| < \langle \eta_{X1} - \eta_{X2} \rangle \tag{4.3}$$

The first criterion ensures that the $\Delta\mathbb{R}$ induce by two ephrins are both are greater than the $\Delta\mathbb{R}$ between to bound states. The second criterion assign a tolerance to difference between the $\Delta\mathbb{R}$ of two ephrins which in this case is mean absolute difference (MAD). The second criteria applied after applying the first criterion. The advantage of using such a dynamics cutoffs is the there is no need for preexisting relationship between allosteric signal and the extent of $\Delta\mathbb{R}$. The results of this statistical analysis is showed in Figure 4.6. One of the surprising results of this analysis is that mutation of just 2 residue in ephrin B2 changes the conformational density of almost half of the residues of G protein. This simply highlights that the primary sequence of ephrin is not correlated to the extent/nature of conformational density changes it induces. As a results of the statistical analysis only 106 residues belong to intersecting pathway that their conformational densities are modified statistically equivalently by the three ephrin binding.

To further analyze what type of changes ephrins induce to conformational ensemble of these residues we calculate the correlations between eta and mean backbone deviations of them. The mean backbone deviation is defined as,

$$d = 1/3 \sum_{X} || \langle r_{CoM} \rangle_{G()} - \langle r_{CoM} \rangle_{G(X)} || \tag{4.4}$$

Where $\langle r_{CoM} \rangle_{G(X)}$ is the average position of center of mass of residue backbone over ensemble G(X). We find that conformational density changes and backbone deviation of residues belong to the intersecting pathway are not correlated perfectly (Pearson correlation coefficient = 0.77). Several residues belong to this set have high changes of conformational entropy and/or side-chain orientation changes. Three such representative case are illustrated in Figure 4.7 which indicate that examining changes in summary statistics such as mean position displacement and change in fluctuation will not provide a complete conformational density changes.

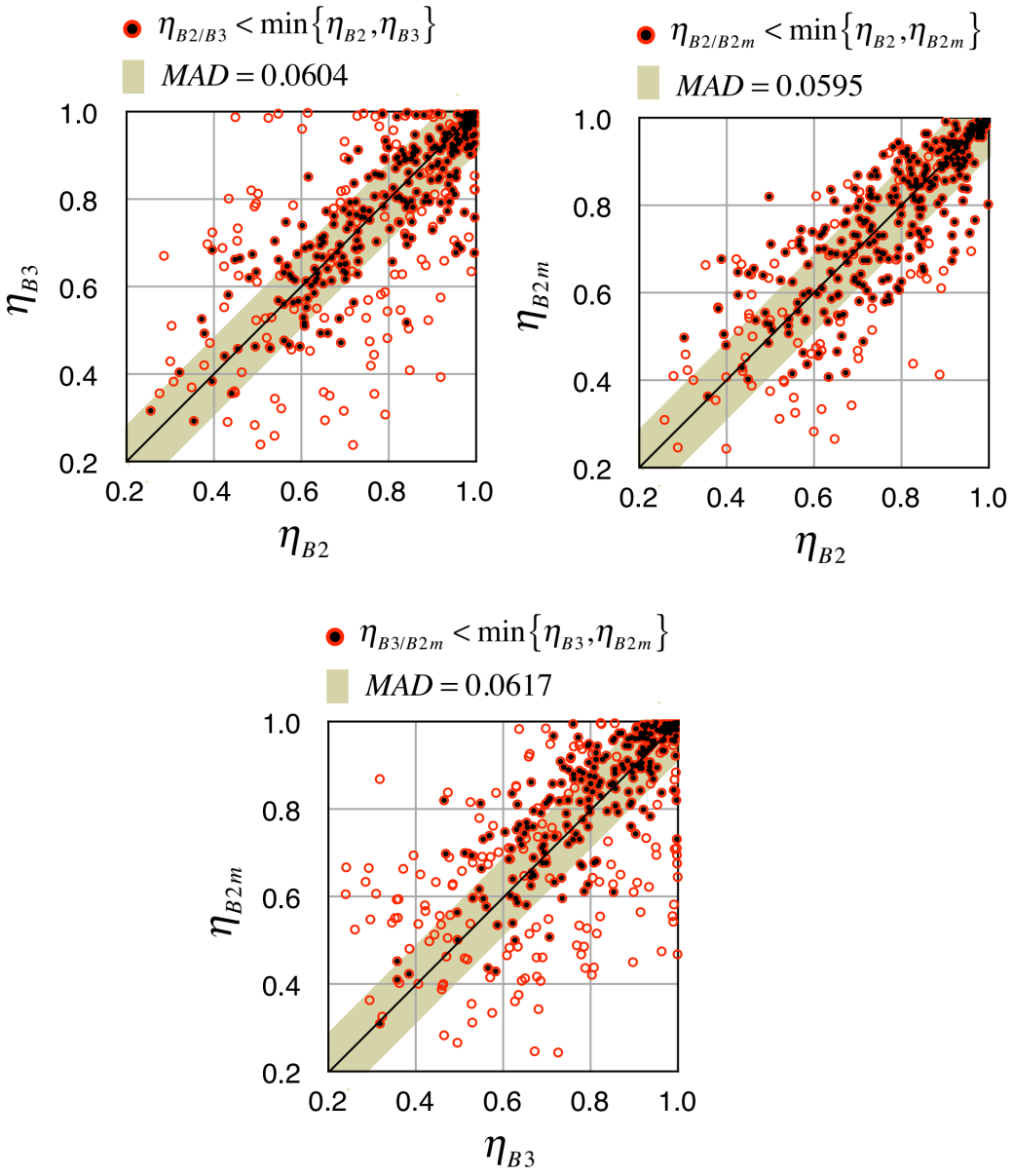

Figure 4.6. Comparison of conformational density shifts of G residues induced by different ephrins. Each plot contains 416 circles which represent residues and filled circles are those that satisfy the condition of equation that is mentioned above each image. These residues where used for MAD calculation which itself is used to find the subset of residues that are shifted statistically equivalently by ephrin X1 and X2(Reprinted with permission form [36]).

Finally the analysis detect 8 out of 14 residues the mutagenesis study showed their effect on F regulation [140]. This suggest that intersecting subset consists of at least one signaling

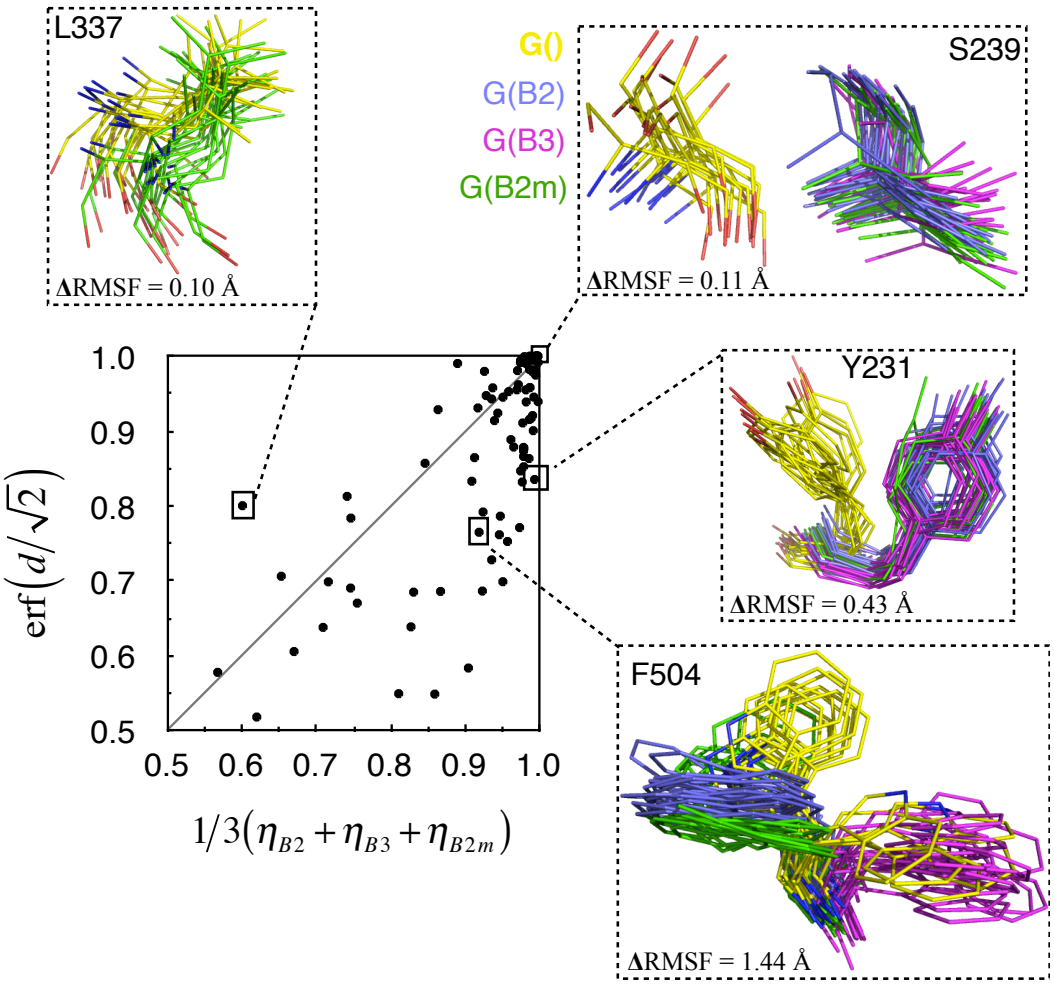

Figure 4.7. Correlation between the conformational density shifts (η) of residues belong to intersecting pathway and their backbone deviation d. The backbone deviation calculated by Eq. 4.4 and transformed to the same Hilbert space where η were calculated by erf(d/$\sqrt{2}$). The figures surrounding the correlation plots showing conformational density of four residues. The ensemble for each residue composed of 15 frames and color coded. The ΔRMSF is the average of differences between root-mean-square fluctuations (RMSFs) of G and G(X)s. The residues such as S239 which are close to the diagonal mostly undergo backbone deviation. Residues such as Y231 and F504 that are below the diagonal undergo side-chain rotation and/or changes in fluctuations. Finally for residues above the diagonal change in backbone fluctuation is dominant mode of change.

pathway. In addition, it is likely that there exist other signaling pathways unique to one or common to two ephrins.

4.4 Corss-comparison of multiple ensembles III: Effect of mutations on regulation

Experimental studies show that during viral activation-fusion process G interacts with ephrin by its receptor binding domain (RBD) and with F by its t F activation domain (FAD). There is not any model that explains the coupling. Due to structural organization of G the allostric coupling must involve in at least one the two RBD-FAD and/or RBD-RBD interfaces [37]. Our previous MD and ensemble comparison analysis of monomeric RBD suggested that intersecting potentially allosteric pathway of three fusion-inducing ephrins involves RBD-RBD interface [36]. Another experimental study based on cellular assays and monoclonal antibody binding also suggest this pathway [143]. Additionally, rearrangement of RBD-RBD interface in a manner that facilitate solvent-exposure of FAD and following interaction of FAD with F was proposed by experimental approaches [144, 145]. This mechanism is particularly intriguing because it assumes that despite of small structural changes in RBD domain binding of ephrin can induce extensive rearrangement of nonoverlapping RBD-RBD interface. A mutagenesis study showed triple mutation of V209V210G211 -> AAA can disrupt F-activating signal of G without effect on expression as well as binding of G to ephrin [140]. The VVG residues are part of RBD and are distant from both RBD-RBD and RBD-FAD interfaces. For further investigation of the allosteric signaling mechanism we carry out MD simulations of RBD-RBD dimer in free and ephrin-bound state as well as VVG mutant in both states. To find the mutation induced shifts in the allosterc signal we calculate $\Delta\mathbb{R} := \Delta\mathbb{R}_{apo} \rightarrow \Delta\mathbb{R}_{bnd}$ as well as $\Delta\mathbb{R}^m := \Delta\mathbb{R}^m_{apo} \rightarrow \Delta\mathbb{R}^m_{bnd}$ therefore the subset of residues where $\Delta\mathbb{R} \neq \Delta\mathbb{R}^m$ are affected by mutation. These residues should meet at least of the following conditions:

$$
\begin{aligned}
|\eta - \eta^m| &> 2 \times MAE, \\
\eta_{apo} &> erf(1/\sqrt{2}), and \\
\eta_{bnd} &> erf(1/\sqrt{2})
\end{aligned}
\tag{4.5}
$$

Where MAE is the mean absolute error of the system, therefore the first condition ensures that the difference is greater than error of the method. The other two inequalities place a tolerance on the mutation induced ensemble shift in both free and bound states. This tolerance corresponds to 1Åshift of the center of mass of residue. Figure 4.8.a shows the residues that satisfy these conditions which are about 50% of all residues of the RBD. This suggest that mutation has a global effect on the conformational ensemble of RBD. Visualization on the 3D structure depict this spread better (see Figure 4.8.b). This figure also shows conformational ensembles of few other residues, including those near to the RBD-FAD interface. Experiment show that mutation of D468 negatively impact stimulation of G [146]. However, since the ensemble shift because of mutation is negligible, this residue could be only important for protein structural integrity and not necessarily part of the pathway. The conformational ensemble of other five residues but perturbed by mutation which are proximal to RBD-FAD interface. These results suggest VVG mutation disrupt the binding signal via RBD-FAD interface.

We also analyzed the relationship between the extent of the shift and distance from mutation Figure 4.9 shows that there is no relationship between these two.

