Protein sequence-structure-dynamics-function relationships: The close association of dynamics with protein function

by

Sambit Kumar Mishra

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Robert L. Jernigan, Co-major Professor
Guang Song, Co-major Professor
Drena L. Dobbs
Mark S. Hargrove
Laura Jarboe

Iowa State University

Ames, Iowa

2018

DEDICATION

To my dear parents, my little brother and my loving wife for their unfailing faith and unconditional support

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS

The journey of past five and half years has been eventful and rewarding! These years have helped me evolve, in every aspect, into a promising researcher and more importantly, a sincere, dedicated and hardworking individual. The task I undertook in 2012 – that of pursuing my PhD in Bioinformatics and Computational Biology at Iowa State University, was challenging and required sheer motivation, patience and determination. It would not be an overstatement if I were to say that this phase of my life has tested me in every possible way. However, it was not solely on the basis of my effort that I successfully faced these challenges. On many occasions there were people who helped me, individuals who supported me and my family who backed me up!

I am sincerely thankful to Dr. Robert L. Jernigan, my PhD advisor, for his support and guidance. His suggestions related to research and also concerning my prospective scientific career have always been valuable. Through his unique mentoring style, he taught me to how to be an independent researcher and more importantly, the art of doing research in science. One of the important lessons which I learned under his guidance is that to make a research project interesting, it is often necessary to thoroughly scrutinize the results and pose more questions. The outcome of research is only as interesting as the questions it attempts to answer.

I would also like to thank my committee members Dr. Guang Song, Dr. Drena Dobbs, Dr. Mark Hargrove and Dr. Laura Jarboe for their support, guidance and valuable suggestions. I am extremely thankful to them for being a part of my committee and also for their questions and comments concerning my research, which has helped in making it fine-tuned.

I am thankful to my former and current lab members – Kannan, Jie, Ataur, Kejue, Sayane, Dan, Ambuj and Pranav for their suggestions to improve my work. Besides, I am thankful to all my fellow BCB students for their support.

Thanks to Trish Stauble whose incredible support through all these years has been remarkable! Thank you Trish for organizing all the BCB dinners, socials and also for our delightful discussions!

Thanks to my friends in Ames – Gaurav, Pulkit, Ashish, Akshay, Viraj, Gokul and Anindya. Your support has been incredible! I also wish to thank some of my friends in India who inspired me to pursue a PhD – Sachin, Gopi and Hemanth.

My foundation in science was laid by my teachers and mentors back in India. In this context, I would like to thank my tutors at Shri Jagannath Tutorials - Rupesh Sir, Rajesh Sir, Sunil Sir, Kumar Sir and all other staff. My thankfulness and gratitude also extends to all my primary school teachers at Blessed Sacrament High School. I also extend my gratefulness to my professors at my undergraduate alma mater, Bharath Institute of Higher Education and Research – especially, Swaminathan Sir, Vijay Sir and Subhashini Madam for inciting my interest and passion in Bioinformatics. I am also thankful to Dr. Sameer Hassan for being the supervisor of my undergraduate project and for his technical guidance.

My life has been a gift from my parents to whom I owe all my successes and achievements. Their support through all these years, their prayers and their faith in my abilities has been my strength. My little brother (he is still a little fellow to me despite growing up!!!) has been my constant companion in my moments of grief, struggle and a great friend during moments of joy. My wife, who recently stepped into my life, has been a blessing from my parents and The Almighty. Her presence and support has been motivating during the last phase of my PhD.

Last, but not the least, my gratefulness to the Almighty for being there with me all these years, guiding me in my moments of desperation and motivating me to tread the path of the virtuous.

ABSTRACT

The intrinsic dynamics of globular proteins is the key to the understanding of their function, being a consequence of protein structure and geometry. The view of protein structures has recently changed from native structures being considered to be a single rigid, static object into one where conformational ensembles coexist. Besides, allostery, the transmission of signals from a distant site to the active site, is a direct outcome of the detailed dynamics of a given protein. Investigating how dynamics controls protein function is one of the overall aims of our studies. It is essential to probe protein function by combining information from all three types of data: sequence, structure and dynamics, which combine to define their functions. The abundance of protein sequence data in repositories like UniProt and Pfam is huge and is strongly complementary to the rich data of protein structures in PDB. Exploiting this wealth of information and coupling it with molecular simulations that provide information on protein dynamics, facilitates the understanding and predicting of protein function, which is the underlying motivation and overall objective of the present work.

The dynamic behavior of proteins is often altered upon the binding of ligands, partner proteins or other biological macromolecules such as DNA and RNA. This work describes the influence of binding on the intrinsic dynamics of proteins through studies on homooligomeric protein assemblies which are comprised of multiple subunits of the same protein. Specifically, this work compares the dynamics of functionally important residues of a single subunit in isolation with those in its assembled form. Next, is presented a systematic investigation of the extent of similarity between the protein dynamic communities obtained from molecular dynamics with those from a simpler molecular simulation method, the elastic network models. The focus is on the separate dynamic communities, which are those groups of residues, highly cohesive in terms of their motions and which move like a rigid unit. Elastic network models are models for protein cohesion and are particularly appropriate for application to this task. We also show how they can effectively capture the differences in community distributions for mutant and wild type forms of T4 lysozyme. Finally, a machine learning classification method is developed wherein protein dynamics information is coupled with structure, evolutionary and physicochemical properties to predict regulatory and functional binding sites.

This work emphasizes the collective interplay between sequence, structure and dynamics as the key to the understanding of protein function. It also highlights the use of simplified molecular representations for simulations, i.e., the elastic network model, which can often be suitable as a substitute for atomic molecular dynamics. The machine learning models developed as a part of this work strongly point up the importance of including protein dynamics to improve predictions. The methods developed have potential practical applications, for instance as predictive models for identification of hot spot residues for site-directed mutagenesis or even for the prediction of sites where potential therapeutics could bind to restore dynamics and other disturbed functions, or even to suggest ways to generate new functions.

CHAPTER 1. INTRODUCTION

Proteins, the cellular nanomachines, manifest in diverse structural forms and participate in a multitude of functional roles. Some of these roles involve providing structural integrity to the cell with cytoskeletons, transporting molecules into and out of the cell with membrane transporters, facilitating intracellular signal transduction through G-proteins and even facilitating their own expression via RNA polymerase and ribosomes, and degradation with proteasome. The role of proteins is not just limited to the intracellular boundary. These molecules also facilitate intercellular communication in the form of hormones and are bulwarks of host defense mechanisms against pathogens.

The building blocks of all proteins are amino acids which associate with each other in a linear sequence of peptide bonds and later fold into three dimensional structures, referred to as the tertiary structure. The functional form of proteins is their tertiary structure. However, proteins can self-associate into oligomers or even form large macromolecular complexes to execute their function. These complex structural forms are referred to as protein quaternary structures. Based on their structural invariabilities proteins can be broadly classified into two types: fibrous and globular. Fibrous proteins have a rigid architecture, formed of repetitive linear strands of polypeptide units and are responsible for providing structural robustness (e.g., keratin and collagen). Globular proteins, on the other hand, are more dynamic and exhibit structural plasticity. They frequently undergo structural transitions and thus, exist in a continuum of conformational ensembles. Most functional proteins are globular by nature and the key to their function is the close association between their sequence, structure and dynamics.

This dissertation presents a set of computational approaches aimed at exploiting the available sequence and structure information for globular proteins to model their dynamics and understand their functional mechanisms.

1.1. Background

1.1.1. Paradigm Shift: From Protein Sequence-Structure-Function to Sequence-Structure-Dynamics-Function

Protein sequences have been used extensively to understand the extent of homology between organisms. Proteins showing considerable sequence similarity are usually considered as homologs that is, they evolve from the same ancestor and have similar or the same function. Even before the first protein structure was available, Linderstrom-Lang (Linderstrøm-Lang, 1952) proposed that the amino acid sequence of a protein can order into structurally discrete motifs (secondary structure elements), which would then fold into a more compact three dimensional structure. The connecting link between sequence and structure of proteins, which was debated for many decades, was established by Anfinsen through his experiment on ribonuclease A, revealing that the amino acid sequence is the key to a protein's native three dimensional structure (ANFINSEN & HABER, 1961). Later, Chothia and Lesk (Chothia & Lesk, 1986) showed that the amino acid sequence of proteins can be highly variable while, the protein structure is more conserved during evolution. Many examples of proteins showing weak sequence similarity but bearing strong structural resemblance and homology have since been reported (Rost, 1999; Sander & Schneider, 1991). These studies emphasized protein structure as the key determinant of its function and led to the inception of databases like CATH and SCOP (Hubbard, Ailey, Brenner, Murzin, & Chothia, 1999; Orengo et al., 1997) which assigned homology based on protein structure similarity instead of sequence similarity.

Proteins were previously viewed as static and rigid molecules, having well-defined structures and geometries. The lock-and-key hypothesis (Behr, 2007; Fischer, 1894) considered proteins having cavities (locks) of fixed shapes to which, only specific ligands (keys) bind. Such static representations of proteins were gradually replaced by a view that regards proteins as dynamic entities. Koshland introduced the induced fit model to explain the effect of ligand binding to a protein and suggested that binding can induce a change in shape of the cavity, aiding complementarity (Koshland, 1958). In quaternary structures, such a change induced in one subunit could be transmitted to adjacent subunits. A similar model was also proposed by Monod, Wyman and Changeux (Changeux & Edelstein, 2005) which assumes that subunits in proteins can either be in the tensed or relaxed state, the latter favoring ligand binding, wherein the subunits of quaternary structures are in the same conformational state (either tensed or relaxed) at any point in time. It was however, the availability of a large number of protein structures determined using advanced X-ray crystallography and NMR techniques that provided a clear picture about a protein’s dynamic state (Mannige, 2014). In stark contrast to the first crystallized structure of myoglobin (Kendrew et al., 1958), the structures of numerous proteins resolved thereafter showed disorder (missing regions), providing substantial evidence for the dynamic nature of proteins. Structures of proteins from same organisms were reported in multiple conformations further confirming protein flexibility and conformational variability (Damm & Carlson, 2007). To explain such conformational variability, sometimes even in the absence of ligands, the modern view considers the ensemble nature of protein structures.

According to the modern view, proteins don’t just exist as a single structure or conformation. The notion of protein structure from being referred to as a single structure or conformation has evolved into an ensemble one, in which a structure is an ensemble of conformations. Each conformation (or a conformer) is associated with a certain amount of energy and the ensemble can be outlined by a Boltzmann distribution of the conformational energies. Low energy conformations are more favorable than high energy ones. Proteins can frequently transition between conformations. The probability of such transitions strongly depends on the difference in energy between two conformational states; lesser the difference, more probable is the transition. Pathways for such conformation transitions are comprised of multiple intermediate conformers each associated with a certain amount of energy and inspire the notion of protein energy landscape (Dill, 1999) – a rugged contour having all possible conformers and their energies mapped between a starting and an ending conformation (Tavernelli, Cotesta, & Di Iorio, 2003).

The conformational ensemble model for proteins is frequently used to explain the effect of ligand binding or in general, any binding partner. Ligands can bind to pre-existing conformations, a process referred to as conformation selection, triggering a shift in the ensemble distribution towards the favorable binding state (Lindorff-Larsen, Best, DePristo, Dobson, & Vendruscolo, 2005; Vértessy & Orosz, 2011). Or they can even bind intrinsically disordered proteins and induce order and folding. The conformational plasticity and intrinsic dynamics of proteins is by virtue of their structure, sometimes referred to as "structure-encoded dynamics". Such conformational variability gives rise to promiscuous binding in which, a protein has multiple binding partners or can execute multiple functions, such as catalyzing reactions other than the ones they are evolved for (referred to as catalytic promiscuity or moonlighting). Examples of promiscuous binding proteins include antibodies such as immunoglobulin G, major histocompatibility class 1, aminoglycoside kinase, protein kinase A and aminoglycoside kinase (Chang, McLaughlin, Baron, Wang, & McCammon, 2008; James & Tawfik, 2009; Sundberg & Mariuzza, 2000) while, some proteins with moonlighting functions are gephyrin and cytochrome c (Copley, 2003). These works and many more, set the foundations for a strong association between protein dynamics and function and led to the paradigm shift, *sequence-structure-dynamics-function*.

1.1.2. Role of Dynamics in Regulating Protein Function

Conformation variability has been shown to strongly correlate with protein function (McCammon, 1984). Slow collective motions, also referred to as global motions, involve strong coordination between majority protein residues and have been shown to be evolutionarily conserved within many protein families (Haliloglu & Bahar, 2015), including the amino acid kinase family (Marcos, Crehuet, & Bahar, 2010). Investigations on mesophilic and thermophilic adenylate kinase reveal that the enzyme's catalytic efficiency can be greatly influenced by the dynamic opening/closing frequency the nucleotide-binding lids (Wolf-Watz et al., 2004). In another study for the same enzyme, local atomic fluctuations in hinge regions were shown to facilitate the collective opening/closing motions of the lid (K. A. Henzler-Wildman et al., 2007). Intrinsic protein flexibility also plays a key role in mediating ligand induced allosteric effects in which binding of effector molecules at an allosteric site can induce conformational change at a distant site. Examples of such conformational changes have been shown for myosin V and Hsp 70 (Coureux, Sweeney, & Houdusse, 2004; General et al., 2014). Other functional significance of large structural changes brought about by collective motions are described in (Grant, Gorfe, & Mccammon, 2010) and the coupling between enzyme catalytic site and collective dynamics in (L. W. Yang & Bahar, 2005) and (Z. Kurkcuoglu, Bakan, Kocaman, Bahar, & Doruker, 2012).

Given the emerging and underlying importance of dynamics in determining protein function, Hensen and co-workers (Hensen et al., 2012) introduced the novel concept of "protein dynasome" which hypothesizes that proteins having similar functions share similar dynamics and hence, a common dynamic fingerprint.

1.1.3. Molecular Simulations: Methods for Investigating Protein Dynamics

The term dynamics refers to any intrinsic motion within proteins. This includes motions at the atomic level as well as collective motions within subunits or even movement of domains with respect to one another. While the atomic motions occur at time scales of a few nanoseconds, collective motions such as domain motions occur in the microsecond or millisecond scale (K. Henzler-Wildman & Kern, 2007; Teilum, Olsen, & Kragelund, 2009). At the crux of every molecular simulation method is its potential, sometimes referred to as force field. Potentials are used to assess the net energy of a biological system and are broadly of two types: physics-based and knowledge-based. Physics-based potentials model inter-atomic interactions through harmonic springs and typically include bonded interactions (bond lengths, bond angles and dihedral angles) and non-bonded interactions (electrostatic and van der Waal interactions). The parameters for these potentials are derived from quantum mechanical (QM) calculations. Knowledge-based potentials or statistical potentials however, use the available protein structures to model interactions within a protein. Simply stated, such approaches count the frequency of contacts for different amino acid pairs and come up with a table that describes the favorable and unfavorable contact pairs. The methods available to study dynamics of molecular systems can be broadly divided into three categories. Each method differs from the other with respect to the underlying potential it uses to represent the protein structure and also with respect to how the simulation is carried out.

1.1.3.1. Molecular dynamics

Developed first by Fermi *et al*., (Fermi, Pasta, & Ulam, 1955) molecular dynamics (MD) simulations are one of the primary tools used to investigate the dynamics of biomolecular systems. The very first MD simulations on biological systems were performed by Levitt and Warshel (Levitt & Warshel, 1975), while the first MD simulation on “a macromolecule of biological interest” was on the bovine pancreatic trypsin inhibitor by Karplus and McCammon (McCammon, Gelin, & Karplus, 1977). A typical MD simulation takes into consideration two types of atomic forces that govern molecular motions: forces arising from the interactions of chemically bonded atoms and forces from non-bonded atoms. Consequently, an MD potential can be broken down into terms that describe bonded interactions (bond lengths, bond angles and dihedral angles) and non-bonded interactions (electrostatic and van der Waal interactions). In a typical MD run, the molecule of interest is first minimized with respect to the underlying potential. A standard MD simulation involves solving Newton’s equations of motion for the system of interacting molecules, in this case a protein. Each particle in the system is assigned an initial velocity and acceleration and then, the position of the particles after a short time *t* (a few femtoseconds) is calculated from Newton’s equations. Using the new positions for each particle, the force on each particle is calculated from the slope of the potential and from the force, the acceleration is calculated. This information is used to obtain the next positions of the particles, and so on. The procedure is repeated for a certain length of time, providing a trajectory which describes the dynamics of all the atoms of the system. Besides investigating protein dynamics, MD simulations have also been used to study protein folding (Scheraga, Khalili, & Liwo, 2007). Reviews by Karplus (Karplus & McCammon, 2002) and Durrant (Durrant & McCammon, 2011) shed light on some of the applications of MD.

1.1.3.2. Normal mode analysis

Normal mode analysis (NMA) is based on the vibrational motions of molecules and is used to characterize the motions of oscillating systems near the equilibrium state (Goldstein, Poole, & Safko, 2002). A normal mode is defined as a motion in which all particles in a system (in this case all atoms of a protein) move with the same frequency and in the same phase, sinusoidally. The underlying principle for such simulations for biological systems is that by virtue of its structure, a biological entity, such as a protein, can vibrate in certain ways. These vibrations represent accessible conformations for the protein and have been shown to strongly correlate with protein functional dynamics. The slow vibrating modes (ones with less frequency) are energetically more favorable than higher modes.

Classical NMA uses physics-based potentials which, similar to MD, have both bonded and non-bonded terms. The protein is first energy minimized with respect to this potential and then, a 3N by 3N dimensional Hessian matrix (N is the total number of atoms) is obtained by taking the second derivatives of the potential. The Hessian is a matrix of force constants describing the force on one atom owing to another. The eigen vectors of this matrix are referred to as normal modes. Each normal mode describes the displacement of atoms from their equilibrium position in X, Y and Z directions. The eigen values correspond to the square of mode frequency. Slower modes i.e., those having low frequency, describe collective motions such as inter-domain motions, while fast modes describe local motions as side chain vibrations. For motion in 3 dimensions, the first 6 modes correspond to rigid body motions and are usually overlooked.

1.1.3.3. Elastic network models

Elastic network models (ENMs) are simpler formulations of NMA which assume that a resolved protein structure is a local energy minimum thus, eliminating the requirement for energy minimization. Instead of modeling the system with a complex all-atom potential, ENM resorts to using the simple Hookean potential. In the ENM formulation, each atom is represented as a point mass and atoms within a certain cutoff distance are connected by Hookean springs. The total potential of this system is then the sum of the distance of separation between all atom pairs weighted by stiffness of the springs connecting them. The initial formulation of ENM was proposed by Tirion (Tirion, 1996) who showed that the theoretical B factors of residues calculated with ENM shows strong correlation with those obtained experimentally.

Gaussian network model (GNM) (I Bahar, Atilgan, & Erman, 1997; Rader, Chennubhotla, Yang, & Bahar, 2006), a more simplified version of ENM was later proposed by Bahar and co-workers. GNM assumed that residues fluctuations were isotropic by nature and followed a Gaussian distribution, representing proteins as coarse-grained systems of residue C^{α} atoms. GNM used an N x N Kirchhoff matrix instead of the 3N x 3N Hessian matrix. The off-diagonal elements of the Kirchhoff matrix denote the interacting residue pairs with -1 and non-interacting pairs with 0. The diagonal of the matrix gives the coordination number for each residue. Mean square fluctuations (MSF) of residues are given by the diagonal elements of the inverse of the Kirchhoff matrix while off diagonal elements correspond to residue cross-correlations. To account for the anisotropy in residue fluctuations, a theme not addressed by GNM, the anisotropic network model (ANM) was formulated. The ANM (Atilgan et al., 2001) provided positional displacement in X, Y and Z directions for each atom, described by the eigen vector (normal mode) of the 3N x 3N

Hessian matrix. The first 6 normal modes are usually eliminated as they correspond to rigid body motions. The inverse of the Hessian provides information about the residue mean square fluctuations and inter-residue cross-correlations. Excellent reviews on the working principles of ENMs and their applications are provided in the articles - (Ivet Bahar, Lezon, Bakan, & Shrivastava, 2010 and Timothy R. Lezon, Indira H. Shrivastava, 2009).

ENMs have two adjustable parameters: the distance cutoff r_c at which atoms are considered to be interacting and the spring constant γ describing the stiffness of springs connecting an atom pair. These parameters are chosen so as to best reproduce the experimental B factors. The distance cutoff r_c is usually set to 7Å for GNM and 13Å for ANM, while γ is set to 1 assuming uniform interaction strengths. Instead of representing amino acids just by their C^{α} atoms, other schemes of mixed coarse-grained representations have also been attempted (Doruker, Jernigan, & Bahar, 2002; O. Kurkcuoglu, Jernigan, & Doruker, 2004; Sinitskiy, Saunders, & Voth, 2012; Tozzini, 2005; Zhang et al., 2008). Several formulations of ENMs which take into consideration the different types of interactions and assign different γ based on the interaction type have been proposed (Jeong, Jang, & Kim, 2006; Kim et al., 2013). Besides, another popular version of ENM is the parameter free ENM (pfENM) that weighs the stiffness of springs as $\left(\frac{1}{r_c}\right)^k$ (L. Yang, Song, & Jernigan, 2009).