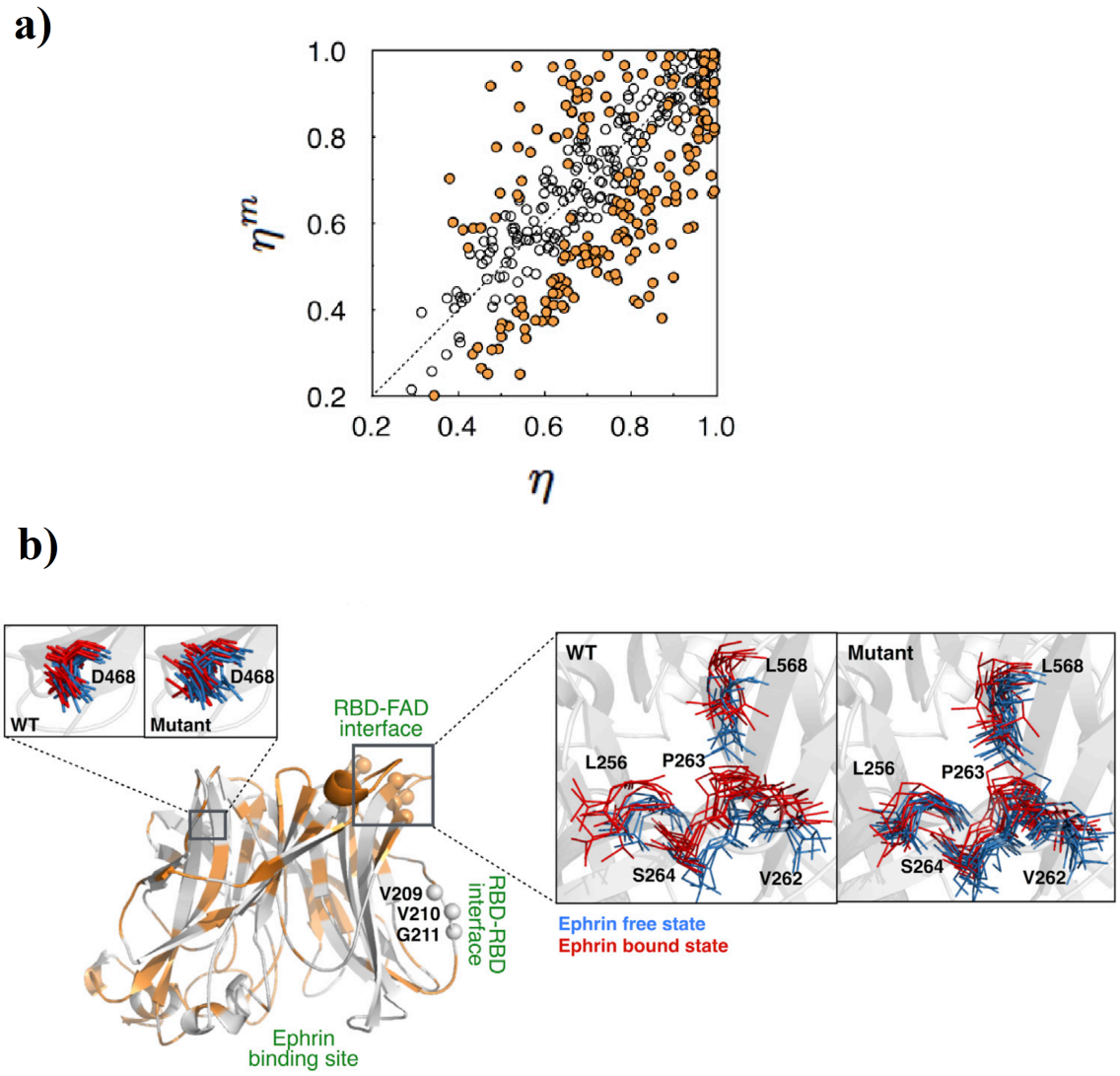

Figure 4.8. a) Comparison of ligand binding induced shifts on wild-type RBD (η) versus similar shift in mutant RBD (η^m). Residues that meet the condition of Eq. 4.5 are highlighted in orange. b) These residues are highlighted on the X-ray structure and ensemble of some of them also were provided including those proximal to the RBD-FAD. (Reprinted with permission from [37])

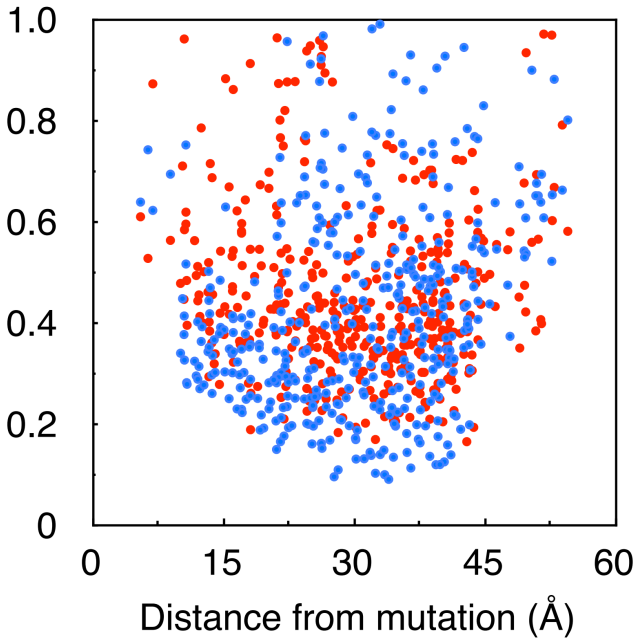

Figure 4.9. The correlation between mutations-induced conformational density shifts and distance from mutation site (Reprinted with permission from [37])

CHAPTER 5

DEVELOPMENT OF NEW METHOD FOR CONNECTING ENSEMBLE SHIFTS TO REGULATION

5.1 Theoretical background and existing method

Methods have been developed to relate inter-state differences to allosteric regulation [21, 22, 23], which have also contributed to design of new customized proteins [6, 93, 94, 4, 95]; however, none account for thermal fluctuations. These methods typically rely on average structural differences between states, which renders them unsuitable for studying proteins in which inter-state differences in structure are comparable to thermal fluctuations; but we note that these methods were not intentionally designed to account for thermal fluctuations.

Methods have also been developed to connect correlations in thermal fluctuations to signaling [24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. Given time-dependent conformations of two sites in a given protein state, f_i and f_j, their fluctuation correlations are determined as

$$C_{ij} = \frac{1}{\sigma_i \sigma_j} \int (f_i - \overline{f_i})(f_j - \overline{f_j}) dt, \tag{5.1}$$

where the bar denotes average, and the σ denote fluctuations in individual sites. These inter-site fluctuation correlations can be combined with each other and with the spatial organization of the protein to yield insight into how different spatial regions communicate with each other (intra-state signaling). However, since no information on divergence from a reference state is incorporated, these approaches cannot theoretically provide insight into regulatory mechanisms.

New methods are required for understanding mechanisms in proteins regulated by dynamic allostery. Toward this end, we and others have recently developed methods [147, 107, 148, 109, 34, 36, 35, 37] to compare conformational ensembles of different states against each other, and obtain inter-state differences in terms of physically meaningful metrics. These methods, in general, overcome the challenge of finding appropriate feature spaces (or summary statistics) that distinguishes ensembles, and provide a comprehensive difference between ensembles that naturally embodies differences in thermal fluctuations. In particular, we have shown that these methods can be used to tease out protein regions affected by regulators and statistically analyze similarities and differences between different states [34, 36, 35, 37]. However these methods, by themselves, do not provide direct insight into regulatory signaling networks as they do not relate induced conformational ensembles changes in one site to another site.

It is, therefore, not surprising that several fundamental biophysical questions in dynamic allostery still remain unanswered. For example, is "dynamic allostery" aptly termed in that regulation occurs due entirely to induced changes in dynamics or do small changes in energy-minimum structures also contribute? In either case, can we define cutoffs in structural changes, such as in center-of-mass (CoM), below which their contributions to regulation are insignificant? Are there relationships between a residue's propensity to contribute to regulation, and its spatial location or hydrophobicity? If a residue contributes significantly to spatial communication within a state (intra-state signaling), then is it justified to assume that it is also important to propagation of regulatory signals? Do stimulator-binding and unbinding responses occur in the same manner? In general, how different are activating signals from deactivating signals?

Addressing such questions requires an understanding of how stimulation at one site of a protein produces conformational ensemble shifts at another site. Theoretically, this requires determination of inter-site correlations in ensemble shifts, that is, for two ensembles, f_i and g_i

of a given site i, it requires us to determine how ensemble shifts in this site, $g_i^* = g_i \setminus (f_i \cap g_i)$, are correlated with ensemble shifts in another site (g_j^*). Mathematically, it requires us to determine

$$C_{ij}^{f \to g} = \frac{1}{\sigma_i^* \sigma_j^*} \int (g_i^* - \overline{f_i})(g_j^* - \overline{f_j}) dt. \tag{5.2}$$

In the equation above, the bar denotes average and σ^* are fluctuations in shifts. Similarly, we can also define inter-site correlations in f's shift with respect to g as

$$C_{ij}^{g \to f} = \frac{1}{\sigma_i^* \sigma_j^*} \int (f_i^* - \overline{g_i})(f_j^* - \overline{g_j}) dt, \tag{5.3}$$

where $f^* = f \setminus (f \cap g)$ represents ensemble shift in f with respect to g. Note that $C_{ij}^{f \to g}$ and $C_{ij}^{g \to f}$ are expected to be identical only if the distributions in f^* and g^* for both residues i and j are symmetric about their interface (5.1). Computation of $C_{ij}^{f \to g}$ and $C_{ij}^{g \to f}$ require that f and g are repartitioned such that conformations corresponding to the overlap region $f \cap g$ are identified and then removed from f and g to get the residuals f^* and g^*, respectively. To our knowledge this is an unresolved problem, and here we develop a machine learning based method to accomplish this high-dimensional repartitioning task, which then enables calculation of $C_{ij}^{f \to g}$ and $C_{ij}^{g \to f}$. These pairwise correlations can be combined with each other and with the spatial organization of sites, just as C_{ij} are combined [30, 26, 28, 29, 149], to discern regulatory signaling networks. Moreover, in this work we implement a new parameter-free version of the graph theory approach to combine pairwise correlations with each other according to the spatial organization of proteins.

5.2 Ensemble repartitioning and inter-site correlations

As the theory of our SVM-based method is explained in chapter 3, the support vectors can be used to estimate the overlap of two distributions. The visualization as an example Figure 5.2a shows the distribution of support vectors in a test case of two partially overlapping

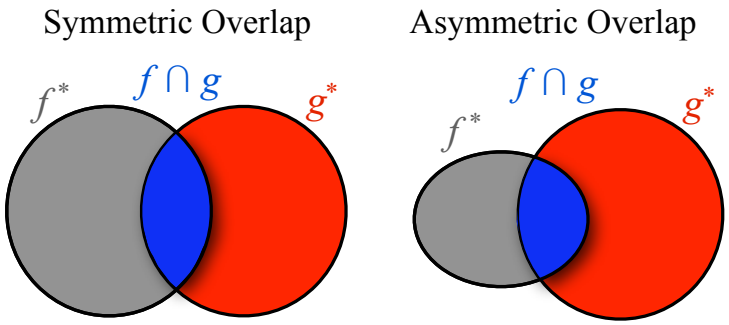

Figure 5.1. Venn diagram of symmetric and asymmetric overlapping distributions. The overlap region $f \cap g$ is shaded blue, the $g^* = g \setminus (f \cap g)$ region is shaded red and the $f^* = f \setminus (f \cap g)$ region is shaded grey.

2D Gaussian distributions. Indeed, we find that the majority of the support vectors are part of the overlap region. Then they can be simply removed from f and g, respectively, to obtain f^* and g^*. However, a fraction of the support vectors do not belong to the overlap region, and instead belong to f^* and g^*. This would imply that the ratio $s/2m$ overestimates the overlap, and consequently the computed η is smaller that the analytical value. This is, in fact, what we noted previously [34, 36, 37] – for almost all of our test cases involving various distribution types (unimodal, bimodal, trimodal and quadrimodal), we found that the computed η are systematically underestimated ($< 6\%$) with respect to exact values.

Now if f^* and g^* were constructed by simply removing the support vectors from f and g, then f^* and g^* would, at worst, suffer from partial omissions of instances. More importantly, f^* and g^* will not be contaminated by instances belonging to $f \cap g$. 5.2b shows the average omission error in 50 random pairs of Gaussian distributions, one of which is shown in 5.2a. We find that as the ensemble size (m) increases, the omission error reduces and for $m \geq 10000$, the average omission error is below 4%, and the worst case error is also below 7%, which are similar to errors we reported earlier [37] in the estimation of η.

If f_i and f_j represent the distributions of two sites in a protein, then their fluctuation correlations C_{ij} are determined as 5.1. When the distribution is discrete and the data are

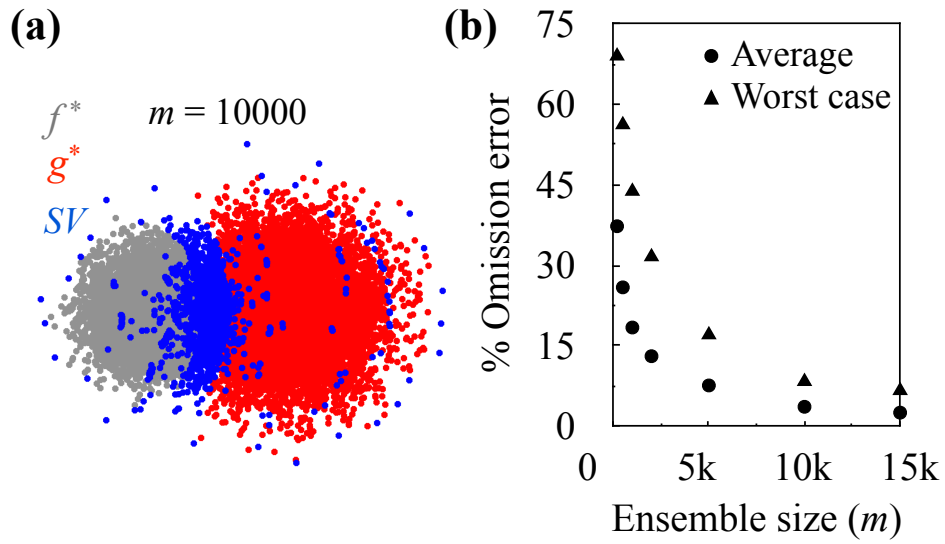

Figure 5.2. (a) Distribution of support vectors (SV) in a representative case of two partially overlapping 2D Gaussian distributions. Each of the two distributions, f and g comprise of $m = 10000$ data points. The remaining instances in the two distributions, f^* and g^* are colored grey and red, respectively. (b) Percent omission error in 50 random pairs of Gaussian distributions. It is computed as a ratio of the number of incorrectly assigned support vectors in the f^* (and g^*) region and the total instances that belong to the f^* (and g^*) region. In other words, Omission error $= FP/(TP+FP)$, where FP and TP are abbreviations for false positives and true positives.

vectors, that is, $f_i = \{\mathbf{f}_{i1}, \mathbf{f}_{i2}, ..., \mathbf{f}_{im}\}$, such as that obtained from molecular simulations, then 5.1 takes the following form:

$$C_{ij} = \frac{1}{\sigma_i \sigma_j} \sum_{k}^{m} \|\mathbf{f}_{ik} - \overline{\mathbf{f}}_i\| \|\mathbf{f}_{jk} - \overline{\mathbf{f}}_j\|, \tag{5.4}$$

where $\|...\|$ denote the magnitudes of vectors. When applied to proteins in the context of constructing signaling networks [28], the vectors $\mathbf{f}_{ik}$ and $\mathbf{f}_{jk}$ are generally the centers of mass of two different amino acids.

If f_i and g_i represent two different distributions of the same site i, but under the influence of different external potentials, and if f_j and g_j represent the corresponding distribution of site j, then inter-site correlations in ensemble shifts $C_{ij}^{f \to g}$ and $C_{ij}^{g \to f}$ are given by 5.2 and

5.3, respectively. Calculation of $C_{ij}^{f\to g}$ and $C_{ij}^{g\to f}$ require that the ensemble data, f and g, are repartitioned such that conformations corresponding to the overlap region $f \cap g$ are identified separately for each site i and then removed from f and g to get the residuals f^* and g^*, respectively. Below we show that such a high-dimensional repartitioning task can be accomplished using the mathematical framework of support vector machines (SVMs). The development below follows from our SVM-based method to compute quantitative estimates for overlaps between conformational ensembles [34], which we also describe briefly for clarity.