Owing to their reduced nature, it is easier to formulate and implement ENMs than classical NMA and MD. Also, the underlying assumption of ENM that a protein structure resolved using experimental methods such as X-ray crystallography and NMR is a local energetic minimum excludes the requirement for energy minimization, often a time consuming process. Though formulated in different forms, the underlying principle of all

ENMs is the same: that protein dynamics is a consequence of the structure and geometry of proteins. The conformational transitions that a protein can make are in accordance with its geometry. This principle has been extensively used to study the functional dynamics of diverse biomolecular systems, some of which even include large macromolecular assemblies like virus capsids (Tama & Brooks, 2006) and ribosomes (Kurkcuoglu, Doruker, & Jernigan, 2009; Wang, Rader, Bahar, & Jernigan, 2004). Studies using ENMs have also been carried out to bridge the gap between structure, dynamics and function for tryptophan synthase (I Bahar & Jernigan, 1999) and protein-dna and protein-rna complexes (O. Kurkcuoglu, Turgut, Cansu, Jernigan, & Doruker, 2009). Other applications and development of ENMs have been described in the review article by Lopez-Blanco (López-Blanco & Chacón, 2016). Importantly, the dynamics obtained using the simple ENM approach corresponds strongly to the dynamics obtained with sophisticated methods such as MD and NMA.

1.1.4. Investigating Dynamics using Structure Ensembles

Another strategy for obtaining knowledge about the intrinsic protein dynamics is by principal component analysis (PCA) of protein structure ensembles. This requires the availability of either a dynamic simulation trajectory of a protein (can be obtained from MD) or even an ensemble of structures obtained from a repository such as PDB. The underlying principle of such an approach is to have diverse conformational states of a protein and then identify the dominant conformational changes (Howe, 2001).

First a structural alignment is performed for all members of the ensemble by superimposing each conformational state on a representative structure. This is done to transform structures from different coordinate frames to a single frame. Then, based on either multiple sequence alignment or multiple structure alignment, only the core subset of residues from each structure is retained (Leo-Macias, Lopez-Romero, Lupyan, Zerbino, & Ortiz, 2005). These residues correspond to those positions that are not aligned to any gaps. A covariance matrix is then created from the 3D coordinate information of the core residues. Eigen decomposition of this covariance matrix gives principal components (the eigen vectors) and the variance explained by each component (eigen value). Each principal component explains the positional fluctuation of residues in the X, Y and Z dimensions and the associated eigen value tells the extent to which that motion dominates the structure ensemble (larger eigen values imply greater importance). Dynamics obtained using such ensemble approaches are often referred to as the essential dynamics of proteins (Amadei, Linssen, & Berendsen, 1993). Previous works show strong association between the dynamics uncovered using PCA and other methods such as MD and ENM (L. Yang, Song, Carriquiry, & Jernigan, 2008). Other works using PCA to understand global dynamics of protein structures include investigations on GroEL subunit (Skjaerven, Martinez, & Reuter, 2011) and on λ-repressor mutants (Maisuradze & Leitner, 2006).

1.1.5. Data-driven Approaches for Understanding Protein Functions

The protein databank (PDB) (Berman et al., 2000) currently holds 138878 structures of which 128935 are protein structures and 6720 are protein-nucleic acid complexes. The sequence data available for proteins is many folds greater than the structure data. Such a wealth of data can be used to carry out large scale computational studies to provide general conclusions. Data-driven approaches such as machine learning (ML) can be greatly exploited to identify hotspots of functional residues and make predictions for protein functions using structural data. A number of machine learning methods exist till date and can range from the simple linear regression to more sophisticated random forests and deep learning methods.

Machine learning models predict a target property (response variable) using a set of pre-calculated features (predictors) considered to be important for the prediction (Kohavi & Provost, 1998). Based on the type of predictions they make, ML models may be categorized as classification models or regression models. Classification models are used when the response variable is categorical such as predicting enzyme class, oligomeric state and so on. Regression models are used for response variables that are continuous such as predicting the catalytic efficiency of an enzyme, enzyme half-life and so on. ML models can also be classified into supervised and unsupervised learning models, based on how they are trained. Supervised learning models are usually trained on a single or multiple datasets and are used to make predictions on test data (data which the model has not seen before). The data used for these models are usually labelled i.e., the correct value of the response variable is known. Unsupervised learning models do not require training and are used to make predictions on unlabeled data, where the exact labels are not known (Hastie, Tibshirani, Friedman, & others, 2009). Common examples of supervised learning algorithms include generalized linear models, logistic regression, decision trees and random forests while, cluster detection is a good example of unsupervised learning. A review on the available supervised learning methods are provided by Kotsiantis, 2007 and Ng, 2012.

The applications of machine learning to problems in computational biology are numerous. Machine learning has been used for gene predictions, protein secondary structure predictions and even for predicting protein-protein interactions with considerable success. In systems biology, they have been used in predicting signaling networks and metabolic pathways (Dale, Popescu, & Karp, 2010). They have also been used to predict regulatory (Demerdash, Daily, & Mitchell, 2009) and active sites (Petrova & Wu, 2006) in proteins and also in protein tertiary structure predictions (Cheng, Tegge, & Baldi, 2008). A more elaborate discussion on the applications of machine learning to bioinformatics is presented by Larrañaga et al. 2006.

1.2. Specific Aims

The availability of diverse structure and sequence data lays the foundation for using data driven methods such as machine learning in addition to computer aided simulations to get deeper insight into the working principles of proteins. Biology continues to remain a largely unexplored field and there remain quite a few unsolved, unaddressed problems. Solving each problem through an experiment is often inconvenient owing to the lack of available techniques or the expense at which they are available. Computational simulations have two advantages in this context: first they take relatively lesser time and expenditure to solve a given problem and second, previous results have shown strong correlation between results from computational simulations and experiments.

In this dissertation, I address three specific aims, each referring to an existing biological problem, and I outline the computational schemes to solve these problems. A common theme connecting each problem is the essence of dynamics in the context of protein function.

Aim 1. Understanding the Effects of Oligomerization on Intrinsic Protein Dynamics

As intrinsic dynamics of proteins can be greatly influenced by protein structure, it is necessary to understand how the dynamics of oligomeric assemblies differ from their monomeric forms. Though previous studies have investigated the effect of oligomerization for specific proteins, a large scale study on diverse oligomeric states is yet to be carried out. We compile a diverse set of oligomeric proteins and compare their dynamics for the monomeric and oligomeric forms, using elastic network models. For the same set of proteins we also verify the effect of oligomerization on the dynamics of key functional residues (those which are evolutionarily conserved). Using the specific case of triosephosphate isomerase, we investigate changes to the highly correlated parts in the monomeric form (dynamic communities) upon oligomerization.

Aim 2. A Simpler Method for Mining Protein Dynamic Communities

Dynamic communities are the highly correlated parts of a protein whose residues are highly cohesive and exhibit rigid body motions. Identifying these communities is critical for the understanding of functional mechanism of proteins. Typically, MD simulations (~ 100 ns or longer) have been used to identify these communities (McClendon, Kornev, Gilson, & Taylor, 2014). We show that the communities obtained from elastic network models closely correspond to the communities from MD. Our study also reveals that atomic formulations of ENMs can be used to distinguish between deleterious and stable mutants for a protein.

Aim 3. Predicting Regulatory and Active Site Residues

Identification of allosteric and active site residues in proteins is a widely acknowledged important biological problem. Numerous prediction schemes have been implemented for the prediction of allosteric and active site residues (Lu, Huang, & Zhang, 2014; Sankararaman, Sha, Kirsch, Jordan, & Sjölander, 2010; Singh, Biswas, & Jayaram, 2011). However, there is a disjoint between most allosteric and active site prediction methods in that they are usually trained on separate datasets and make no definite connection between the two categories. To address this problem, we use a dataset that has both allosteric and active site residues labelled. Also, given the underlying importance of dynamics for protein function prediction, we hypothesize that incorporating information on protein dynamics should in principle lead to improved detection of allosteric and active site residues. We develop separate machine learning models for the prediction of allosteric and active site residues using a common subset of features which involve dynamic, evolutionary, physicochemical and structural information. We also perform comparisons with existing methods to verify our method's performance and also compare the predictions for active and allosteric residues for a single protein, establishing a close relationship between the predictions.

1.3. Dissertation Organization

This dissertation is comprised of six chapters. The background to this work was provided in **Chapter 1**. The content of the other chapters are briefly summarized below.

Chapter 2 relates to specific aim 1 with the underlying objective to understand the effects of oligomerization on the intrinsic protein dynamics and its consequences for functionally important residues. This chapter has been published in a peer reviewed journal under the title "*Altered Dynamics upon Oligomerization Corresponds to Key Functional Sites*" by *Sambit Kumar Mishra, Kannan Sankar and Robert L. Jernigan* in *Proteins Struct. Funct. Bioinformatics 85(8), April 2017.*

Chapter 3 corresponds to specific aim 2 where an exhaustive study that compares dynamical communities from MD with ENM has been carried out. It has been formatted as a manuscript and submitted with the title "*Protein Dynamic Communities from Elastic Network Models Align Closely to the Communities Defined by Molecular Dynamics*" by *Sambit K. Mishra* and *Robert L. Jernigan* to *PLoS One* and is currently under review.

Chapter 4 implements machine learning classifiers that consider protein dynamic information along with protein structure, evolutionary and physicochemical properties to predict allosteric and active site residues. The paper is currently being formatted into a manuscript for submission to a peer reviewed journal.

Chapter 5 compares different formulations of ENMs and their performance in terms of predicting experimentally determined conformational dynamics. It has been published under the title "*Comparisons of Protein Dynamics from Experimental Structure Ensembles, Molecular Dynamics Ensembles, and Coarse-Grained Elastic Network Models*" by *Kannan Sankar, Sambit K Mishra and Robert L. Jernigan in J. Phys. Chem. B, January 2018*.

Chapter 6 summarizes the overall findings in the dissertation and suggests the scope and improvements that can be made to the dissertation in future work.

Appendices A, B, C and D contain the supplemental information for Chapters 2, 3, 4 and 5, respectively.

CHAPTER 2. ALTERED DYNAMICS UPON OLIGOMERIZATION CORRESPONDS TO KEY FUNCTIONAL SITES

Paper published in Proteins 2017 85(8):1422-1434.