The support vectors can, therefore, be used to construct f^* and g^*, and without need for fitting the underlying distributions to assumed mathematical forms. f^* and g^* can be used to determine inter-site correlations in ensemble shifts. For discrete distributions and when the data are vectors, 5.2 takes the following form

$$C_{ij}^{f\to g} = \frac{p_{ij}^{f\to g}}{\sigma_i^* \sigma_j^*} \sum \|\mathbf{g}_{ik}^* - \overline{\mathbf{f}_i}\| \|\mathbf{g}_{jk}^* - \overline{\mathbf{f}_j}\|. \tag{5.5}$$

Note that the summation does not run over all conformations k in g_i^* and g_j^*. Instead it runs only over a subset of protein conformations that are common to both g_i^* and g_j^*. Consequently, we introduce $p_{ij}^{f\to g}$, which denotes the probability of finding protein conformations that are part of both g_i^* and g_j^*. Similarly, 5.3 takes the form

$$C_{ij}^{f\to g} = \frac{p_{ij}^{g\to f}}{\sigma_i^* \sigma_j^*} \sum \|\mathbf{f}_{ik}^* - \overline{\mathbf{g}_i}\| \|\mathbf{f}_{jk}^* - \overline{\mathbf{g}_j}\|. \tag{5.6}$$

Note also that $\mathbf{f}_{ik}$ and $\mathbf{f}_{jk}$ represent $3n_i$ and $3n_j$ dimension vectors, where n_i and n_j are the numbers of atoms in the two amino acids i and j. Consequently, the support vectors that are generated are representative of entire conformations of amino acids. After repartitioning conformational ensembles of amino acids, we then represent $\mathbf{f}_{ik}$ and $\mathbf{f}_{jk}$ by their respective CoMs and compute $C_{ij}^{f\to g}$ and $C_{ij}^{g\to f}$.

5.3 Parameter free network definition

The inter-site correlations described in the previous section are combined with each other and also connected with the spatial organization of sites using undirected weighted graphs $G(V, E)$ [30, 26, 28, 29, 149] comprising of V nodes and E edges that connect the nodes. Nodes on graphs represent points on the proteins that serve as receivers and/or transmitters of information in signaling pathways. Since signal transduction generally needs to be understood at the level of amino acids, nodes on graphs typically represent amino acids [30, 26, 28, 29, 149], and we define them as CoMs of amino acids. From a physical standpoint, direct signal communication is expected to occur between only those nodes whose conformational spaces are directly influenced by each other [30, 26, 28, 29]. To implement this, one typically measures the distance between the CoMs of two nodes, and if that distance is less than a predefined cutoff, which is generally in the range 4-6 Å, [30, 26, 28, 29] then the two nodes are connected by a edge. Otherwise, the two nodes remain unconnected. Instead of using cutoffs, we implement a parameter-free approach that uses the same physical logic, but determines connectivity on the basis of overlap between node volumes – if Γ_i is the volume of the conformational ensemble of node i , then it will be connected to node j only if $\Gamma_i \cap \Gamma_j > 0$.

Edges weights represent a quantity that tells us how nodes communicate with each other. We define edge weights as the inverse of the inter-site correlations.

5.4 Shortest paths analysis

Allosteric signal propagation is an example of information transmission from the binding site into another functional site. On the other hand network communication is dynamic, with altered preferred routes. This alteration of communication routes in different regulatory states is probably leads to a higher efficiency and better control of the transmitted

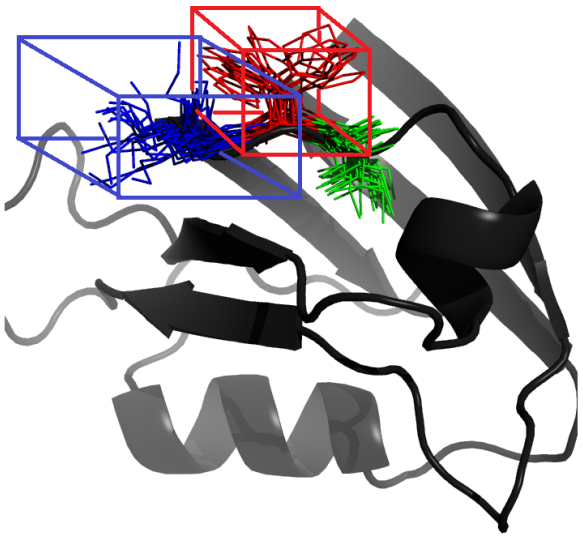

Figure 5.3. The parameter-free definition of neighbors. It considers two nodes connected if the conformational ensemble volume of two residues overlap.

information [29]. Considering this concept Van Wart et al., made a weighted network base on correlated motion on MD model of imidazole glycerol phosphate synthase (IGPS). Then they proposed the single shortest path connecting binding site to the functional site as the potential allosteric pathway [98]. This pathway included two residues that experimental studies had shown are involve in allostery. Later they expanded their work by developing Weighted Implementation of Suboptimal Paths (WISP) algorithm which in addition to single optimum shortest path finds other near-optimal paths [150]. This algorithm is based on the idea that while allostrey may occur through a single path for many proteins it could be summation of synergy of several paths. In this work we use Dijkstra's algorithm [151] implemented in Igraph [152] to solve for shortest paths between all $V(V-1)/2$ node pairs.

CHAPTER 6

APPLICATION OF NEW METHOD FOR CONNECTING ENSEMBLE SHIFTS TO REGULATION

In order to test the new method we need a test-case model with relatively known dynamical allostery behavior. As mentioned above this behavior is reported for large list of proteins from different families.

6.1 PDZ domains

We used the PDZ2 domain of human phosphatase PTP1E for this purpose (see Figure 6.1). This system has several characteristics which makes it ideal model for studies on dynamic allostery and it has been used in many allostery pathway prediction models as a benchmark [153, 56, 154, 155, 156, 157, 158, 57, 105, 159, 160, 161, 162, 163, 164, 165]. These characteristics are: it is a signaling module of many proteins, it has high resolution 3D structures with small change due to activation, it has been subject of many experimental and computational studies, and finally it is a small domain with less than 100 amino acids which makes the ensemble generation and interpretation of the results in the molecular level easy.

6.2 Generating ensembles using molecular dynamics

6.2.1 Molecular dynamics

The starting coordinates for molecular dynamics of the apo and the GEF2-bound states are taken from crystallographic structures [159] deposited in the Protein Databank (PDB

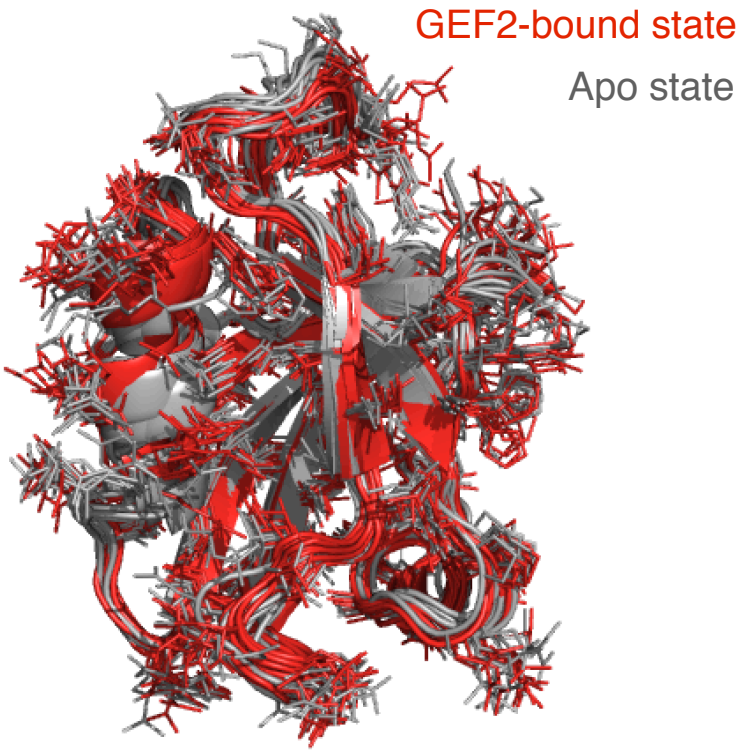

Figure 6.1. Superimposed conformational ensembles of the PDZ2 domain in the apo and GEF2-bound states. Each of the two conformational ensembles is represented using 11 snapshots taken at regular intervals from their respective molecular dynamics trajectories (see methods). For the sake of clarity, the GEF2 peptide is not shown.

IDs: 3LNX and 3LNY). Hydrogens are added and their positions optimized using PDB2PQR [166]. The N- and C-termini of the GEF2 peptide and the protein are capped by adding ACE and NME, respectively. The apo and the GEF2-bound structures are placed in cubic boxes containing ∼11K water molecules, including those resolved crystallographically. KCl concentration is set at 75 mM, and there are extra K^+ ions compared to Cl^- ions to compensate for the charge on the GEF2 peptide. MD simulations are carried out in duplicates (different random seeds for velocities) for both the apo and the GEF2 bound states of PDZ2, and each MD simulation is 0.5 μs long.

All four MD simulations are carried out under isothermal-isobaric boundary conditions, and using Gromacs version 5 [167]. Temperature is maintained at 298 K using an extended

ensemble approach [168, 169] and with a coupling constant of 0.2 ps. An extended ensemble approach is also used for maintaining pressure [170]. Pressure is maintained at 1 bar using a coupling constant of 1 ps and a compressibility of 4.5×10^{-5} bar^{-1}. Electrostatic interactions are computed using the particle mesh Ewald scheme [171] with a Fourier grid spacing of 0.1 nm, a fourth-order interpolation, and a direct space cutoff of 10 Å. van der Waals interactions are computed explicitly for interatomic distances $\leq$ 10 Å. The bonds in proteins and the geometries of water molecules are constrained [172, 173], and consequently an integration time step of 2 fs is employed. The protein and ions are described using Amber99sb-ILDN parameters [174], and water molecules are described using SPCE parameters [175]. Convergence is administered by tracking time evolutions of backbone RMSDs, pressure and potential energies, and consequently only the second half of each trajectory (0.25 μs) is used for analysis.

6.2.2 NMR data reproduction

NMR spin relaxation parameters uniquely suited for proper benchmarking MD simulations against quantitative experimental measurement specially internal protein dynamics[from NMR order parameter Dter.]. The employed MD protocol reproduces well the methyl deuterium order parameters obtained from NMR [56, 155] (6.3). The deuterium order parameters (S^2_{axis}) are computed by modeling the autocorrelation function based on the model-free approach [176].

$$C(t) = \frac{1}{2}(3\langle\hat{\mu}(0)\cdot\hat{\mu}(t)\rangle^2 - 1) \tag{6.1}$$

as an exponential decay

$$C(t) = S^2_{axis} + (1 - S^2_{axis})e^{-t/\tau}; \tag{6.2}$$

an assumption also used in estimating order parameters from NMR spectral densities. In the expressions above $\hat{\mu}$ are the unit vectors of C–C bonds in which the latter carbon is part of the methyl group, and τ is the relaxation time. We assume here that the order parameters of the C–C(H_3) bonds represent those of the hydrogens in the CH_3 groups. Figure 1.2 depicts the single exponential curve-fitting on computed autocorrelations of the second half of MD trajectories (250ns) for residue Valine 84. The estimated (S^2_{axis})s were compared against experimental counterparts in Figure 2.2 a) which showed good agreement. Figure 2.2 b) shows the convergence of computed (S^2_{axis})s.

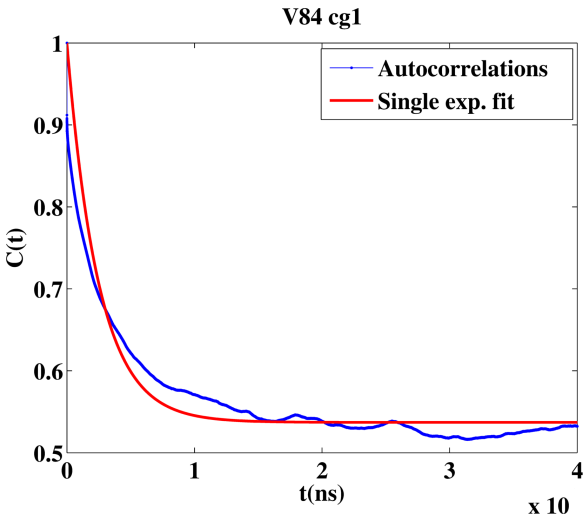

Figure 6.2. autocorrelations and single exponential curve-fitting of Valine 84.

6.3 Ensemble difference quantification and repartitioning

We generate duplicate MD trajectories of the PDZ2 domain in its apo and GEF2-bound states. Each trajectory is 0.5 μs long, and we use the second halves of these trajectories to construct conformational ensembles for analysis. To determine whether the latter halves of these trajectories provide adequate representations of conformational ensembles, we compute residue-wise differences between conformational ensembles constructed from duplicate

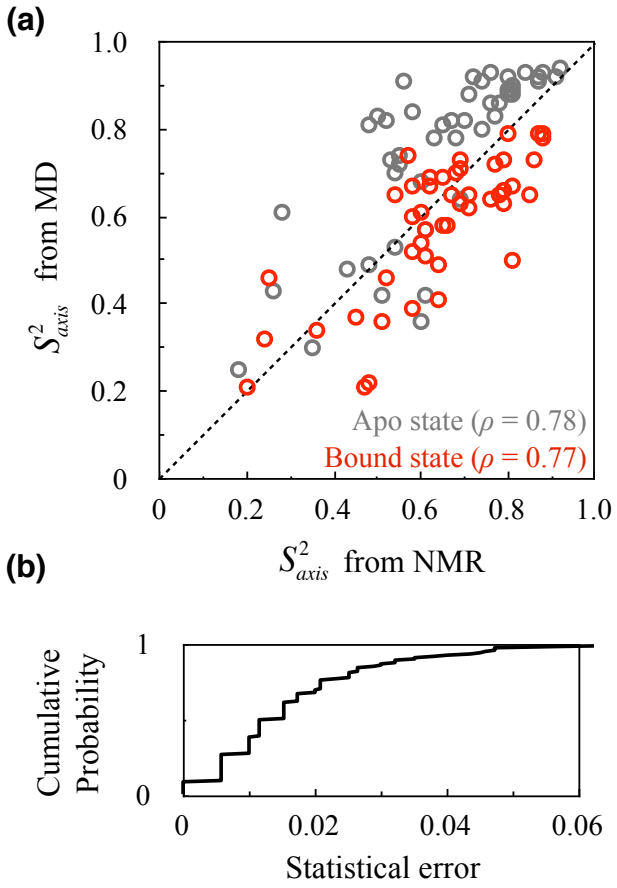

Figure 6.3. (a) Methyl deuterium order parameter (S^2_{axis}) computed from the final 250 ns of MD compare well with those estimated from NMR [155]. ρ denotes the Pearson correlation coefficient between the computed and experimental S^2_{axis} values. (b) Distribution of statistical error in estimating (S^2_{axis}) from MD, determined from block averaging over the final 100 ns of MD [177]. Note that for almost all cases the error < 0.05, indicating that the S^2_{axis} values are statistically converged.

trajectories, that is, for each residue i in the PDZ2 domain we determine,

$$\eta_i^{1\leftrightarrow 2} = 1 - \|f_i^1 \cap f_i^2\|, \tag{6.3}$$

where f_i^1 and f_i^2 are the ensembles of the same residue i extracted from duplicate trajectories, and $\|f_i^1 \cap f_i^2\|$ is the physical overlap between the ensembles. We determine η using a SVM based method we developed previously [34], which is also described briefly in the

methods section. η is bounded, that is, $\eta \in [0, 1)$, and takes up a value closer to unity as the difference between ensembles increases. Each ensemble contains 5000 snapshots extracted at regular time intervals from their respective trajectories. Note that prior to extracting the coordinates of a residue from a conformation of the PDZ2 domain, the entire conformation of the PDZ2 domain is least-square fitted on to the starting structure, which removes the bias against whole molecule rotation and translation [178]. 6.4 shows the cumulative distribution of residue-wise η computed separately for both the apo and GEF2-bound state ensembles. We find that the 90% of the residues have η values smaller than 0.35, which is equivalent to a mean position difference of less than $erf(0.35/\sqrt{2}) = 0.27$ Å, [36], showing that the differences between the duplicate trajectories are small. We also note that the η of a few residues, especially in the apo state, are large, but an inspection of residue identities reveals that they belong to the N- and C- termini of the PDZ2 domain. We exclude these residues from further analysis. Instead of discarding the data from the duplicate trajectories, we combine the ensembles from the duplicate trajectories, and create one representative 10000-conformation ensemble for each of the apo (f) and GEF2-bound states (g). We then estimate the difference between these ensembles $\eta_i^{f \leftrightarrow g}$ and compare them to the $\eta_i^{1 \leftrightarrow 2}$ estimated for duplicate trajectories. We find, in general, that $\eta_i^{f \leftrightarrow g} >> \eta_i^{1 \leftrightarrow 2}$, which shows that the statistical differences between duplicate trajectories is smaller than the GEF2-induced shifts in conformational ensembles. Together, this analysis shows that the latter half of the trajectories provide adequate representations of conformational ensembles of the two states.

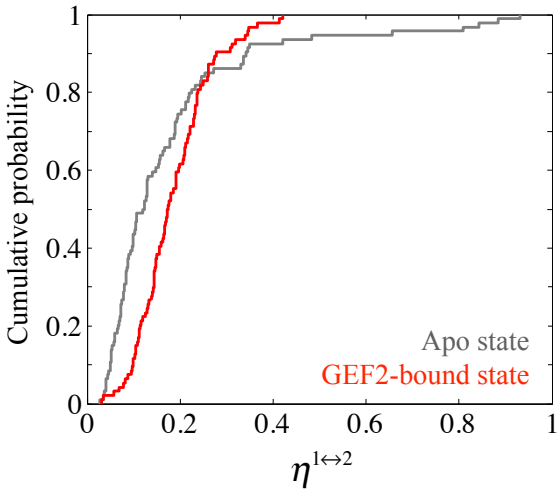

Figure 6.4. Cumulative probability distribution of residue $\eta^{1\leftrightarrow 2}$ between duplicate trajectories.

6.4 Results

6.5 compares the GEF2-induced shifts in residue centers-of-mass (CoMs) and root mean square fluctuations (RMSFs). In general, we note that GEF2 affects the structure and dynamics of residue side chains more than their respective backbones, a result consistent with previous studies [56, 159]. Such a form of induced changes have contributed to the hypothesis [56] that allosteric regulation in PDZ2 occurs primarily due to changes in side chain structure and dynamics. However, such a mechanistic model downplays the contributions of residues that undergo relatively smaller changes in backbone structure and dynamics. As such, there is no formal theory that relates signaling propensity to the extent of induced shifts, and so understanding regulatory mechanisms requires estimation of shifts and many-body correlations in conformational ensemble.

Toward this end, we first determine all pairs of residues that physically interact with each other. Typically, this is achieved by measuring distances between the average CoMs of two residues, and if that distance were less than a predefined cutoff, which is generally around 5 Å, [30, 26, 28, 29] then the two residues are assumed to physically interact with each

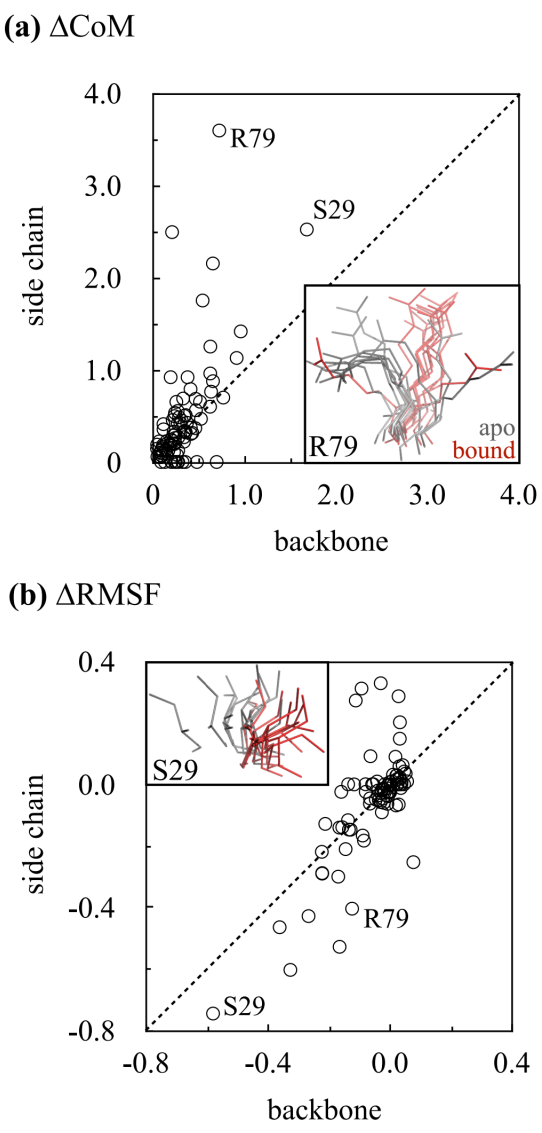

Figure 6.5. GEF2-induced shifts (Å) in residue centers-of-mass (ΔCoMs) and root mean square fluctuations (ΔRMSFs). GEF2 affects the structure and dynamics of residue side chains more than their respective backbones. The inset in (a) compares the conformational ensembles (11 equally spaced representative snapshots) of R79, the residue whose side chain undergoes the highest change in CoM. The inset in (b) compares the conformational ensembles of S29, the residue that undergoes the highest change in RMSF.

other. Instead of using pre-defined cutoffs, we compute the overlap between the volumes of residue conformational ensembles, and two residues are considered to physically interact if their volume overlap is non-zero. We assemble together these pairwise connectivities using

undirected graphs. We construct three such connectivity graphs, one using residue ensembles from the receptor-free state (G_f), another using residue ensembles from the GEF2-bound state (G_g), and the third (G_{fg}) using the union of G_f and G_g. All graphs have the same number of $V = 89$ nodes (residues), but they have different numbers of connected edges (interacting residue pairs) – G_f has 810 edges, G_g has 714 edges and G_{fg} has 846 edges. On average, each node in G_f has 9.1 edges, each node in G_g has 8.0 edges and each node in G_{fg} has 9.5 edges.

For all residue pairs in G_f and G_g, we then compute inter-site correlations C_{ij}^f and C_{ij}^g, respectively, using 5.4. For all pairs in G_{fg}, we compute inter-site correlations in ensemble shifts, which we do in two ways, one using 5.5 and the other using 5.6, which yield, respectively, $C_{ij}^{f\to g}$ and $C_{ij}^{g\to f}$. We then use the inverse of these correlations as numerical weights on the edges of the graph. Note that we get two separate G_{fg} graphs, $G_{f\to g}$ and $G_{g\to f}$, depending on whether we use $1/C_{ij}^{f\to g}$ or $1/C_{ij}^{g\to f}$ as edge weights. We then solve for shortest weighted paths between all $V(V-1)/2$ pairs of nodes. After solving for shortest paths, we count how many times each node appears in the $V(V-1)/2$ shortest paths. We indicate node-occurrences by the symbol Ω_i. Note that $(V-1) \leq \Omega_i \leq V(V-1)/2$. We do this separately for each of the 4 graphs, and so for each graph, we obtain a separate set of node-occurrences $\{\Omega_i\}$.

We assume that a residue that has a higher Ω contributes more to allosteric signaling [30, 26, 28, 29], and so we rank all residues in decreasing order of their Ω. This yields, for each of the four graphs or signaling models, an ordered set of node-occurrences $\{\Omega_i^{rank}\}$. 6.1 shows the Pearson correlation coefficients between residue ranks in the four signaling network models. We note first that the correlations are small. The correlations are even smaller if only the top ranked (25%) residues are considered in each model. Now if we ignore the relative ordering in the top ranked residues, we find that the pairwise identity overlaps between the four models are around 50%. Taken together, these observations imply that

if a residue contributes significantly to signaling in the apo state, it does not necessarily imply that it will also contribute significantly in the bound state. Furthermore, inter-state regulatory signals, which we compute from residue ordering in the $G_{f\to g}$ and $G_{g\to f}$ models, are not necessarily propagated by residues that contribute to intra-state signaling in G_f, G_g. This finding cautions against the reliance on single state models for garnering molecular insight into regulatory mechanisms. Finally, since the correlation between residue ranks in the $G_{f\to g}$ and $G_{g\to f}$ models is small, we conclude that the GEF2-binding and GEF2-unbinding response signals propagate through different networks.

Table 6.1. Pearson correlation coefficient between residue ranks in the four signaling models G_f, G_g, $G_{f\to g}$ and $G_{g\to f}$.

	G_f	G_g	$G_{f\to g}$	$G_{g\to f}$
G_f	1	0.18	0.23	0.31
G_g		1	0.18	0.28
$G_{f\to g}$			1	0.20
$G_{g\to f}$				1

The relative contribution of each residue to the overall signaling network can be given by the fraction

$$\bar{\Omega}_i = \frac{\Omega_i - (V-1)}{\sum_{i=1}^{V}(\Omega_i - (V-1))}. \tag{6.4}$$

Note that Ω_i are rescaled and this rescaling is phenomenologically equivalent to discarding occurrences of residue i in paths where they serve as end points, and so this rescaling yields a residue's contribution to signaling that is not initiated by that residue. These contributions can be rank ordered (highest to lowest contribution) and then summed to determine the subset of residues that carry out the bulk of the signaling. We, therefore, define cumulative signaling as

$$\Theta_k = \sum_{rank=1}^{k} \bar{\Omega}_i^{rank} \tag{6.5}$$

which approaches unity as k approaches the total number of residues (nodes) V in the network. 6.6 compares the cumulative signaling of the four signaling models. We note that regulatory signaling, that is, signaling due to GEF2 binding/unbinding, requires a considerably smaller set of residues than signaling within an individual state – while 75% of the intra-state signaling (in G_f and G_g) is carried out by 30 residues, 75% of the signaling in $G_{f\to g}$ and $G_{g\to f}$ require only 19 residues. This is opposite to what we would expect given that there are more edges in $G_{f\to g}$ and $G_{g\to f}$ compared to G_f or G_g. This surprising result can be explained by comparing inter-residue correlations in the four signaling models (6.7 and 6.8), which show that inter-site correlations in thermal fluctuations are, in general, more widespread and stronger than inter-site correlations in ensemble shifts.

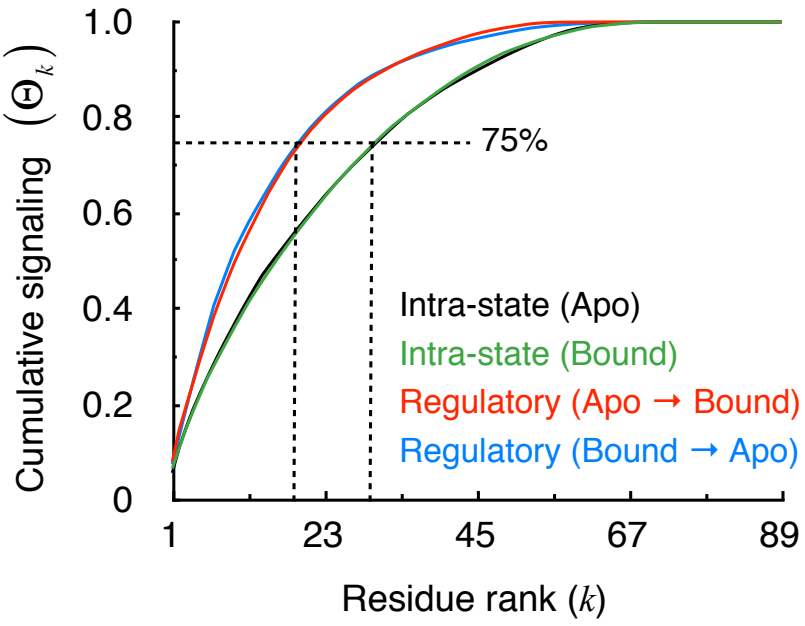

Figure 6.6. Comparison of cumulative signaling (6.5) in graphs weighted using intra-state correlations in thermal fluctuations and graphs weighted using inter-state correlations in ensemble shifts.

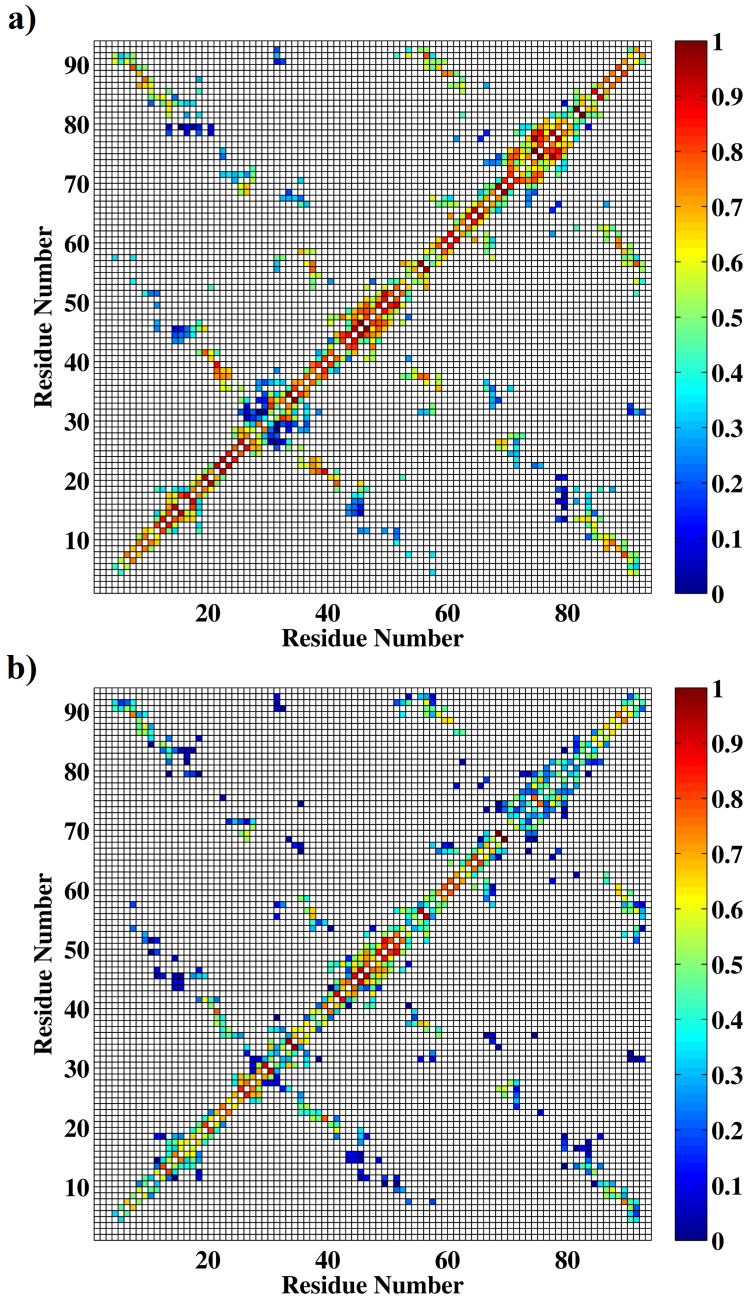

Figure 6.7. Heat maps of inter-site correlations in thermal fluctuations (C_{ij}^{f} and C_{ij}^{g}). The correlations are normalized by dividing each set by their respective highest values.