Sambit Kumar Mishra[1,2], Kannan Sankar[1,2], Robert L. Jernigan[1,2]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa 50011, USA

[2]Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011, USA

Abstract

It is known that over half of the proteins encoded by most organisms function as oligomeric complexes. Oligomerization confers structural stability and dynamics changes in proteins. We investigate the effects of oligomerization on protein dynamics and its functional significance for a set of 145 multimeric proteins. Using coarse-grained elastic network models, we inspect the changes in residue fluctuations upon oligomerization and then compare with residue conservation scores to identify the functional significance of these changes. Our study reveals conservation of about ½ of the fluctuations, with ¼ of the residues increasing in their mobilities and ¼ having reduced fluctuations. The residues with dampened fluctuations are evolutionarily more conserved and can serve as orthosteric binding sites, indicating their importance. We also use triosephosphate isomerase as a test case to understand why certain enzymes function only in their oligomeric forms despite the monomer including all required catalytic residues. To this end, we compare the residue

communities (groups of residues which are highly correlated in their fluctuations) in the monomeric and dimeric forms of the enzyme. We observe significant changes to the dynamical community architecture of the catalytic core of this enzyme. This relates to its functional mechanism and is seen only in the oligomeric form of the protein, answering why proteins are oligomeric structures.

2.1. Introduction

Proteins are critical for diverse cellular functions, including structural integrity, transport, and catalysis of biochemical reactions. Some function as independent monomeric units and others in multimers, or even form large biological complexes. The process of forming oligomers, oligomerization, often confers increased stability and the ability to perform complex functions (Ali & Imperiali, 2005; Marianayagam, Sunde, & Matthews, 2004). Oligomers can exist either as an assembly of identical subunits, homo-oligomers, or can combine in a mosaic of hetero-oligomers. Previous work reveals that homo-oligomers often tend to display structural symmetry that is generally associated with greater stability and robustness (Goodsell & Olson, 2000; Healy, 2015). Apart from their specific architecture, oligomers can also be classified based on whether or not complexation is required for their biological activity. Obligate cases require oligomerization in order to execute their functions, while non-obligate oligomers are transient complexes with the subunits capable of performing their functions in isolation (Griffin & Gerrard, 2012).

Oligomeric complexes can perform complex functions, a role often not possible for monomers. For example, the homo-oligomeric complexes Hsp90 and calreticulin play significant roles in affecting protein folding (Matthews & Sunde, 2012); the oligomeric forms of these proteins are known to bind misfolded proteins with higher affinity than their monomeric counterparts. Moreover, most oligomeric complexes exhibit longer-range allosteric regulation than in the monomer, which can be important for signal transduction (Ali & Imperiali, 2005; Changeux & Edelstein, 2005). Hemoglobin is a classic example that has been investigated frequently to elucidate aspects of allostery and cooperativity with respect to protein oligomerization. Also, the increased stability of protein complexes by oligomerization is an essential modification for thermophiles to prevent their dissociation under extreme temperatures (Walden et al., 2001)

The dynamics of individual monomers persist in most oligomeric assemblies. However, some complexes can develop novel dynamics after oligomerization, especially when some critical motions are not accessible to the monomeric form. Previously Voth *et al.*, (G. Song, Doruker, Jernigan, Kurkcuoglu, & Yang, 2008), showed that the dimeric form of triosephosphate isomerase was required to obtain appropriate motions of the closing loop, while the monomer does not show such motions. Bahar *et al.* (Marcos, Crehuet, & Bahar, 2011) investigated the low frequency normal modes accessible to an individual subunit of amino acid kinases in the monomeric and oligomeric forms and proposed that changes to the dynamics upon oligomerization facilitate allostery and ligand binding. A molecular dynamics simulation of tryptophan synthase revealed that in its monomeric form the enzyme is more rigid and cannot undergo conformational transitions that are seen after oligomerization (Qaiser Fatmi & Chang, 2010). In addition, oligomerization is known to increase the catalytic efficiency of this enzyme in contrast to the isolated monomer. In another study, we reported a similar finding where the functional loops of triosephosphate isomerase preserve their dynamics in both natively dimeric and natively tetrameric forms (Katebi & Jernigan, 2014).

The conformational flexibilities of globular proteins have often been considered to be a central factor for their function (Ivet Bahar, Lezon, Yang, & Eyal, 2010; Goodman, Pagel, & Stone, 2000; Henzler-Wildman & Kern, 2007). Soft modes from elastic network models have frequently been used to predict energetically favorable conformational changes upon substrate binding, and these predictions bear a strong similarity to the different experimentally resolved structures (Haliloglu & Bahar, 2015). Previous studies indicated strong correlations between dynamic flexibility and conservation levels of amino acids, with the most conserved residues showing the smallest fluctuations. These studies emphasized the significance of regions having high packing density, low mobility and low solvent accessibility by their high level of conservation; this also underscores how important probing conformational dynamics is to decipher protein function (Liao, Yeh, Chiang, Jernigan, & Lustig, 2005; Liu & Bahar, 2012; Marsh & Teichmann, 2014). These studies, however, have assigned functional significance based on residue flexibility in the native protein structure. For a native oligomeric protein, subunits of an assembly will exhibit different residue flexibility profiles when in isolation than when in the assembly owing to the differences in packing densities. A comparative study on residue flexibilities in the monomeric and oligomeric forms of a protein was not previously carried out for a diverse set of proteins - the aim of the present study, which will inform us about the importance of oligomerization for functional sites.

To understand the changes in dynamics that oligomerization introduces, we investigate a diverse set of 145 homo-oligomers with oligomeric states ranging from two (homo-dimer) to six (homo-hexamer). For each protein, we compute the change in mean square fluctuations (MSFs) of all residues in the monomer upon oligomerization. We then compare the residue conservation profiles of each protein with the MSF changes to ascribe functional significance to the changes in dynamics. We limit this study to a consideration of only homooligomers, owing to their greater abundance. We investigate the specific cases for four enzymes: glutamate dehydrogenase, arginase 1, glycine N-methyltransferase and D-amino acid oxidase to probe the functional importance of regions showing altered dynamics and then, provide more general results that associate changes in dynamics with functional significance. Our study reveals the importance of regions with dampened fluctuations following oligomerization. Using the specific cases of the four enzymes, we further confirm that the residues in regions with dampened mobilities often play a key role in the catalytic activity of the enzyme and hence, are orthosteric by nature. In the final section, we also address the question of why certain enzymes function only in their oligomeric state with triosephosphate isomerase (TIM) as a case study. Specifically we compare the residue communities (blocks of residues which are most highly correlated in their motions) for the monomeric and oligomeric forms of TIM. We observe a substantial shift in the community architecture of the catalytic core in the oligomer, the fundamental characteristic change necessary for the enzyme's activity, and a further change upon substrate binding.

2.2. Materials and Methods

2.2.1. Protein Structures

The initial dataset comprises Protein Data Bank (PDB) files of 174 different homooligomers downloaded from PDB. For each protein, the number of subunits in its functional quaternary state (biological assembly) ranges from two to six. We identify the biological assembly for each protein based on the assignment made by the authors and software in the PDB entry of the protein.

2.2.2. Homolog Selection and Multiple Sequence Alignment

For each protein in the initial dataset, we extract the sequence corresponding to a single chain (by default, we consider just the first chain) in the PDB file. We refer to these as query sequences. For each query sequence we search for homologous sequences using BLAST against the non-redundant protein sequence database with an e-value cutoff of 0.01, percentage identity in the range of ≥ 35% and ≤ 95% and query coverage of 80%. To filter duplicates, we then cluster the initial set of homologs with CD-Hit (Huang, Niu, Gao, Fu, & Li, 2010) at 95% sequence identity and then select only the representative sequences from each cluster. Our final dataset has 145 symmetric homooligomeric proteins (Supporting Information file ds145.xlsx), each having a minimum of 50 representative hits from BLAST. The diversity of the dataset in terms of oligomeric state and residues is depicted in Figure A.1 (A and B).

We then perform Multiple Sequence Alignment (MSA) for the representative homologs collected for each protein with Clustal Omega (Sievers & Higgins, 2014) with its default parameters.

2.2.3. Conservation Scores

Using Rate4Site (Pupko, Bell, Mayrose, & Glaser, 2002) with its default parameters for the evolutionary model (JTT) and rate inference method (Bayesian), we calculate the conservation scores for each protein from its respective MSA file. Rate4Site reports the extent of conservation at a position as a z-score, where a lower score indicates higher conservation.

2.2.4. Mean Square Fluctuations (MSF) from Elastic Network Model (ENM)

The fluctuations derived from ENM show remarkable agreement with the experimental fluctuations in B-factors (Atilgan et al., 2001; I Bahar, Atilgan, & Erman, 1997; Tirion, 1996). Here, we use the Anisotropic Network Model (ANM) (Atilgan et al., 2001) to study the protein dynamics. We model individual proteins as coarse-grained elastic networks by representing each residue by its C^{α} atom and connecting residue pairs by harmonic springs. In equilibrium, the potential of this system is given as

$$V = \frac{1}{2}\, \Delta R^{T} H \Delta R \tag{2.1}$$

Here, ΔR is the vector of change in position for all residues, ΔR^{T} is the transpose of this vector and H is a $3N$ by $3N$-dimensional Hessian matrix that has the second derivatives of the potential function. We vary the strength of spring γ between a residue pair by the inverse of their separation distance (d_{ij}), given by the following equation.

$$\gamma = \left(\frac{1}{d_{ij}}\right)^{a} \tag{2.2}$$

Diagonalizing the Hessian matrix results in 3N-6 modes (V) and eigen values (λ) which correspond to the non-rigid body dynamics of the system and we use these to calculate the MSF of residues with the following equation.

$$< \Delta R_i^2 > = \frac{3K_B T}{\gamma} \sum_{j=1}^{3N-6} \frac{1}{\lambda_j} \sum_{i=3k-2}^{3k} V_{ji}^2 \tag{2.3}$$

Here, K_B is the Boltzmann constant and T (set to 300) is the temperature in Kelvin. We then compute the theoretical B-factors ($Bfactor^{MSF}$) from these mean-square fluctuations (MSFs) as

$$B_i = \frac{8\pi^2 < \Delta R_i^2 >}{3} \tag{2.4}$$

and use them to describe residue positional fluctuations. We set a to 3 as it gives the highest median correlation with the experimental B-factors (Figure A.1.C).

2.2.5. MSF of Monomer and Oligomer

We use an approach similar to that of Bahar (Marcos et al., 2011) and Chang (Qaiser Fatmi & Chang, 2010) to obtain a protein's monomeric form from the oligomeric assembly. For each protein, we extract only the first chain from the PDB file and consider it to be the isolated monomer ($Monomer^{isolated}$) and when the same chain is in the oligomeric assembly, we refer to it as $Monomer^{oligomer}$. Comparing the fluctuation profiles of $Monomer^{isolated}$and $Monomer^{oligomer}$ will give us insight into the changes in dynamics after oligomerization. As the dataset comprises only symmetric proteins, we assume that there is a high overlap in the dynamics of individual chains in the oligomer and thus, we proceed with monitoring the change in dynamics for the first chain only.

We calculate the $Bfactor^{MSF}$vectors for the isolated monomer and oligomer of each protein and refer to these as $Bfactor^{MSF}_{Monomer^{isolated}}$ and $Bfactor^{MSF}_{Oligomer}$ respectively. We then consider the MSF values of only the first chain of the oligomer to study the fluctuation profile of the monomer in the oligomer.

2.2.6. Z-score Transformation of Raw MSF and Fold Changes

For individual proteins, we standardize the raw fluctuation values obtained in the $Bfactor^{MSF}_{oligomer}$ and $Bfactor^{MSF}_{Monomer^{isolated}}$ vectors by converting them to *z*-scores. Transforming the raw scores into z-score helps express both vectors on the same scale, i.e. the number of standard deviations the fluctuation of a given residue is from the mean fluctuation value over all residues. It also helps eliminate any potential bias that may be introduced due to the difference in the number of residues in the $Monomer^{isolated}$and the $Monomer^{oligomer}$. From the standardized $Bfactor_{oligomer}^{MSF}$ and $Monomer^{isolated}$ we obtain the standardized scores for the monomer in assembly and in isolation respectively. We refer to these standardized vectors as $Z_{Monomer^{isolated}}$ and $Z_{Monomer^{oligomer}}$. We then convert these vectors into a positive scale as follows:

$$Z^N{}_{Monomer^{isolated}} = Z^N{}_{Monomer^{isolated}} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} . \left| \min\left(Z^N{}_{Monomer^{isolated}}, Z^N{}_{Monomer^{oligomer}}\right) \right| \quad (2.5)$$

$$Z^N{}_{Monomer^{oligomer}} = Z^N{}_{Monomer^{oligomer}} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} . \left| \min\left(Z^N{}_{Monomer^{isolated}}, Z^N{}_{Monomer^{oligomer}}\right) \right| \quad (2.6)$$

The *min* function takes the minimum of the two vectors. We then define *Fold Change Ratio (FCR)* as the ratio of the *z*-scores of the monomer in assembly to the z-scores of the monomer in isolation.

$$Fold\ Change\ Ratio\ (FCR) = \frac{Z^N{}_{Monomer^{oligomer}}}{Z^N{}_{Monomer^{isolated}}} \quad (2.7)$$

To identify residues with significant increases or decreases in fluctuations, we use a cutoff of 1.5 FCR. A FCR greater than or equal to 1.5 indicates that the residue shows increased fluctuations upon oligomerization whereas, a FCR less than or equal to $\frac{1}{1.5}$ suggests a significant reduction in fluctuations after oligomerization. As the problem of finding residues with significant change in fluctuations has some similarity to the problem of identifying differentially expressed genes in RNA-Seq and microarray assays, we proceed with the cutoff of 1.5 which was shown to provide significant results for those types of experiments (Dalman, Deeter, Nimishakavi, & Duan, 2012; Tibshirani, 2007).

2.2.7. Identifying Interface Residues

For a particular chain, we identify interface residues as those whose heavy atoms are within 4.5Å from the atoms of residues from any of the other subunits (Bordner & Abagyan, 2005; Ofran & Rost, 2003).

2.2.8. Packing Density Calculations

Residue level packing densities are computed from the atomic structure of each protein in the dataset. The packing density values are obtained using the software Voronoia (Rother, Hildebrand, Goede, Gruening, & Preissner, 2009).

2.2.9. Residue Community Analysis

We define residue communities as groups of residues which are highly correlated in their fluctuations and exhibit motion as rigid units. We perform community analysis for the monomer of TIM for 4 cases: the isolated monomer without substrate, the isolated monomer with substrate, the monomer in the context of the dimer without the substrate and the monomer as part of the dimer with substrate. We use the PDB 1tph as the substrate bound form and 8tim as the unbound form. For the substrate bound form of TIM (PDB 1tph) we coarse-grain the protein at the C^{α} level while, retaining the substrate in its all-atom form. We set the exponent for spring strength *a* to 2 as this gives high correlation with the experimental B Factors and model the interaction strength as given in Equation 2.2.

After diagonalization of the Hessian of this system, we use the first twenty low frequency modes to construct the inverse hessian matrix as follows.

$$H^{-1} = \sum_{i=1}^{20} \lambda_i^{-1} V_i V_i^T \qquad (2.8)$$

Here, V_i is the *i*th low frequency mode vector, V_i^T the transpose of V_i and λ_i is the corresponding eigenfrequency of this mode. The H^{-1} has dimensions 3N by 3N, N being the number of residues and gives the correlations between residue fluctuations in the x,y and z directions. Like the Hessian, the H^{-1} can also be viewed as an N-dimensional matrix of sub-elements, these having a dimension of 3 by 3. We then calculate the correlation between the fluctuations of residues *i* and *j* as

$$c_{ij} = \frac{trace\ H_{ij}^{-1}}{\sqrt{trace\ H_{ii}^{-1}\ trace\ H_{jj}^{-1}}} \qquad (2.9)$$

In the above equation, H_{ij}^{-1} is a 3 by 3 block element of the inverse Hessian corresponding to residues *i* and *j* and it gives the correlation between the fluctutions of residues *i* and *j* in the x,y and z directions. H_{ii}^{-1} and H_{jj}^{-1} are the block elements corresponding to the self-correlations of residues *i* and *j*. The trace is the sum of the diagonal elements of each block matrix. In taking the trace of the block matrices, we are only accounting for the correlations of residue fluctuations in the same directions. Performing the above operation results in an N-dimensional symmetric correlation matrix, C.

We then express the above correlation matrix as a dissimilarity matrix by subtracting each element of C from 1.

$$c_{ij}^{dissimilarity} = 1 - c_{ij} \qquad (2.10)$$

Hierarchical clustering of the dissimilarity matrix with complete linkage then yields a dendrogram, grouping residues which are correlated to similar extents in their motions. We cut the dendrograms for the ligand free form (8tim) and ligand bound form (1tph) of TIM at manually selected heights to generate two, three and four clusters and then map these clusters onto the structure for comparisons. To perform hierarchical clustering and generate the dendrograms, we use the MATLAB clustering module (https://www.mathworks.com/help/ /stats/hierarchical-clustering.html).

2.2.10. Probability Distribution Fit

We fit the residue conservation and packing density data to different distributions using the MATLAB function allfitdist (https://www.mathworks.com/matlabcentral/fileexchange/34943-fit-all-valid-parametric-probability-distributions-to-data/content/allfitdist.m).

2.2.11. Non-parametric Test of Significance

We perform the non-parameteric Kruskal-Wallis test to evaluate the significance of residue conservation scores for different levels of MSF change using the MATLAB kruskalwallis (Kruskal & Wallis, 1952) function.

2.2.12. Protein Structure Visualization and Mapping of Critical Residues onto Structures

We use Pymol to map and visualize the key functional residues and clusters on the protein structure (DeLano, 2002).

2.3. Results

Our dataset includes 145 different homo-oligomeric proteins having between two (homo-dimer) and six (homo-hexamer) subunits. For each protein, we choose a single subunit (the first chain from the PDB file) to represent its monomeric form. This method of using a single subunit from the oligomeric protein assembly to represent the isolated monomer is similar to the approach taken by Bahar (Marcos et al., 2011) and Chang (Qaiser

Fatmi & Chang, 2010). We also verify the reliability of this approach by considering the case of protein tyrosine phosphatase that has been crystallized in both monomeric (PDB 1L8G) and oligomeric (PDB 2CM3) forms. Comparison of the dynamics of the crystallized monomer with that of the monomer extracted from the oligomer shows a strong correlation in the residue fluctuations for the two forms (Figure A.2), which further verifies this approach.

To investigate the effect of oligomerization on protein dynamics, for each protein, we compute the Mean Square Fluctuations (MSF) for residues when it is an isolated monomer and compare with the fluctuations when the same monomer is in its oligomeric assembly. Then, we simply look at the ratio of changes in the scalar mobilities (fold change ratio or FCR); with an arbitrary cutoff at 50% either reduced or increased, we identify residues that have undergone significant changes in their mobilities upon oligomerization. We attribute functional significance to the changes in mobilities by considering them together with the degree of conservation of residues, which we compute using Rate4Site (Pupko et al., 2002) Regions which are critical to the protein's function, such as catalytic sites, evolve more slowly and hence, are usually more conserved.

2.3.1. Influence of Oligomerization on Key Functional Residues

First we inspect the effect of oligomerization for four enzymes: bovine glutamate dehydrogenase (an enzyme known for its allosteric behavior), arginase 1 (a critical enzyme in the urea cycle), glycine N-methyltransferase (playing a critical role in methionine metabolism) and D-amino acid oxidase (oxidizes D amino acids and enables yeast to use D-amino acids for nutrition). For each, we identify from the literature those residues known to have functional significance and map them onto the protein structure to focus on the changes in fluctuations for these. Here, we address the fundamental question: does the mobility of the identified critical residues for a protein change significantly upon oligomerization? If it does, then do these functional residues undergo significant reductions or increases in their mobilities upon oligomerization?

2.3.1.1. Glutamate dehydrogenase

Glutamate dehydrogenase (GDH) plays a pivotal role in the metabolism of ammonia and is universal throughout most domains of life. It catalyzes the inter-conversion of L-glutamate into α-ketoglutarate and ammonia. In mammals, enriched GDH activity is found in liver, kidney, brain and pancreas, and the ammonia produced from glutamate is utilized in the urea cycle (Peterson & Smith, 1999). GDH in mammals exists as a homohexamer with

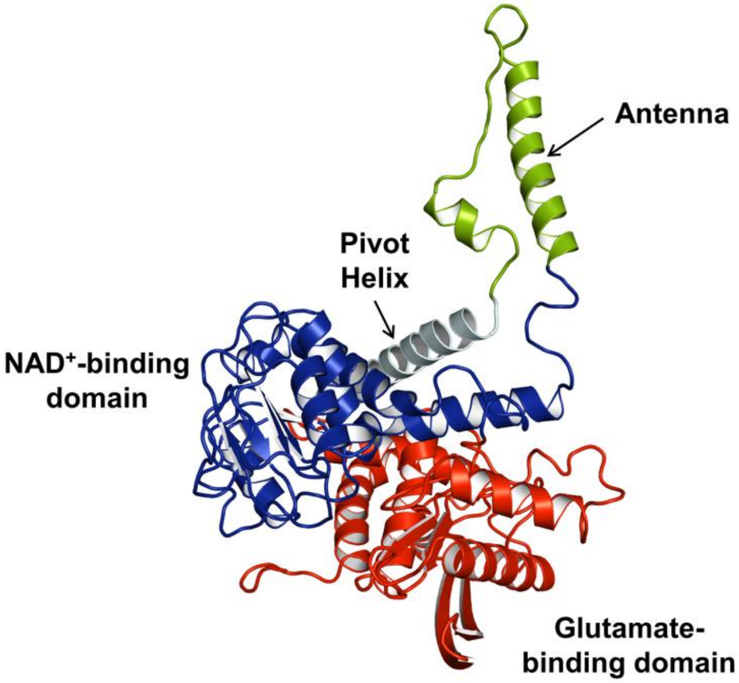

Figure 2.1. Domains and structural aspects of bovine glutamate dehydrogenase (GDH). The mobility of the NAD+-binding domain mediates allostery in the enzyme. Glutamate-binding domain is responsible for binding the substrate glutamate. The antenna feature is unique only to animal GDH and is also hypothesized to play some role in the allosteric behavior of the protein. Table A.1 provides details of the functionally significant residues and their roles.

dihedral symmetry and is comprised of about 500 residues. It has two structural domains: the NAD$^+$-binding domain where, the coenzyme NADH binds and the glutamate-binding domain where the substrate glutamate binds. In contrast to its isoforms in other life forms, mammalian GDH demonstrates allostery (Li, Li, Allen, Stanley, & Smith, 2012). Previous studies have shown that the mobility of the enzyme's NAD$^+$-binding domain (Figure 2.1) is essential to mediate the enzyme's allosteric behavior (Li et al., 2012; Smith, Peterson, Schmidt, Fang, & Stanley, 2001). Also, the 'antenna' protrusion in the enzyme's structure is present only in mammalian GDH, and its role has been implicated in the allosteric regulation of the enzyme (Allen, Kwagh, Fang, Stanley, & Smith, 2004).