6.9 shows the identities and conformational summary statistics of the 19 residues that provide 75% of the signaling in $G_{f\to g}$ and $G_{g\to f}$. We see not only a weak correlation between

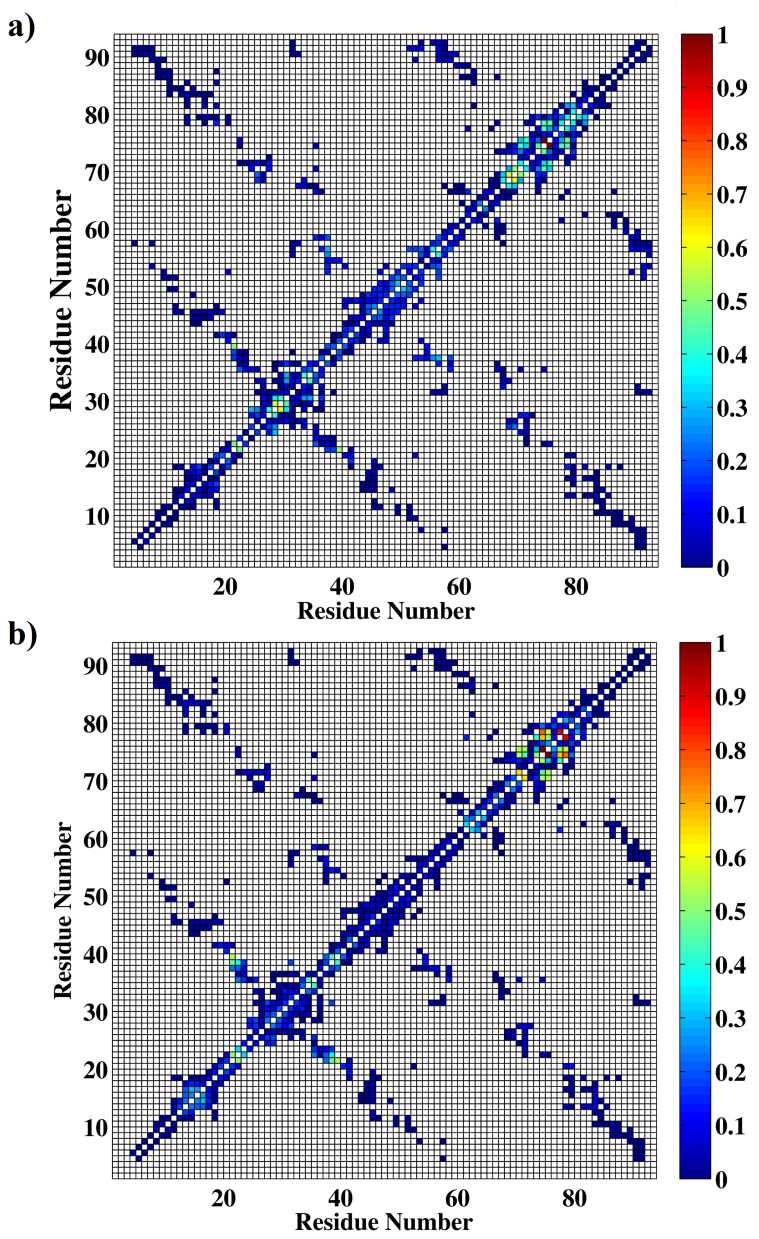

Figure 6.8. Heat maps of inter-site correlations in ensemble shifts ($C_{ij}^{f\to g}$ and $C_{ij}^{g\to f}$). The correlations are normalized by dividing each set by their respective highest values.

residue ranks in the two subnetworks, but also just a partial overlap in residue identities. Notably, while residue D56 has the highest contribution in $G_{f\to g}$, it is ranked 19^{th} in $G_{g\to f}$. Conversely, while residue T70 is ranked 2^{nd} in $G_{g\to f}$, it is ranked 17^{th} in $G_{f\to g}$. There is

also no direct relationship between the spatial location of residues and their contribution to signaling.

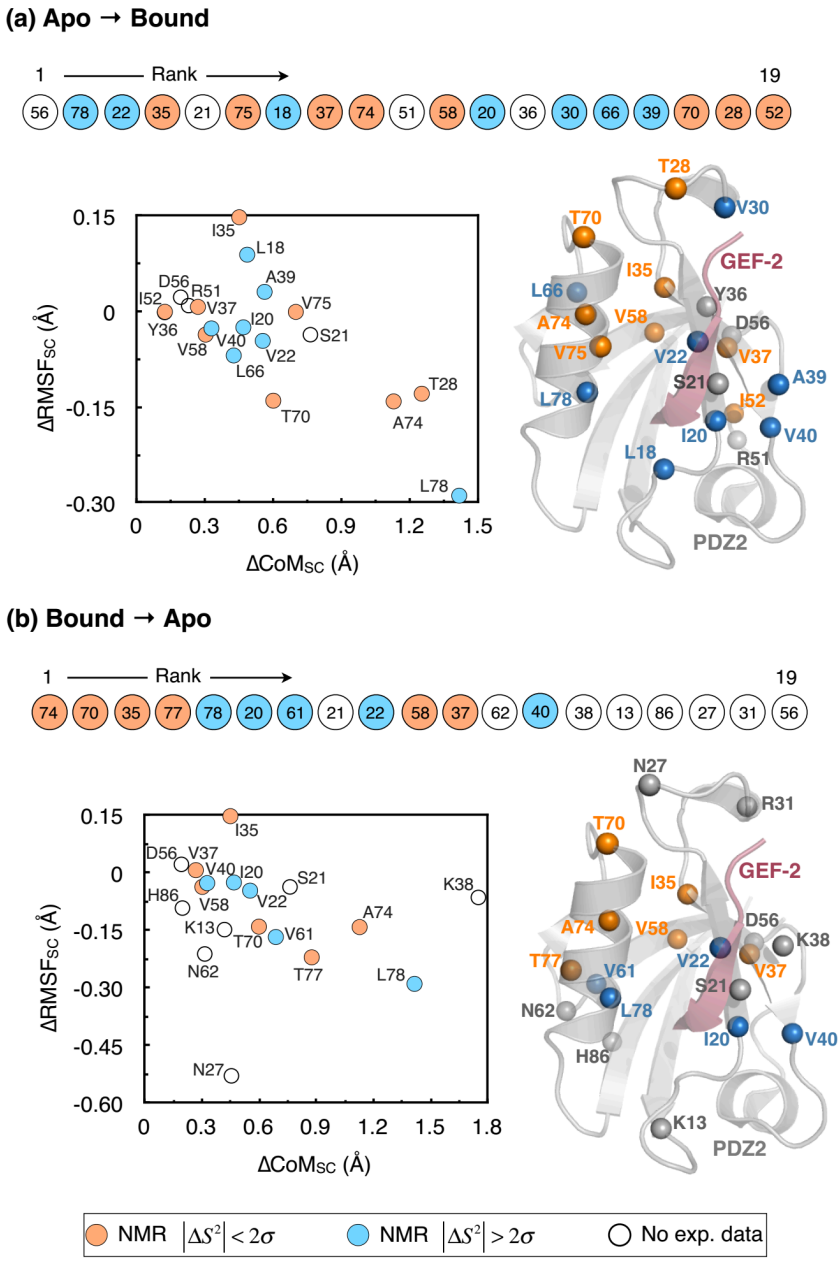

Figure 6.9. Identities, ranks and conformational summary statistics of residues that contribute to 75% of cumulative signaling. The residues are also color-coded according to whether their NMR order parameters change upon GEF2 binding.

We also find no correlations between a residue's rank and its ΔCoM and ΔRMSF, recommending that cutoffs on summary statistics should be avoided when identifying residues important to allosteric regulation. For example, D56 has the highest contribution in $G_{g\to f}$, but undergoes relatively small changes in both CoM and RMSF. We attribute its high contribution to the high number of moderately-correlated connections it makes in the signaling network. 6.10 shows the local connectivity of D56 in $G_{f\to g}$ and $G_{g\to f}$. We note that almost all correlations in $G_{f\to g}$ are relatively stronger than the respective correlations in $G_{g\to f}$, and this is perhaps why D56 emerges as the highest contributor in $G_{f\to g}$, but not in $G_{g\to f}$. 6.10 shows the local connectivity of another residue, T70, which contributes more to $G_{g\to f}$ than $G_{f\to g}$. Relative to D56, T70 has fewer connections, but several of T70's connections in $G_{g\to f}$ are highly correlated ($C_{ij} > 0.2$), which rationalize its high contribution to $G_{g\to f}$.

The two examples above, however, appear to suggest that a residue's contribution to signaling depends more on the strengths of correlated connections rather than the number of spatial connections. To examine this further, we compute for each edge in $G_{f\to g}$ the total number of times it occurs in all shortest paths (Ω_{ij}). This differs from Ω_i in that the number passes are computed over edges, instead of over nodes. Just as in the case of node-occurences Ω_i, we assume that an edges having higher Ω contribute more to the signaling network. Figure 7.2.a shows the 3D X-ray crystal structure of the PDZ2 bound to GEF2 where 7.2.b shows the edges color coded based on number of pass. 7.2.c shows 31 edges that has more than 50% of all the passes. Finally, 6.12 shows that there is only a weak relationship between Ω_{ij} and the strengths of the correlations $C_{ij}^{f\to g}$ (Pearson correlation $= -0.21$). Therefore, we conclude that a residue's contribution to signaling depends on both the strengths of correlated connections and the number of spatial connections.

In 6.9, we also note that some residues undergo GEF2-induced changes primarily in CoMs, some undergo changes primarily in RMSF and others undergo changes in both CoM and RMSF. This leads us to conclude that regulation in PDZ2 emanates from a combination

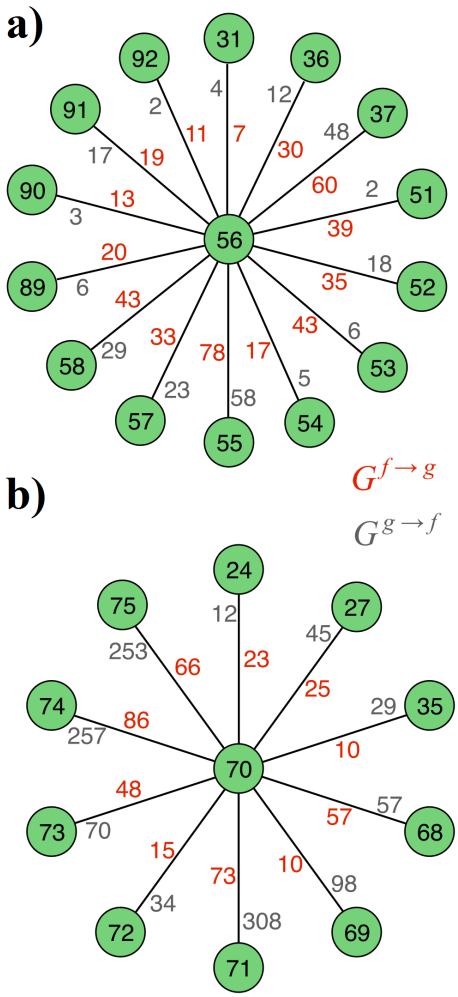

Figure 6.10. Local connectivities of residues D56 and T70 in $G_{f\to g}$ and $G_{g\to f}$. The nodes are represented as filled circles, and the edges are represented as lines. The two numbers on the lines represent correlations ($\times 10^{-3}$) in $G_{f\to g}$ (red) and $G_{g\to f}$ (gray).

of changes in structure and dynamics, and not just changes in dynamics, as is occasionally argued [157, 179, 164]. In other words, dynamics does play a key role in allosteric regulation, but it is not the sole mode of signal transduction. Additionally, we note that not all residues that undergo changes in NMR S_2^{axis} contribute to regulation. Out of the 14 residues that were found to undergo changes in S_{axis}^2, only about half of them contribute significantly to signaling. Conversely, residues that are not found to undergo changes in S_2^{axis} can still

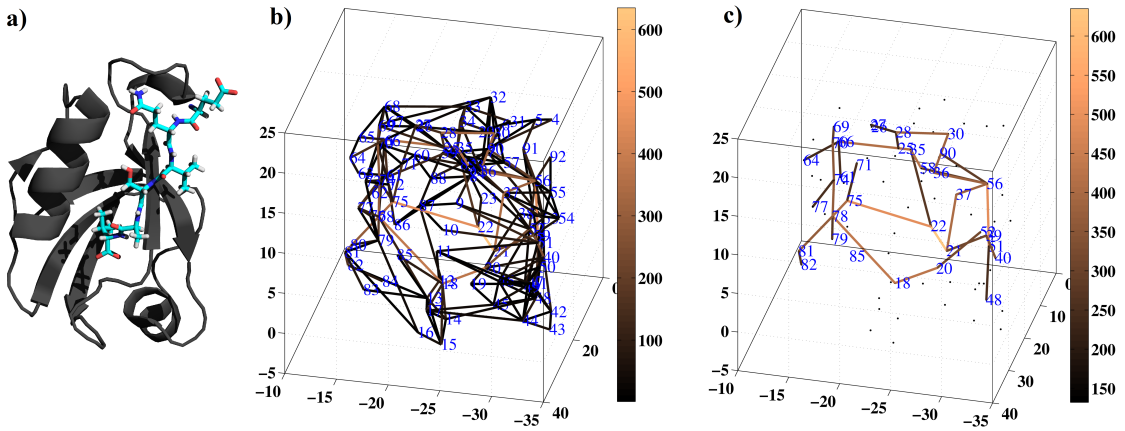

Figure 6.11. a) PDZ2 crystal structure in active state with the RA-GEF2 C-terminal peptide in blue. b) all of the edges color coded with number of visits. c) 14 % of the edges with the highest number of visits that have more than 50% of total number of passes.

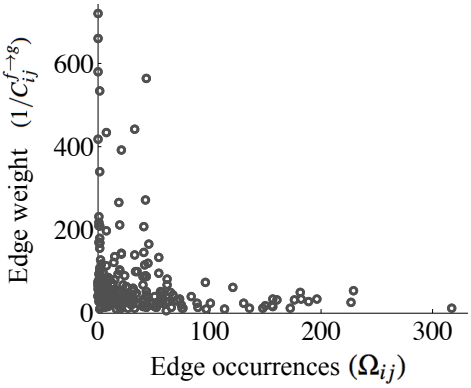

Figure 6.12. Correlation between edge weights $(1/C_{ij}^{f\to g})$ and edge occurrences (Ω_{ij}) in shortest paths in $G_{f\to g}$.

contribute to signaling through changes in structure and dynamics. Its important to note here that changes in S^2_{axis} are not equivalent to changes in RMSF – S^2_{axis} and RMSF can be related, but they are fundamentally two different summary statistics of dynamics.