The most commonly known allosteric effectors for the enzyme are ADP, GTP and NADH, while the enzyme is also known to be regulated by other metabolites such as leucine and monocarboxylic acids (Li et al., 2012) GTP and NADH regulate the enzyme by facilitating its conformational transition to the inactive state in which the NAD$^+$-binding domain has a closed conformation and helps in the modification of the glutamate substrate. ADP on the other hand is responsible for activating the enzyme to release the substrate during which the NAD$^+$-binding domain attains the open conformation. While GTP binds on the NAD$^+$-binding domain below the pivot helix, the binding site for ADP is uncertain (Peterson & Smith, 1999; Smith et al., 2001).

We probe the influence of oligomerization on the dynamics of glutamate dehydrogenase using the PDB structure 3mw9. We observe that the NAD$^+$-binding domain becomes more flexible upon oligomerization while the glutamate binding domain undergoes considerable reduction in its mobility (Figure 2.2.A). Residues K90, K114, K126, R211 and S381 have been shown to interact with glutamate (Peterson & Smith, 1999) and are of prime

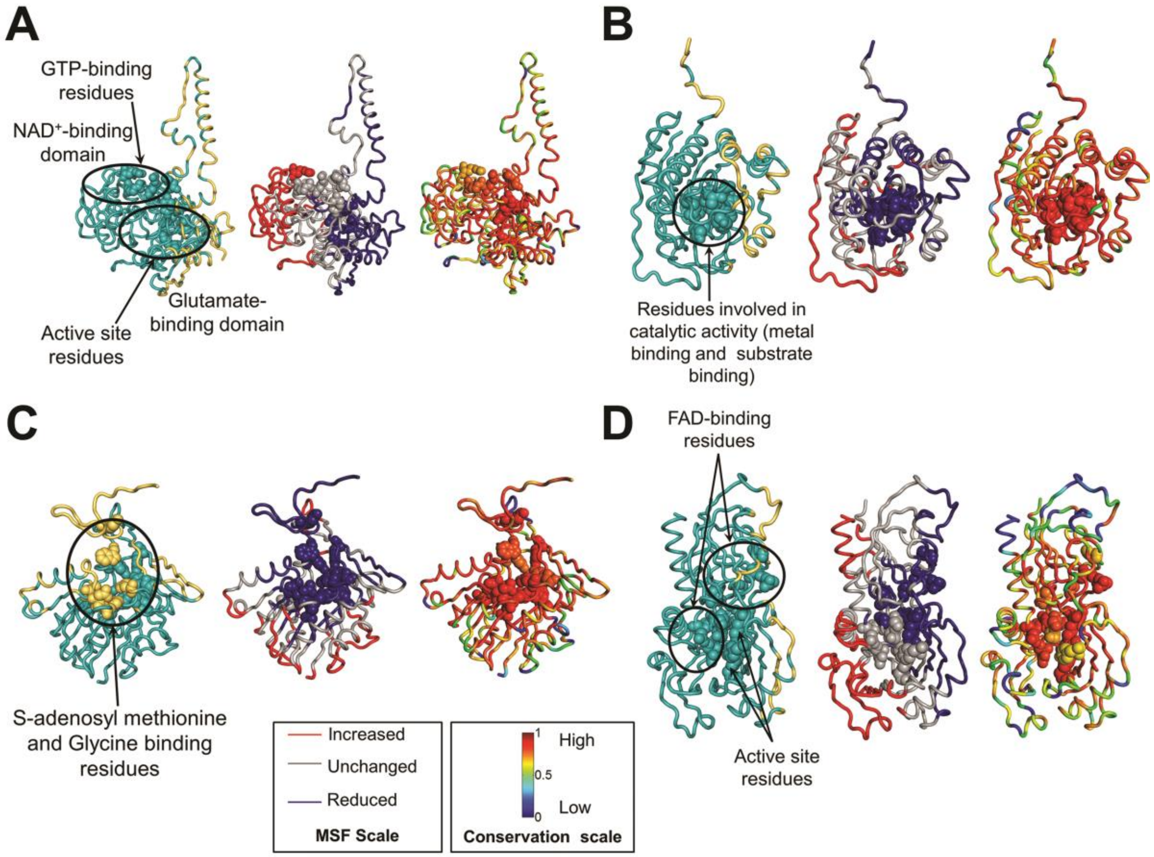

Figure 2.2 Flexibility change and sequence conservation of four enzymes. For each enzyme (A, B, C and D), a figure has three parts. The first part (Left) has the enzyme colored by interface (pale yellow) and non-interface (teal). Next, it is colored by change in residue fluctuations (Middle). Regions with increases in MSF (1.5 fold increase or more) are shown in red, regions with reduced MSF (1.5 fold decrease or more) in blue and those without any significant changes in gray. The third part of the figure (Right) shows the enzyme colored by residue conservation scores with blue and red marking the lower and upper end of the conservation, respectively. In all the three parts, the key functional residues of each enzyme are shown as spheres. (A) Bovine GDH, (B) Arginase 1, (C) Glycine n-methyltransferase (GNMT), and (D) D-aminoacid oxidase. The details of the key functional residues for each enzyme are provided in the Supporting Information.

importance for the enzyme's catalytic activity. Interestingly, four of these residues (K90, K114, K126 and S381) map to regions with reduced fluctuations (Figure 2.2.A and Table A.1). Residues that bind to allosteric regulator GTP, on the other hand, are found either in regions with increased fluctuations or where there are no significant changes in fluctuations.

Residues H209, R217, R261 and R265 which interact with the allosteric inhibitor GTP fall into this category.

Oligomerization increases the packing density of interface residues and as a consequence, it is reasonable to speculate that the flexibility of these residues will be diminished in the assembly. However, in Figure 2.2.A we observe that some residues that are not in the interface also undergo reductions in their fluctuations and some of these residues are orthosteric (involved in the catalytic activity of the enzyme) by nature. Our findings also corroborate results from previous studies which suggest the importance of the mobility of the NAD^+ binding domain for the enzyme's allosteric behavior (Peterson & Smith, 1999). Importantly, the mobility of this domain is significantly higher in the oligomer than in the monomer.

2.3.1.2. Arginase I

Mammalian arginase plays a vital role in the urea cycle, a cascade of chemical reactions that help to eliminate toxic chemicals inside the body. The enzyme is known to exist in two isoforms: arginase I, which catalyzes the hydrolysis of *L*-arginine to form ornithine and urea in the final step of the urea cycle, and arginase II, which regulates the concentrations of arginine and ornithine. Both enzymes have significant roles in maintaining homeostasis inside the body and in facilitating the elimination of toxic chemicals (Kanyo, Scolnick, Ash, & Christianson, 1996). We investigate the effect of oligomerization on the arginase I enzyme from *Rattus norvegicus* (PDB 1rla), which is active as a trimer.

On comparing the MSFs of residues of the independent monomer with the monomer taken as part of its trimeric assembly, we observe that residues located at the interface undergo significant reductions in their mobilities. Some exposed residues which are not part of the interface exhibit increases in fluctuations following oligomerization. We also note that there are residues not at the interface undergoing reduced mobilities. Arginase I uses Mn^{2+} as a cofactor to catalyze the hydrolysis of arginine. Residues H101, D124, H126, D128, D232 and D234 form the manganese binding cluster in the enzyme while, H141 and E277 have been shown to interact with the substrate and are responsible for its catalytic modification (Table A.2) (Cama, Emig, Ash, & Christianson, 2003). Previously, mutation studies of these sites indicated that these severely impair the enzyme's function either by reducing the binding affinity of the enzyme for the cofactor or by reducing its catalytic activity (Cama et al., 2003; Lavulo, Emig, & Ash, 2002). We explore whether these residues have a special preference to exist in regions with increased or dampened mobilities upon oligomerization by mapping them onto the structure and verifying their fluctuation changes. All six residues, where mutation studies were carried out, map to regions having reduced fluctuations after oligomerization. Moreover, residues interacting with the substrate are also seen to be further stabilized in the assembly. Interestingly, all of these residues are in the non-interface parts of the enzyme and yet they displayed significant reductions in their mobilities upon oligomerization (Figure 2.2.B).

2.3.1.3. Glycine N-methyltransferase

Glycine N-methyltransferase (GNMT) is an essential enzyme involved in the metabolism of methyl groups. It uses glycine and S-adenosylmethionine (SAM) as substrates and catalyzes their conversion into S-adenosylhomocysteine (SAH) and sarcosine. The reaction involves the transfer of a methyl group from SAM to glycine. The enzyme is known to be active in its tetrameric form and is found in abundance in mammalian liver cells. It maintains the SAM/SAH ratio in the cell and thus, controls methylation in the cell (Luka et al., 2007). Besides, in humans this enzyme is known to play an important role in gluconeogenesis (Kerr, 1972) and the expression of the GNMT gene is also linked to prostate cancer proliferation (Y. H. Song, Shiota, Kuroiwa, Naito, & Oda, 2011).

For the effects of oligomerization on the dynamics of the monomer (PDB 1bhj), we observe, similar to the previous cases, a major fraction of residues at the interface showing reduced fluctuations while, some residues on the surface showing an increase in mobility. Also, we notice that certain residues in the non-interface regions show reductions in their fluctuations. We then study the changes in flexibility of key residues that interact with substrates. For the rat GNMT, residues Y21, W30, R40, A64, D85, N116, W117, L136, H142 have been shown to interact with the substrate SAM while, residues Y33, G137, N138, R175,Y194,Y220 and Y242 are known to interact with glycine (Table A.3) (Takata et al., 2003). We observe a similar pattern as we did for the other enzymes: the key functional residues are located in regions where the flexibility is reduced upon oligomerization. While most residues involved in binding SAM are located in the interface, residues which bind to glycine are found in non-interface parts of the enzyme and show stabilization upon oligomerization (Figure 2.2.C). Mutations to certain glycine and SAM-binding residues (Y21, Y33, Y194 and Y220) have been shown to be important in contributing to the catalytic efficiency of the enzyme (Takata et al., 2003), and of these, three residues map to regions with reduced fluctuations (Table A.3).

2.3.1.4. D-amino acid oxidase

D-amino acid oxidase catalyzes the dehydrogenation of D-amino acids into their corresponding imino acids. The reaction uses flavine adenine dinucleotide (FAD) as the cofactor and results in the reduction of the cofactor. It is an important enzyme in yeast where cell growth is dependent on the effective utilization of D-amino acids. In mammals, the enzyme is found in a few organs and is known to be catalytically less efficient than its yeast counterpart. In yeast, the enzyme exists as a stable homodimer. Previous studies provided evidence that the enzyme dimerizes upon addition of the cofactor FAD, suggesting that the transition from the apo to the holoenzyme is essential for dimerization (Pollegioni et al., 2002; Pollegioni, Langkau, Tischer, Ghisla, & Pilone, 1993; Porter, Voet, & Bright, 1977).

From the probing of the dynamic effects of oligomerization on a single subunit (PDB 1c0k), we observe that most of the enzyme shows no significant changes in mobility. Interestingly, we do not find regions with increased mobility for the cutoff of 50% change. However, by relaxing the change cutoff to only 25%, we observe that residues on the surface and distal to the oligomerization interface exhibit greater flexibilities than in their isolated form (Figure 2.2.D). While the dynamic flexibility of most of the residues that form the catalytic chamber of the enzyme and interact with substrate (Y1223, Y1238, R1285 and S1335) (Pollegioni et al., 2002) remain relatively unchanged, a larger fraction of residues that bind to the FAD coenzyme (S1012, S1015, A1034, R1035, A1047, S1048, G1052, N1054, V1162, S1334, S1335, G1337, Y1338, Q1339) show reduction in their mobilities (Table A.4 and A.5 and Figure 2.2.D). All of these critical residues map onto the non-interface regions of the enzyme.

2.3.2. Functional Significance of Dynamic Change

Is there a general consensus as observed for the four enzymes above, with most of the functional sites undergoing a significant dampening in their fluctuations upon oligomerization? Or put conversely, are regions with reduced mobilities more conserved? To answer these questions, we consider the residue conservation profiles for all the proteins in the dataset calculated using Rate4Site and investigate the underlying distributions of the conservation scores. On fitting the residue conservation scores to different distributions, we observe that the conservation scores are best fit with the generalized extreme value distribution (Figure A.5) as has often been observed for biological sequences (Bastien & Maréchal, 2008).

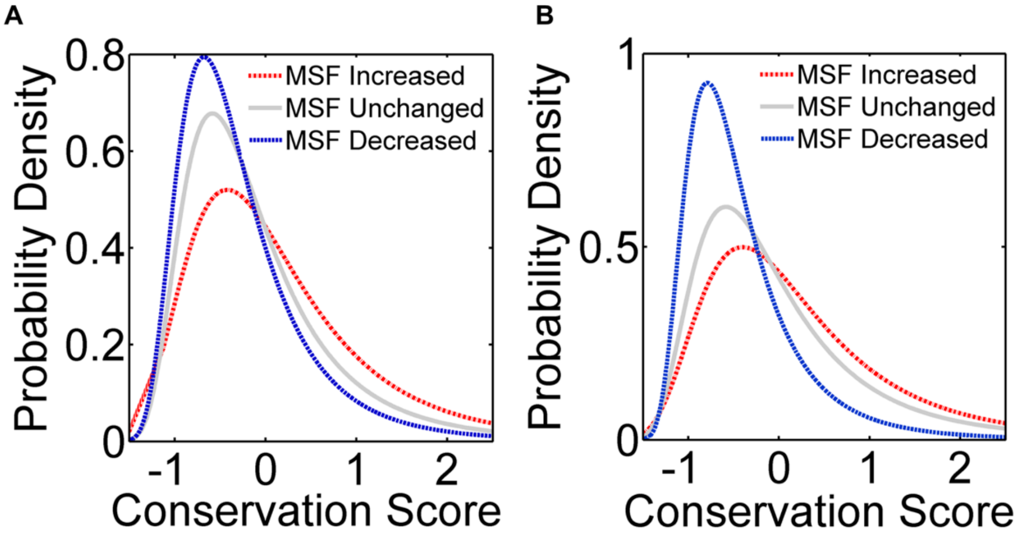

Figure 2.3. Relationship between changes in MSF and residue conservation. (A) For interface residues, the distribution of conservation scores is sharper for regions with reduced MSF, followed by regions with no relative change. The regions which show increases in flexibility upon oligomerization are least conserved and have a broader distribution of conservation scores. (B) For non-interface residues, the same pattern is observed i.e. residues with reduced fluctuations are observed to be more conserved than their counterparts. Conservation scores are computed from Rate4Site with lower scores indicating higher conservation.

In Figure 2.3, we classify residues as either interface (A) or non-interface (B) and for each category we report the distribution of residue conservation scores for the following three classes.

i. Residues with significant increases in MSF (MSF Increased)

ii. Residues with significant decreases in MSF (MSF Decreased)

iii. Residues with no significant changes in MSF (MSF Unchanged)

We identify a residue as an interface residue if any of its heavy atoms is within 4.5Å of the heavy atoms of the residues from an adjacent subunit of the oligomer. We observe that the extent of conservation is higher for both interface and non-interface residues showing reduced fluctuations than the residues that show either increases or no significant changes in their mobilities following oligomerization. Figure 2.3 also suggests that residues with increased mobilities upon oligomerization have a tendency to evolve more quickly than others. We also evaluate the statistical significance of the observed results. A non-parametric test for statistical significance reveals that the observed differences in residue conservation between the three classes is significant both for interface and non-interface residues (Figure A.7). Also, the distributions are similar for the choice of different fold change ratio (FCR) cutoffs (Figure A.6). To verify the consistency of these observations, we create two smaller data sets from the ds145 set, having 40 and 80 structures each, and repeat the calculations at FCR cutoff 1.5. For both sets we observe a similar distribution of the conservation scores (Figure A.8, A and B) which suggest that the results are consistent across multiple data sets.

2.3.3. Global Changes in Dynamics upon Oligomerization

Oligomerization not only reduces the mobilities, but also increases the mobilities of certain residues. This is seen in the four enzymes we described first. We then ask what fraction of residues in the entire dataset have significantly reduced or increased mobilities upon oligomerization. We investigate the changes in residue fluctuations for the threshold of 1.5 FCR that is, 50 % or more increase or decrease in fluctuations. In this way, we observe that 51.5 % of residues across all the proteins in our dataset show no significant changes in their mobilities upon oligomerization, while 26.2 % of the residues undergo a substantial reduction in their mobilities upon oligomerization (Figure A.3 and Table 2.1).

Table 2.1. Extent of Changes in Mobilities. Counts of the number of interface and non-interface residues showing increased, decreased and unchanged mobilities for 145 proteins. Changes indicate at least a 50% gain or loss in mobility.

Class	Oligomer Interface Residues (Counts and Percentage)	Oligomer Non-Interface Residues (Counts and Percentage)	Total (Counts and Percentage)
MSF Increased	238 (2.7%)	7780 (28.2%)	8018 (22.2%)
MSF Unchanged	2409 (28.2%)	16185 (58.6%)	18594 (51.5%)
MSF Decreased	5871 (68.9%)	3622 (13.12%)	9493 (26.2%)
Total	8518 (23.5%)	27587 (76.4%)	36105 (99.9%)

This aligns with one of the most widely accepted consequences of oligomerization, i.e., the dampening of residue mobilities at the binding interface. However, we also observe that 22.2 % of all residues exhibit increases in their flexibilities. 86 % of the proteins (124/145) in the dataset exhibit an increase for at least 10 percent of their residues. Interestingly, a small percentage of the residues (~ 3 %) with increased fluctuations are actually located at the interface of the oligomeric assembly (Table 1, Figure A.4.A). These interface residues with increased fluctuations are found in regions with a significantly lower packing density in contrast to the other interface residues having reduced fluctuations (Figure A.4.B and A.4.C). We also perform this analysis on individual cases, that is, by identifying fractions of residues with increased, decreased or unchanged fluctuations upon oligomerization for each protein and then plotting the results for each category as box plots (Figure 2.4). We still observe that while almost half the residues for each protein show no change in their fluctuations, about a quarter show reduced and another quarter show increased mobilities. These observations suggest that oligomerization is not just a mechanism that dampens the mobility of residues, but is also a means of increasing the flexibility of certain regions of the protein, very nearly a conservation of the extent of internal mobility. Those regions with increased mobilities, as we saw for bovine glutamate dehydrogenase can play an important role in regulating the allosteric behavior of the protein.

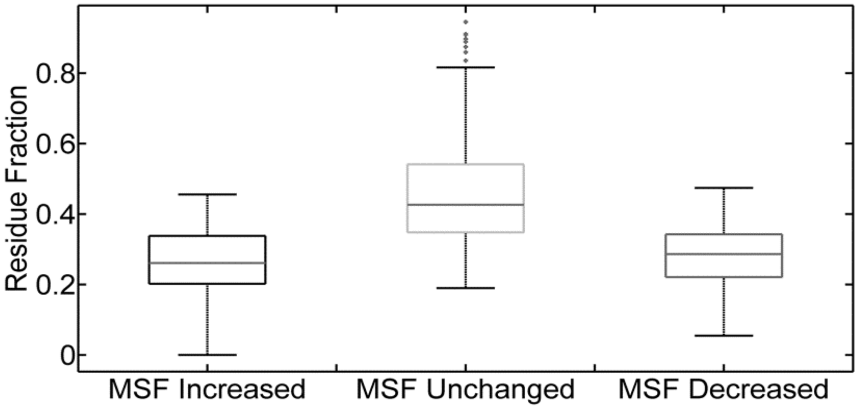

Figure 2.4. Boxplot showing fraction of residues with increased, unchanged and decreased fluctuations across all proteins. Residues with no significant changes in fluctuations have the highest mean fraction (0.474) while, the average fraction of residues with reduced MSF are nearly the same as the fraction of residues with increased (0.278 and 0.246 respectively).