CHAPTER 7

CONCLUSION AND FUTURE DIRECTIONS

Modeling Dynamic Allostery in Proteins Enabled by Machine Learning Conclusion: This work demonstrates the applicability of machine learning enabled approach to characterize the dynamic allostery mechanism. Since dynamic allostery comes with small structural changes the proper approach for modeling dynamic allostery requires to quantify the changes in ensemble of conformations. At the start of the thesis, Leighty and Varma had developed the very first method based on a machine learning technique (SVM) to quantify differences in ensembles at residue level [34]. Figure 7.1 shows the schematic representation of the method. This method unlike existing methods does not require any ad hoc fitting with specific assumptions on underlying distributions, and yielded differences in terms of a normalized metric that made the cross comparisons of ensembles possible.

The accuracy of the method was tested against 5 different conventional class separability measures. A new indexing scheme was used, software pipeline was optimized and parallel processing capability was added to the method implementation which made it more than 100x faster. We validate the efficiency and robustness of the method on a vast range of distributions as well as the internal coordinate system. The method is now publicly available at $https://simtk.org/home/conf\_ensembles$ to use.

We applied the method to computational molecular biology problems which showed dynamic allostery behavior such as: (a) identify the effects of employed different force fields on conformational ensembles [35], (b) cross-comparison of multiple ensembles to determine the

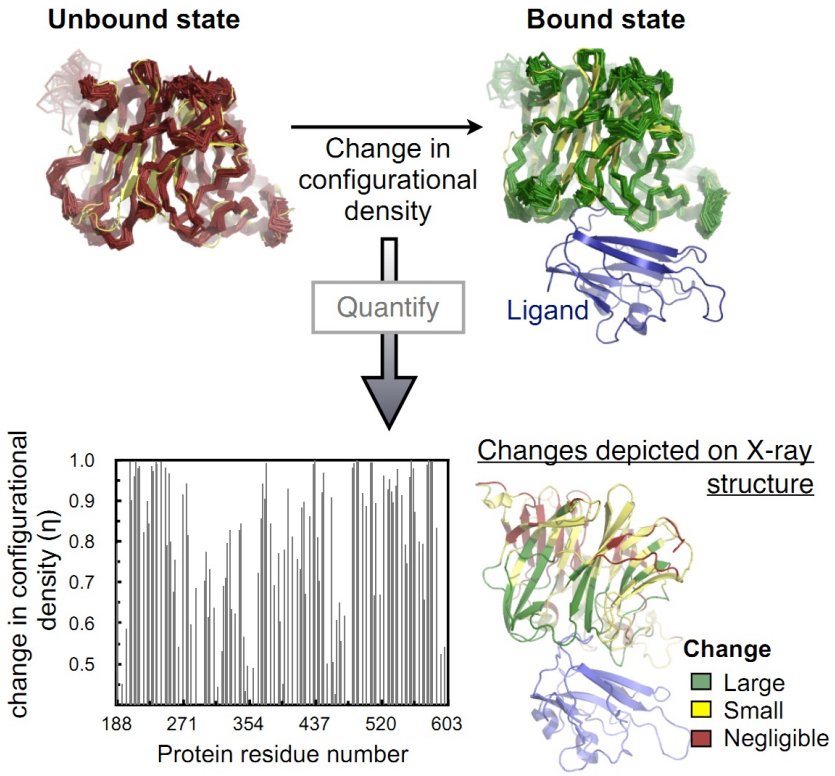

Figure 7.1. On top twenty representative structures of NiV-G superimposed on X-ray structure in yellow. On bottom quantification of ensemble changes on residue level due to binding of a ligand (Adapted with permission form [34]).

common signaling pathways induced by different effectors [36], (c) characterization of effects of point mutations on conformational ensemble shifts in proteins [37].

To gain more insight into dynamic allostery mechanism we need another method to actually relate the induced conformational ensembles changes in one site to another site. Although several method were proposed to model this phenomenon they are not capable to model the mechanism accurately. The existing methods for regulatory signal prediction, model this pathway/network with information of only one state of the protein.

We developed a new method to use information of both state for relating the ensemble population shifts into inter-site signal communications. This method uses mathematical framework of SVM for repartitioning the conformational ensemble shifts that enables us to

calculate inter-site correlation of population shifts. Then we uses shortest path algorithm to find optimum communication pathways between all amino acid pairs. This analysis followed by an enumeration of number of visits of these optimum pathways from amino acids and the edges between them.

We applied this method on hPTP1E's PDZ2 domain which is the most used test-case model for studying dynamic allostery. The results of the shortest path analysis and enumeration on the new two-state model were compared against the results of similar analysis on the single-state models (PDZ in apo state as well as bound state). It showed that there is a sub-network depicted in figure 7.2 with high density of optimum pathways that existed only in the new two-state model. Which demonstrate the dramatic alteration of preferred pathways due to binding an effector. In other words the regulatory networks are very different from the inter-site communication networks present in individual states, highlighting that a residue's role in regulation cannot be projected from its contribution to signaling in a given state. Consistent with earlier predictions, we report that the regulatory network in the PDZ2 domain indeed emerges from a combination of changes in structure and dynamics, and even small changes in structure contribute significantly. Moreover, there is a very weak correlation between the extent of inter-site correlations and the number of visits on that specific edge which suggest caution in using thresholds for interactions specially for system with small changes.

In future works method can be tested on PDZ domains with mutation of hubs we found in this study to check the capability of method on potential forward prediction that can be validated experimentally. It also can be test on similar PDZ domains such as PDZ3 with very similar structure but slightly different proposed activation mechanism. There are other protein systems such as viral attachment proteins which show dynamic allostery behavior and we plan to apply this method on them to demonstrate the generality of the developed approach.

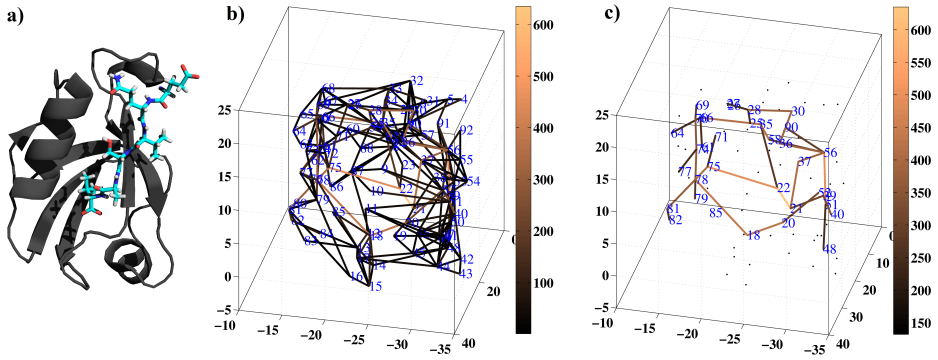

Figure 7.2. a) PDZ2 crystal structure in active state with the RA-GEF2 C-terminal peptide in blue. b) all of the edges color coded with number of visits. c) 14 % of the edges with the highest number of visits that have more than 50% of total number of passes.

REFERENCES

[1] Jacques Monod and François Jacob. General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation. In *Cold Spring Harbor symposia on quantitative biology*, volume 26, pages 389–401. Cold Spring Harbor Laboratory Press, 1961.

[2] Jean-Pierre Changeux. The feedback control mechanism of biosynthetic l-threonine deaminase by l-isoleucine. In *Cold Spring Harbor symposia on quantitative biology*, volume 26, pages 313–318. Cold Spring Harbor Laboratory Press, 1961.

[3] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: a plausible model. *Journal of molecular biology*, 12(1):88–118, 1965.

[4] Daniel J Mandell and Tanja Kortemme. Computer-aided design of functional protein interactions. *Nature Chemical Biology*, 5(11):797–807, 2009.

[5] Jeffrey R Wagner, Christopher T Lee, Jacob D Durrant, Robert D Malmstrom, Victoria A Feher, and Rommie E Amaro. Emerging computational methods for the rational discovery of allosteric drugs. *Chemical reviews*, 116(11):6370–6390, 2016.

[6] Andrei V Karginov, Feng Ding, Pradeep Kota, Nikolay V Dokholyan, and Klaus M Hahn. Engineered allosteric activation of kinases in living cells. *Nature biotechnology*, 28(7):743–747, 2010.

[7] DE Koshland Jr, G Nemethy, and D_ Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5(1):365–385, 1966.

[8] Qiang Cui and Martin Karplus. Allostery and cooperativity revisited. *Protein science*, 17(8):1295–1307, 2008.

[9] K Gunasekaran, Buyong Ma, and Ruth Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics*, 57(3):433–443, 2004.

[10] Gregorio Weber. Ligand binding and internal equilibiums in proteins. *Biochemistry*, 11(5):864–878, 1972.

[11] A Cooper and DTF Dryden. Allostery without conformational change. *European Biophysics Journal*, 11(2):103–109, 1984.

[12] Hesam N Motlagh, James O Wrabl, Jing Li, and Vincent J Hilser. The ensemble nature of allostery. *Nature*, 508(7496):331–339, 2014.

[13] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

[14] Ivet Bahar, Timothy R Lezon, Lee-Wei Yang, and Eran Eyal. Global dynamics of proteins: bridging between structure and function. *Annual review of biophysics*, 39:23–42, 2010.

[15] Shiou-Ru Tzeng and Charalampos G Kalodimos. Protein dynamics and allostery: an nmr view. *Current opinion in structural biology*, 21(1):62–67, 2011.

[16] Chung-Jung Tsai and Ruth Nussinov. A unified view of how allostery works. *PLoS Comput Biol*, 10(2):e1003394, 2014.

[17] Kateri H DuBay, Gregory R Bowman, and Phillip L Geissler. Fluctuations within folded proteins: implications for thermodynamic and allosteric regulation. *Accounts of chemical research*, 48(4):1098–1105, 2015.

[18] Peter S Shenkin, Batu Erman, and Lucy D Mastrandrea. Information-theoretical entropy as a measure of sequence variability. *Proteins: Structure, Function, and Bioinformatics*, 11(4):297–313, 1991.

[19] Gregory Manley and J Patrick Loria. Nmr insights into protein allostery. *Archives of biochemistry and biophysics*, 519(2):223–231, 2012.

[20] Galen Collier and Vanessa Ortiz. Emerging computational approaches for the study of protein allostery. *Archives of biochemistry and biophysics*, 538(1):6–15, 2013.

[21] Lei Yang, Guang Song, and Robert L Jernigan. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences*, 106(30):12347–12352, 2009.

[22] Michael D Daily and Jeffrey J Gray. Local motions in a benchmark of allosteric proteins. *Proteins: Structure, function, and bioinformatics*, 67(2):385–399, 2007.

[23] Michael D Daily and Jeffrey J Gray. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput Biol*, 5(2):e1000293, 2009.

[24] Yifei Kong and Martin Karplus. The signaling pathway of rhodopsin. *Structure*, 15(5):611–623, 2007.

[25] Chih-Peng Lin, Shao-Wei Huang, Yan-Long Lai, Shih-Chung Yen, Chien-Hua Shih, Chih-Hao Lu, Cuen-Chao Huang, and Jenn-Kang Hwang. Deriving protein dynamical properties from weighted protein contact number. *Proteins: Structure, Function, and Bioinformatics*, 72(3):929–935, 2008.

[26] T Lin and Guang Song. Predicting allosteric communication pathways using motion correlation network. In *Proceedings of the 7th Asia Pacific Bioinformatics Conference (APBC)*, pages 588–598. Tsinghua University, 2009.

[27] Wolfram Stacklies, Fei Xia, and Frauke Gräter. Dynamic allostery in the methionine repressor revealed by force distribution analysis. *PLoS Comput Biol*, 5(11):e1000574, 2009.

[28] Amit Ghosh, Reiko Sakaguchi, Cuiping Liu, Saraswathi Vishveshwara, and Ya-Ming Hou. Allosteric communication in cysteinyl trna synthetase a network of direct and indirect readout. *Journal of Biological Chemistry*, 286(43):37721–37731, 2011.

[29] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular systems biology*, 2(1), 2006.

[30] Chakra Chennubhotla and Ivet Bahar. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput Biol*, 3(9):e172, 2007.

[31] Canan Atilgan and Ali Rana Atilgan. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput Biol*, 5(10):e1000544, 2009.

[32] C Atilgan, ZN Gerek, SB Ozkan, and AR Atilgan. Manipulation of conformational change in proteins by single-residue perturbations. *Biophysical Journal*, 99(3):933–943, 2010.

[33] Diego U Ferreiro, Joseph A Hegler, Elizabeth A Komives, and Peter G Wolynes. On the role of frustration in the energy landscapes of allosteric proteins. *Proceedings of the National Academy of Sciences*, 108(9):3499–3503, 2011.

[34] Ralph E Leighty and Sameer Varma. Quantifying changes in intrinsic molecular motion using support vector machines. *Journal of chemical theory and computation*, 9(2):868–875, 2013.

[35] Priyanka Dutta, Mohsen Botlani, and Sameer Varma. Water dynamics at protein–protein interfaces: Molecular dynamics study of virus–host receptor complexes. *The Journal of Physical Chemistry B*, 118(51):14795–14807, 2014.

[36] Sameer Varma, Mohsen Botlani, and Ralph E Leighty. Discerning intersecting fusion-activation pathways in the nipah virus using machine learning. *Proteins: Structure, Function, and Bioinformatics*, 82(12):3241–3254.

[37] Priyanka Dutta, Ahnaf Siddiqui, Mohsen Botlani, and Sameer Varma. Stimulation of nipah fusion: Small intradomain changes trigger extensive interdomain rearrangements. *Biophysical Journal*, 111(8):1621–1630, 2016.

[38] MF Perutz, Wo Bolton, R Diamond, Hilary Muirhead, and HC Watson. Structure of haemoglobin: an x-ray examination of reduced horse haemoglobin. *Nature*, 203:687–690, 1964.

[39] Andrew L Lee et al. Frameworks for understanding long-range intra-protein communication. *Current Protein and Peptide Science*, 10(2):116–127, 2009.

[40] Max F Perutz and LF TenEyck. Stereochemistry of cooperative effects in hemoglobin. In *Cold Spring Harbor symposia on quantitative biology*, volume 36, pages 295–310. Cold Spring Harbor Laboratory Press, 1972.

[41] Max F Perutz, AJ Wilkinson, M Paoli, and GG Dodson. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annual review of biophysics and biomolecular structure*, 27(1):1–34, 1998.

[42] Gürol M Süel, Steve W Lockless, Mark A Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature structural & molecular biology*, 10(1):59–69, 2003.

[43] M Mercedes Silva, Paul H Rogers, and A Arnone. A third quaternary structure of human hemoglobin a at 1.7-a resolution. *Journal of Biological Chemistry*, 267(24):17248–17256, 1992.

[44] Brian F Volkman, Doron Lipson, David E Wemmer, and Dorothee Kern. Two-state allosteric behavior in a single-domain signaling protein. *Science*, 291(5512):2429–2433, 2001.

[45] Anders Malmendal, Johan Evenäs, Sture Forsén, and Mikael Akke. Structural dynamics in the c-terminal domain of calmodulin at low calcium levels. *Journal of molecular biology*, 293(4):883–899, 1999.

[46] Karen L Martinez, Yann Gohon, Pierre-Jean Corringer, Christophe Tribet, Fabienne Mérola, Jean-Pierre Changeux, and Jean-Luc Popot. Allosteric transitions of torpedo acetylcholine receptor in lipids, detergent and amphipols: molecular interactions vs. physical constraints. *FEBS letters*, 528(1-3):251–256, 2002.

[47] Robert G Smock and Lila M Gierasch. Sending signals dynamically. *Science*, 324(5924):198–203, 2009.

[48] Arthur Christopoulos. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nature reviews Drug discovery*, 1(3):198–210, 2002.

[49] Christine Berger, Susanne Weber-Bornhauser, Jolanda Eggenberger, Jozef Hanes, Andreas Plückthun, and Hans Rudolf Bosshard. Antigen recognition by conformational selection. *FEBS letters*, 450(1-2):149–153, 1999.

[50] Sandeep Kumar, Buyong Ma, Chung-Jung Tsai, Neeti Sinha, and Ruth Nussinov. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Science*, 9(1):10–19, 2000.

[51] Terry Kenakin. Efficacy at g-protein-coupled receptors. *Nature reviews Drug discovery*, 1(2):103–110, 2002.

[52] Terry Kenakin and Ongun Onaran. The ligand paradox between affinity and efficacy: can you be there and not make a difference? *Trends in pharmacological sciences*, 23(6):275–280, 2002.

[53] Ernesto Freire. The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody d1. 3 to lysozyme. *Proceedings of the National Academy of Sciences*, 96(18):10118–10122, 1999.

[54] Dorothee Kern and Erik RP Zuiderweg. The role of dynamics in allosteric regulation. *Current opinion in structural biology*, 13(6):748–757, 2003.

[55] Hong Pan, J Ching Lee, and Vincent J Hilser. Binding sites in escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. *Proceedings of the National Academy of Sciences*, 97(22):12020–12025, 2000.

[56] Ernesto J Fuentes, Channing J Der, and Andrew L Lee. Ligand-dependent dynamics and intramolecular signaling in a pdz domain. *Journal of molecular biology*, 335(4):1105–1115, 2004.

[57] Chad M Petit, Jun Zhang, Paul J Sapienza, Ernesto J Fuentes, and Andrew L Lee. Hidden dynamic allostery in a pdz domain. *Proceedings of the National Academy of Sciences*, 106(43):18249–18254, 2009.

[58] Shiou-Ru Tzeng and Charalampos G Kalodimos. Dynamic activation of an allosteric regulatory protein. *Nature*, 462(7271):368–372, 2009.

[59] Shiou-Ru Tzeng and Charalampos G Kalodimos. Protein activity regulation by conformational entropy. *Nature*, 488(7410):236–240, 2012.

[60] Nataliya Popovych, Shangjin Sun, Richard H Ebright, and Charalampos G Kalodimos. Dynamically driven protein allostery. *Nature structural & molecular biology*, 13(9):831–838, 2006.

[61] Lee A Freiburger, Oliver M Baettig, Tara Sprules, Albert M Berghuis, Karine Auclair, and Anthony K Mittermaier. Competing allosteric mechanisms modulate substrate binding in a dimeric enzyme. *Nature structural & molecular biology*, 18(3):288–294, 2011.

[62] Travis P Schrank, D Wayne Bolen, and Vincent J Hilser. Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. *Proceedings of the National Academy of Sciences*, 106(40):16984–16989, 2009.

[63] Jiangang Liu, Narayanan B Perumal, Christopher J Oldfield, Eric W Su, Vladimir N Uversky, and A Keith Dunker. Intrinsic disorder in transcription factors. *Biochemistry*, 45(22):6873–6888, 2006.

[64] Vladimir N Uversky. Intrinsically disordered proteins from a to z. *The international journal of biochemistry & cell biology*, 43(8):1090–1103, 2011.

[65] Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Showing your id: intrinsic disorder as an id for recognition, regulation and cell signaling. *Journal of Molecular Recognition*, 18(5):343–384, 2005.

[66] Peter E Wright and H Jane Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology*, 293(2):321–331, 1999.

[67] Peter Tompa. Unstructural biology coming of age. *Current opinion in structural biology*, 21(3):419–425, 2011.

[68] Abel Garcia-Pino, Sreeram Balasubramanian, Lode Wyns, Ehud Gazit, Henri De Greve, Roy D Magnuson, Daniel Charlier, Nico AJ van Nuland, and Remy Loris. Allostery and intrinsic disorder mediate transcription regulation by conditional cooperativity. *Cell*, 142(1):101–111, 2010.

[69] Alexey G Murzin. Metamorphic proteins. *Science*, 320(5884):1725–1726, 2008.

[70] Natively unfolded proteins. *Current Opinion in Structural Biology*, 15(1):35 – 41, 2005.

[71] Leo C James and Dan S Tawfik. Conformational diversity and protein evolution–a 60-year-old hypothesis revisited. *Trends in biochemical sciences*, 28(7):361–368, 2003.

[72] Nayden Koon, Christopher J Squire, and Edward N Baker. Crystal structure of leua from mycobacterium tuberculosis, a key enzyme in leucine biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8295–8300, 2004.

[73] Patrick A Frantom, Hui-Min Zhang, Mark R Emmett, Alan G Marshall, and John S Blanchard. Mapping of the allosteric network in the regulation of α-isopropylmalate synthase from mycobacterium tuberculosis by the feedback inhibitor l-leucine: solution-phase h/d exchange monitored by ft-icr mass spectrometry. *Biochemistry*, 48(31):7457–7464, 2009.

[74] James E Knapp, Reinhard Pahl, Vukica Šrajer, and William E Royer. Allosteric action in real time: time-resolved crystallographic studies of a cooperative dimeric hemoglobin. *Proceedings of the National Academy of Sciences*, 103(20):7649–7654, 2006.

[75] Gregory M Lee and Charles S Craik. Trapping moving targets with small molecules. *Science*, 324(5924):213–215, 2009.

[76] Anastasia Zhuravleva and Lila M Gierasch. Allosteric signal transmission in the nucleotide-binding domain of 70-kda heat shock protein (hsp70) molecular chaperones. *Proceedings of the National Academy of Sciences*, 108(17):6987–6992, 2011.

[77] James M Lipchock and J Patrick Loria. Nanometer propagation of millisecond motions in v-type allostery. *Structure*, 18(12):1596–1607, 2010.

[78] Patrick J Farber and Anthony Mittermaier. Concerted dynamics link allosteric sites in the pbx homeodomain. *Journal of molecular biology*, 405(3):819–830, 2011.

[79] William E Werner and HK Schachman. Analysis of the ligand-promoted global conformational change in aspartate transcarbamoylase: Evidence for a two-state transition from boundary spreading in sedimentation velocity experiments. *Journal of molecular biology*, 206(1):221–230, 1989.

[80] Brian A Kidd, David Baker, and Wendy E Thomas. Computation of conformational coupling in allosteric proteins. *PLoS Comput Biol*, 5(8):e1000484, 2009.

[81] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[82] S-W Zhang, Y-L Zhang, Q Pan, Y-M Cheng, and K-C Chou. Estimating residue evolutionary conservation by introducing von neumann entropy and a novel gap-treating approach. *Amino acids*, 35(2):495–501, 2008.

[83] Gary D Bader, Doron Betel, and Christopher WV Hogue. Bind: the biomolecular interaction network database. *Nucleic acids research*, 31(1):248–250, 2003.

[84] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–305, 2002.

[85] Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. Mint: a molecular interaction database. *FEBS letters*, 513(1):135–140, 2002.

[86] Frances C Bernstein, Thomas F Koetzle, Graheme JB Williams, Edgar F Meyer, Michael D Brice, John R Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *Archives of biochemistry and biophysics*, 185(2):584–591, 1978.

[87] Andrew A Bogan and Kurt S Thorn. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9, 1998.

[88] Ozlem Keskin, Buyong Ma, and Ruth Nussinov. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of molecular biology*, 345(5):1281–1294, 2005.

[89] Yanay Ofran and Burkhard Rost. Protein–protein interaction hotspots carved into sequences. *PLoS Comput Biol*, 3(7):e119, 2007.

[90] Kyu-il Cho, Dongsup Kim, and Doheon Lee. A feature-based approach to modeling protein–protein interaction hot spots. *Nucleic acids research*, 37(8):2672–2687, 2009.

[91] Joseph D Bryngelson and Peter G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, 84(21):7524–7528, 1987.

[92] Joseph D Bryngelson, José Nelson Onuchic, Nicholas D Socci, and Peter G Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.

[93] James A Brannigan and Anthony J Wilkinson. Protein engineering 20 years on. *Nature Reviews Molecular Cell Biology*, 3(12):964–970, 2002.

[94] Vikas Nanda and William F DeGrado. Computational design of heterochiral peptides against a helical target. *Journal of the American Chemical Society*, 128(3):809–816, 2006.

[95] Robert J Pantazes, Matthew J Grisewood, and Costas D Maranas. Recent advances in computational protein design. *Current opinion in structural biology*, 21(4):467–472, 2011.

[96] Michael V LeVine and Harel Weinstein. Nbit-a new information theory-based analysis of allosteric mechanisms reveals residues that underlie function in the leucine transporter leut. *PLoS Comput Biol*, 10(5):e1003603, 2014.

[97] Anurag Sethi, John Eargle, Alexis A Black, and Zaida Luthey-Schulten. Dynamical networks in trna: protein complexes. *Proceedings of the National Academy of Sciences*, 106(16):6620–6625, 2009.

[98] Adam T VanWart, John Eargle, Zaida Luthey-Schulten, and Rommie E Amaro. Exploring residue component contributions to dynamical network models of allostery. *Journal of chemical theory and computation*, 8(8):2949–2961, 2012.

[99] Irina G Tikhonova, Balaji Selvam, Anthony Ivetac, Jeff Wereszczynski, and J Andrew McCammon. Simulations of biased agonists in the $\beta 2$ adrenergic receptor with accelerated molecular dynamics. *Biochemistry*, 52(33):5593–5603, 2013.

[100] Paul M Gasper, Brian Fuglestad, Elizabeth A Komives, Phineus RL Markwick, and J Andrew McCammon. Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. *Proceedings of the National Academy of Sciences*, 109(52):21216–21222, 2012.

[101] Nigar Kantarci-Carsibasi, Turkan Haliloglu, and Pemra Doruker. Conformational transition pathways explored by monte carlo simulation integrated with collective modes. *Biophysical journal*, 95(12):5862–5873, 2008.

[102] Kateri H DuBay, Jacques P Bothma, and Phillip L Geissler. Long-range intra-protein communication can be transmitted by correlated side-chain fluctuations alone. *PLoS Comput Biol*, 7(9):e1002168, 2011.

[103] Kim Sharp and John J Skinner. Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. *Proteins: Structure, Function, and Bioinformatics*, 65(2):347–361, 2006.

[104] Jordi Silvestre-Ryan, Yuchun Lin, and Jhih-Wei Chu. fluctuograms reveal the intermittent intra-protein communication in subtilisin carlsberg and correlate mechanical coupling with co-evolution. *PLoS Comput Biol*, 7(3):e1002023, 2011.

[105] Yifei Kong and Martin Karplus. Signaling pathways of pdz2 domain: a molecular dynamics interaction correlation analysis. *Proteins: Structure, Function, and Bioinformatics*, 74(1):145–154, 2009.

[106] Ivan Rivalta, Mohammad M Sultan, Ning-Shiuan Lee, Gregory A Manley, J Patrick Loria, and Victor S Batista. Allosteric pathways in imidazole glycerol phosphate synthase. *Proceedings of the National Academy of Sciences*, 109(22):E1428–E1436, 2012.

[107] Kresten Lindorff-Larsen and Jesper Ferkinghoff-Borg. Similarity measures for protein ensembles. *PloS one*, 4(1):e4203, 2009.

[108] Rafael Brüschweiler. Efficient rmsd measures for the comparison of two molecular ensembles. *Proteins: Structure, Function, and Bioinformatics*, 50(1):26–34, 2003.

[109] Christopher L McClendon, Lan Hua, and Mathew P Jacobson. Comparing conformational ensembles using the kullback-leibler divergence expansion. *Journal of chemical theory and computation*, 8(6):2115, 2012.

[110] Michael J Bradley, Peter T Chivers, and Nathan A Baker. Molecular dynamics simulation of the escherichia coli nikr protein: equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *Journal of molecular biology*, 378(5):1155–1173, 2008.

[111] Oliver F Lange and Helmut Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1053–1061, 2006.

[112] Christopher L McClendon, Gregory Friedland, David L Mobley, Homeira Amirkhani, and Matthew P Jacobson. Quantifying correlations between allosteric sites in thermodynamic ensembles. *Journal of chemical theory and computation*, 5(9):2486–2502, 2009.

[113] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[114] Zhiyong Zhang and Willy Wriggers. Local feature analysis: a statistical theory for reproducible essential dynamics of large macromolecules. *PROTEINS: Structure, Function, and Bioinformatics*, 64(2):391–403, 2006.

[115] Angel E García. Large-amplitude nonlinear motions in proteins. *Physical review letters*, 68(17):2696, 1992.

[116] James B Clarage, Tod Romo, B Kim Andrews, B Montgomery Pettitt, and George N Phillips. A sampling problem in molecular dynamics simulations of macromolecules. *Proceedings of the National Academy of Sciences*, 92(8):3288–3292, 1995.

[117] Manel A Balsera, Willy Wriggers, Yoshitsugu Oono, and Klaus Schulten. Principal component analysis and long time protein dynamics. *The Journal of Physical Chemistry*, 100(7):2567–2572, 1996.

[118] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[119] Alex J Smola, Bernhard Schölkopf, and Klaus-Robert Müller. The connection between regularization operators and support vector kernels. *Neural networks*, 11(4):637–649, 1998.

[120] Bernhard Schölkopf and Christopher JC Burges. *Advances in kernel methods: support vector learning*. 1999.

[121] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. 2000.

[122] Robert N Jorissen and Michael K Gilson. Virtual screening of molecular databases using a support vector machine. *Journal of chemical information and modeling*, 45(3):549–561, 2005.

[123] Thorsten Joachims. Making large-scale svm learning practical. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.

[124] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[125] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[126] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Bull. Calcutta Math. Soc*, 1943.

[127] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[128] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[129] Stewart A Adcock and J Andrew McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, 106(5):1589–1615, 2006.

[130] Douglas L Theobald and Deborah S Wuttke. Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput Biol*, 4(2):e43, 2008.

[131] 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, 2008.

[132] Everett C Smith, Andreea Popa, Andres Chang, Cyril Masante, and Rebecca Ellis Dutch. Viral entry mechanisms: the increasing diversity of paramyxovirus entry. *Febs Journal*, 276(24):7217–7227, 2009.

[133] Thomas G Ksiazek, Paul A Rota, and Pierre E Rollin. A review of nipah and hendra viruses with an historical aside. *Virus research*, 162(1):173–183, 2011.

[134] Thomas A Bowden, E Yvonne Jones, and David I Stuart. Cells under siege: viral glycoprotein interactions at the cell surface. *Journal of structural biology*, 175(2):120–126, 2011.

[135] Benhur Lee and Zeynep Akyol Ataman. Modes of paramyxovirus fusion: a henipavirus perspective. *Trends in microbiology*, 19(8):389–399, 2011.

[136] Deborah L Steffen, Kai Xu, Dimitar B Nikolov, and Christopher C Broder. Henipavirus mediated membrane fusion, virus entry and targeted therapeutics. *Viruses*, 4(2):280–308, 2012.

[137] Thomas A Bowden, A Radu Aricescu, Robert JC Gilbert, Jonathan M Grimes, E Yvonne Jones, and David I Stuart. Structural basis of nipah and hendra virus attachment to their cell-surface receptor ephrin-b2. *Nature structural & molecular biology*, 15(6):567–572, 2008.

[138] Matteo Porotto, Feng Yi, Anne Moscona, and David A LaVan. Synthetic protocells interact with viral nanomachinery and inactivate pathogenic human virus. *PloS one*, 6(3):e16874, 2011.

[139] Hector C Aguilar, Vanessa Aspericueta, Lindsey R Robinson, Karen E Aanensen, and Benhur Lee. A quantitative and kinetic fusion protein-triggering assay can discern distinct steps in the nipah virus membrane fusion cascade. *Journal of virology*, 84(16):8033–8041, 2010.

[140] Hector C Aguilar, Zeynep Akyol Ataman, Vanessa Aspericueta, Angela Q Fang, Matthew Stroud, Oscar A Negrete, Richard A Kammerer, and Benhur Lee. A novel receptor-induced activation site in the nipah virus attachment glycoprotein (g) involved in triggering the fusion glycoprotein (f). *Journal of Biological Chemistry*, 284(3):1628–1635, 2009.

[141] Kai Xu, Kanagalaghatta R Rajashankar, Yee-Peng Chan, Juha P Himanen, Christopher C Broder, and Dimitar B Nikolov. Host cell recognition by the henipaviruses: crystal structures of the nipah g attachment glycoprotein and its complex with ephrinb3. *Proceedings of the National Academy of Sciences*, 105(29):9953–9958, 2008.

[142] Oscar A Negrete, Mike C Wolf, Hector C Aguilar, Sven Enterlein, Wei Wang, Elke Mühlberger, Stephen V Su, Andrea Bertolotti-Ciarlet, Ramon Flick, and Benhur Lee. Two key residues in ephrinb3 are critical for its use as an alternative receptor for nipah virus. *PLoS Pathog*, 2(2):e7, 2006.

[143] Qian Liu, Jacquelyn A Stone, Birgit Bradel-Tretheway, Jeffrey Dabundo, Javier A Benavides Montano, Jennifer Santos-Montanez, Scott B Biering, Anthony V Nicola, Ronald M Iorio, Xiaonan Lu, et al. Unraveling a three-step spatiotemporal mechanism of triggering of receptor-induced nipah virus fusion and cell entry. *PLoS Pathog*, 9(11):e1003770, 2013.

[144] Theodore S Jardetzky and Robert A Lamb. Activation of paramyxovirus membrane fusion and virus entry. *Current opinion in virology*, 5:24–33, 2014.

[145] Ronald M Iorio, Vanessa R Melanson, and Paul J Mahon. Glycoprotein interactions in paramyxovirus fusion. *Future virology*, 4(4):335–351, 2009.

[146] Kimberly A Bishop, Tzanko S Stantchev, Andrew C Hickey, Dimple Khetawat, Katharine N Bossart, Valery Krasnoperov, Parkash Gill, Yan Ru Feng, Lemin Wang, Bryan T Eaton, et al. Identification of hendra virus g glycoprotein residues that are critical for receptor binding. *Journal of virology*, 81(11):5893–5901, 2007.

[147] Rommie E Amaro, Anurag Sethi, Rebecca S Myers, V Jo Davisson, and Zaida A Luthey-Schulten. A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. *Biochemistry*, 46(8):2156–2173, 2007.

[148] Kevin C Wolfe and Gregory S Chirikjian. Quantitative comparison of conformational ensembles. *Entropy*, 14(2):213–232, 2012.

[149] KV Brinda and Saraswathi Vishveshwara. A network representation of protein structures: implications for protein stability. *Biophysical journal*, 89(6):4159–4170, 2005.

[150] Adam T Van Wart, Jacob Durrant, Lane Votapka, and Rommie E Amaro. Weighted implementation of suboptimal paths (wisp): an optimized algorithm and tool for dynamical network analysis. *Journal of chemical theory and computation*, 10(2):511–517, 2014.

[151] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[152] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research.

[153] Steve W Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

[154] Paolo De Los Rios, Fabio Cecconi, Anna Pretre, Giovanni Dietler, Olivier Michielin, Francesco Piazza, and Brice Juanico. Functional dynamics of pdz binding domains: a normal-mode analysis. *Biophysical journal*, 89(1):14–21, 2005.

[155] Ernesto J Fuentes, Steven A Gilmore, Randall V Mauldin, and Andrew L Lee. Evaluation of energetic and dynamic coupling networks in a pdz domain protein. *Journal of molecular biology*, 364(3):337–351, 2006.

[156] Stefano Gianni, Tine Walma, Alessandro Arcovito, Nicoletta Calosci, Andrea Bellelli, Åke Engström, Carlo Travaglini-Allocatelli, Maurizio Brunori, Per Jemth, and Geerten W Vuister. Demonstration of long-range interactions in a pdz domain by nmr, kinetics, and protein engineering. *Structure*, 14(12):1801–1809, 2006.

[157] Lieke CJ van den Berk, Elena Landi, Tine Walma, Geerten W Vuister, Luciana Dente, and Wiljan JAJ Hendriks. An allosteric intramolecular pdz- pdz interaction modulates ptp-bl pdz2 binding specificity. *Biochemistry*, 46(47):13629–13637, 2007.

[158] Z Nevin Gerek, Ozlem Keskin, and S Banu Ozkan. Identification of specificity and promiscuity of pdz domain interactions through their dynamic behavior. *Proteins: Structure, Function, and Bioinformatics*, 77(4):796–811, 2009.

[159] Jun Zhang, Paul J Sapienza, Hengming Ke, Aram Chang, Sarah R Hengel, Huanchen Wang, George N Phillips Jr, and Andrew L Lee. Crystallographic and nmr evaluation of the impact of peptide binding to the second pdz domain of ptp1e. *Biochemistry*, 49(43):9280, 2010.

[160] MS Vijayabaskar and Saraswathi Vishveshwara. Interaction energy based protein structure networks. *Biophysical journal*, 99(11):3704–3715, 2010.

[161] Elisa Cilia, Geerten W Vuister, and Tom Lenaerts. Accurate prediction of the dynamical changes within the second pdz domain of ptp1e. *PLoS Comput Biol*, 8(11):e1002794, 2012.

[162] Márton Münz, Jotun Hein, and Philip C Biggin. The role of flexibility and conformational selection in the binding promiscuity of pdz domains. *PLoS Comput Biol*, 8(11):e1002749, 2012.

[163] Brigitte Buchli, Steven A Waldauer, Reto Walser, Mateusz L Donten, Rolf Pfister, Nicolas Blöchliger, Sandra Steiner, Amedeo Caflisch, Oliver Zerbe, and Peter Hamm. Kinetic response of a photoperturbed allosteric protein. *Proceedings of the National Academy of Sciences*, 110(29):11725–11730, 2013.

[164] Sebastian Buchenberg, Volker Knecht, Reto Walser, Peter Hamm, and Gerhard Stock. Long-range conformational transition of a photoswitchable allosteric protein: molecular dynamics simulation study. *The Journal of Physical Chemistry B*, 118(47):13468–13476, 2014.

[165] Germán A Miño-Galaz. Allosteric communication pathways and thermal rectification in pdz-2 protein: A computational study. *The Journal of Physical Chemistry B*, 119(20):6179–6189, 2015.

[166] Todd J Dolinsky, Paul Czodrowski, Hui Li, Jens E Nielsen, Jan H Jensen, Gerhard Klebe, and Nathan A Baker. Pdb2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic acids research*, 35(suppl 2):W522–W525, 2007.

[167] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[168] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics*, 52(2):255–268, 1984.

[169] William G Hoover. Canonical dynamics: equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[170] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[171] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[172] Berk Hess. P-lincs: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation*, 4(1):116–122, 2008.

[173] Shuichi Miyamoto and Peter A Kollman. Settle: an analytical version of the shake and rattle algorithm for rigid water models. *Journal of computational chemistry*, 13(8):952–962, 1992.

[174] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.

[175] HJC Berendsen, JR Grigera, TP Straatsma, et al. The missing term in effective pair potentials. *J. phys. Chem*, 91(24):6269–6271, 1987.

[176] Giovanni Lipari and Attila Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *Journal of the American Chemical Society*, 104(17):4546–4559, 1982.

[177] Henrik Flyvbjerg and Henrik Gordon Petersen. Error estimates on averages of correlated data. *The Journal of Chemical Physics*, 91(1):461–466, 1989.

[178] Vladimir N Maiorov and Gordon M Crippen. Size-independent comparison of protein three-dimensional structures. *Proteins: Structure, Function, and Bioinformatics*, 22(3):273–283, 1995.

[179] Anne Dhulesia, Joerg Gsponer, and Michele Vendruscolo. Mapping of two networks of residues that exhibit structural and dynamical changes upon binding in a pdz domain protein. *Journal of the American Chemical Society*, 130(28):8931–8939, 2008.

APPENDICES

Title: Frameworks for Understanding Long-Range Intra-Protein Communication

Publication: Current Protein and Peptide Science

Publisher: Bentham Science Publishers

Date: Apr 1, 2017

Review Order

Please review the order details and the associated terms and conditions.

No royalties will be charged for this reuse request although you are required to obtain a licen
comply with the license terms and conditions. To obtain the license, click the Accept button b

Licensed Content Publisher	Bentham Science Publishers
Licensed Content Publication	Current Protein and Peptide Science
Licensed Content Title	Frameworks for Understanding Long-Range Intra-Protein Communication
Licensed Content Author	
Licensed Content Date	April 2017
Type of use	Thesis/Dissertation
Requestor type	Academic institution
Format	Print, Electronic
Portion	image/photo
Number of images/photos requested	1
Rights for	Main product
Duration of use	Life of current/future editions
Creation of copies for the disabled	no
With minor editing privileges	yes
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	0 to 499
The requesting person/organization	Mohsen Botlani

THE AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE LICENSE TERMS AND CONDITIONS

Jun 01, 2017

This Agreement between University of South Florida -- Mohsen Botlani ("You") and The American Association for the Advancement of Science ("The American Association for the Advancement of Science") consists of your license details and the terms and conditions provided by The American Association for the Advancement of Science and Copyright Clearance Center.

License Number	4120250549738
License date	Jun 01, 2017
Licensed Content Publisher	The American Association for the Advancement of Science
Licensed Content Publication	Science
Licensed Content Title	Sending Signals Dynamically
Licensed Content Author	Robert G. Smock,Lila M. Gierasch
Licensed Content Date	Apr 10, 2009
Licensed Content Volume	324
Licensed Content Issue	5924
Volume number	324
Issue number	5924
Type of Use	Thesis / Dissertation
Requestor type	Scientist/individual at a research institution
Format	Print and electronic
Portion	Figure
Number of figures/tables	1
Order reference number	
Title of your thesis / dissertation	Modeling Dynamic Allostery in Proteins Enabled by Machine Learning
Expected completion date	Jun 2017
Estimated size(pages)	100
Requestor Location	University of South Florida 4202 East Fowler Ave ISA 2015 TAMPA, FL 33620 United States Attn: Mohsen Botlani
Billing Type	Invoice
Billing Address	University of South Florida 4202 East Fowler Ave ISA 2015 TAMPA, FL 33620

Print This Page

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

Jun 01, 2017

This Agreement between University of South Florida -- Mohsen Botlani ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	4120251402813
License date	Jun 01, 2017
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature
Licensed Content Title	The ensemble nature of allostery
Licensed Content Author	Hesam N. Motlagh, James O. Wrabl, Jing Li, Vincent J. Hilser
Licensed Content Date	Apr 16, 2014
Licensed Content Volume	508
Licensed Content Issue	7496
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	The dynamic continuum of allosteric phenomena.
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Modeling Dynamic Allostery in Proteins Enabled by Machine Learning
Expected completion date	Jun 2017
Estimated size (number of pages)	100
Requestor Location	University of South Florida 4202 East Fowler Ave ISA 2015 TAMPA, FL 33620 United States Attn: Mohsen Botlani
Billing Type	Invoice
Billing Address	University of South Florida 4202 East Fowler Ave ISA 2015 TAMPA, FL 33620 United States Attn: Mohsen Botlani

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jun 01, 2017

This Agreement between University of South Florida -- Mohsen Botlani ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4120261344119
License date	Jun 01, 2017
Licensed Content Publisher	Elsevier
Licensed Content Publication	Journal of Molecular Biology
Licensed Content Title	Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues
Licensed Content Author	Ozlem Keskin,Buyong Ma,Ruth Nussinov
Licensed Content Date	Feb 4, 2005
Licensed Content Volume	345
Licensed Content Issue	5
Licensed Content Pages	14
Start Page	1281
End Page	1294
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Original figure numbers	figures 1 and 7
Title of your thesis/dissertation	Modeling Dynamic Allostery in Proteins Enabled by Machine Learning
Expected completion date	Jun 2017
Estimated size (number of pages)	100
Elsevier VAT number	GB 494 6272 12
Requestor Location	University of South Florida 4202 East Fowler Ave ISA 2015 TAMPA, FL 33620 United States Attn: Mohsen Botlani

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jun 01, 2017

This Agreement between University of South Florida -- Mohsen Botlani ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4120270237975
License date	Jun 01, 2017
Licensed Content Publisher	Elsevier
Licensed Content Publication	Biophysical Journal
Licensed Content Title	Stimulation of Nipah Fusion: Small Intradomain Changes Trigger Extensive Interdomain Rearrangements
Licensed Content Author	Priyanka Dutta,Ahnaf Siddiqui,Mohsen Botlani,Sameer Varma
Licensed Content Date	Oct 18, 2016
Licensed Content Volume	111
Licensed Content Issue	8
Licensed Content Pages	10
Start Page	1621
End Page	1630
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Original figure numbers	figure 3 and 8
Title of your thesis/dissertation	Modeling Dynamic Allostery in Proteins Enabled by Machine Learning
Expected completion date	Jun 2017
Estimated size (number of pages)	100
Elsevier VAT number	GB 494 6272 12
Requestor Location	University of South Florida 4202 East Fowler Ave ISA 2015 TAMPA, FL 33620 United States Attn: Mohsen Botlani

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Jun 01, 2017

This Agreement between University of South Florida -- Mohsen Botlani ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4120280827091
License date	Jun 01, 2017
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Proteins: Structure, Function and Bioinformatics
Licensed Content Title	Discerning intersecting fusion-activation pathways in the Nipah virus using machine learning
Licensed Content Author	Sameer Varma,Mohsen Botlani,Ralph E. Leighty
Licensed Content Date	Oct 9, 2014
Licensed Content Pages	14
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	3
Original Wiley figure/table number(s)	Figure 5,6 and 7
Will you be translating?	No
Title of your thesis / dissertation	Modeling Dynamic Allostery in Proteins Enabled by Machine Learning
Expected completion date	Jun 2017
Expected size (number of pages)	100
Requestor Location	University of South Florida 4202 East Fowler Ave ISA 2015 TAMPA, FL 33620 United States Attn: Mohsen Botlani
Publisher Tax ID	EU826007151
Billing Type	Invoice
Billing Address	University of South Florida 4202 East Fowler Ave ISA 2015 TAMPA, FL 33620 United States Attn: Mohsen Botlani
Total	0.00 USD

ABOUT THE AUTHOR

Mohsen Botlani.