2.3.4. Effect of Oligomerization on Residue Communities: A Case Study on Triosephosphate Isomerase (TIM)

Triosephosphate isomerase (TIM) plays an important role in the glycolytic pathway by catalyzing the reversible interconversion between isomers, dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde 3-phosphate (GAP). The enzyme has a “TIM barrel” fold and is active as a homodimer in most mesophilic organisms. The catalytic chamber of the enzyme is located at the center of each TIM barrel and catalysis is carried out by a Lys-His-Glu triad (Figure 2.5). Glu165 and His95 are critical for proton transfer while, Lys13 bonds with the substrate oxygen (Zhang et al., 1994). Residues 166-176 correspond to the loop 6 which plays a critical role in the presentation and orientation of the ligand to interact with the active site residues. Previous studies showed that the dynamics of this loop is essential for the enzymatic activity, especially in protecting the substrate from solvent and preventing the formation of byproducts (Sampson & Knowles, 1992).

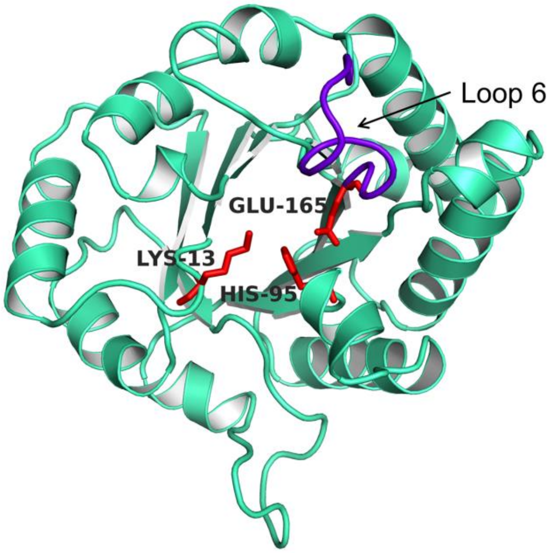

Figure 2.5. Architecture of Triosephosphate isomerase (TIM). The enzyme has a TIM barrel structure with the catalytic residues located at the center of the molecule. The catalytic triad is formed by Lys13-His95-Glu165 (sticks). The mobility of loop 6 plays a key role in bringing in the substrate, protecting the ligand when it is bound, and removing the products.

We study the influence of oligomerization on residue clusters that exhibit significant correlation in their mobilities (referred here as residue communities) in the isolated monomer. Oligomerization, we hypothesize, by changing the geometry of the molecule can facilitate creation of new rigid blocks, often critical for the enzyme’s function. These newly introduced communities, present only in the oligomeric state of the molecule, could possibly explain why some enzymes are functionally active only in their oligomeric form.

Mesophilic TIM is known to be active only in its dimeric form. Interestingly, the enzyme does not form an active site shared between the adjacent subunits at the oligomeric interface (Zhang et al., 1994). The monomeric form of the enzyme is equipped with all the required catalytic residues to carry out its reaction on the substrate. The question then arises, why is oligomerization necessary for TIM if it is catalytically complete in the monomeric form. In this context, we investigate the changes in residue communities upon oligomerization and their importance for the enzyme's function.

We use two forms of TIM: an unbound form (PDB: 8tim) and a substrate bound form (PDB: 1tph). Our aim is to investigate residue communities for four cases: a. single monomer from 8tim as an isolated monomer, b. single monomer from 1tph as an isolated monomer, c. 8tim as a dimer, and d. 1tph as a dimer. For each case, we study the rigid residue blocks in a single chain (by taking the first chain in the PDB file) and observing how they change upon oligomerization. For both forms we coarse-grain the protein by using only the C^{α} atoms, while modeling the substrate in 1tph at an all-atom level. We then model the dynamics of the isolated monomer and the monomer bound in the oligomer as elastic networks, the strength of interactions between residue pairs given by Equation 2.2. We obtain the matrix for correlated fluctuations from the inverse of the Hessian which is constructed using the first twenty soft modes, since these modes convey the most important motions (Equations 2.8 and 2.9). By using a single mode or a combination of these modes, proteins have been shown to undergo conformational transitions essential to their function (Dobbins, Lesk, & Sternberg, 2008; Liu & Bahar, 2012; Marcos et al., 2011). To obtain residue communities, we first transform the matrix of fluctuation correlations into a dissimilarity matrix (Equation 2.10) by subtracting each element from 1 and then perform hierarchical clustering with complete linkage on this matrix (Rokach & Maimon, 2010). The results of hierarchical clustering are displayed as a dendrogram (Figure A.9). We truncate the dendrogram at different levels to obtain two, three and four clusters and treat them as structural blocks having highly correlated fluctuations, refer them as residue communities, and then investigate the influence of oligomerization on these communities.

In Figure 2.6 we have mapped the clusters formed by cutting the dendrograms at 90

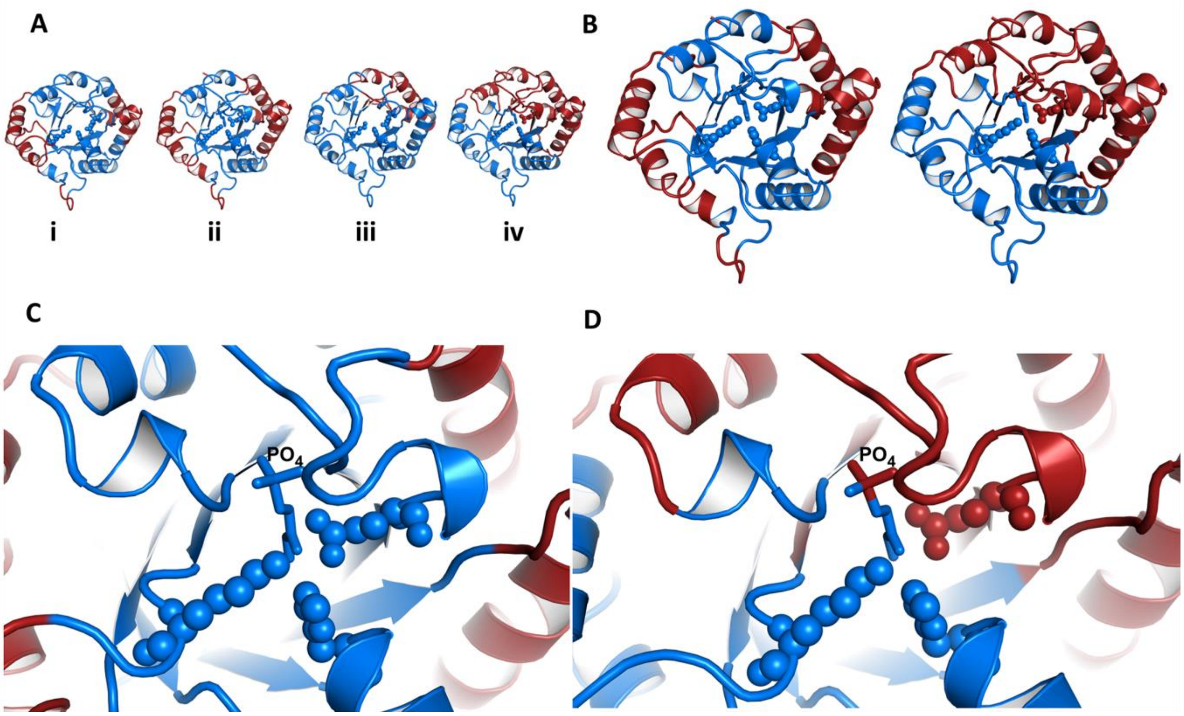

Figure 2.6. Effect of oligomerization on the distribution of residues into correlated communities for TIM. The communities formed upon truncating of the dendrograms at 90 percent (both 8tim and 1tph) are mapped onto the enzyme. (A) The community structure of TIM in isolation without substrate (i), with substrate (ii), as part of the oligomeric complex without substrate (iii), and oligomer with substrate (iv). (B) Communities in 1tph in isolation and in oligomeric association with bound substrate. Close-up view of the architecture of the active site residues and loop 6 for monomeric TIM with substrate (C) and oligomeric TIM with substrate (D) The two communities are colored red and blue. The substrate is shown as sticks and the active site triad as spheres. Glu165 and the phosphate group of the substrate can be seen to be dynamically correlated with loop 6 only in the oligomeric form of the enzyme.

percent of their maximum heights onto the TIM structures. Truncation at this level results in 2 clusters for 1tph and 8tim both in isolation and in their dimeric assembly. Figure 2.6.A shows the mapped residue communities observed for 8tim and 1tph in isolation (i and ii respectively) and when in association with its adjacent unit (iii and iv respectively). As seen in Figure 2.6.A (i and ii), the community structure of TIM in isolation doesn't change much when the substrate is included in the structure. However, the change in the community structure is significant when the molecule is in its oligomeric form and the substrate is included (Figure 2.6.A.iv).

The oligomeric and monomeric forms of TIM show quite different community structures in the presence of substrate (Figure 2.6.B). A close up view of the active site of 1tph in its isolated form (Figure 2.6.C) and in its oligomeric form (Figure 2.6.D) shows the splitting of the active site into two communities (blue and red) in the oligomer while it remains rigid in the monomer. While two of the active site residues (Lys13 and His95) are part of a larger community, Glu165 displays coordinated motion with loop 6 and is part of the second community. We also observe the splitting of the substrate into two communities in the oligomeric form of the enzyme, with the phosphate group of the molecule moving in coordination with Glu165 and loop 6. When the dendrogram for the oligomeric form of TIM was cut to yield 3 clusters, Glu165 still moves in coordination with loop 6 while Lys13 and His95 are still part of the same community. Interestingly, at this level of clustering we begin to observe the coordination of Glu165 with loop 6 even in the unbound oligomeric form of TIM (8tim) as shown in Figure A.10.A.c. However, the observed rigidity of the active site in the monomeric form of TIM is preserved even after cutting the dendrogram for 1tph at different levels to yield three and four clusters (Figure A.10.B).

2.4. Discussion

Dynamics is critical for the functioning of globular proteins. From its native state, a protein can frequently access an ensemble of low energy conformational changes which help it to carry out its function. In many cases, however, there is a set of conformations that cannot be visited from a protein's native state as it incurs a huge increase in the net free energy of the protein. This energy overhead can be overcome through events like ligand binding that can shift the equilibrium population of conformers towards the required conformation by reducing the energy barrier. From the perspective of the Monod-Wyman-Changeux (MWC) model for allostery (Changeux & Edelstein, 2011, 2005), oligomerization is a mechanism to introduce larger scale allostery in proteins through conformational equilibrium shifts. The results presented here, in part, support this hypothesis.

We observe that a major fraction of the proteins in our dataset have a significant number of residues that increase in their mobility upon oligomerization. From the case study on bovine GDH, it is evident that the NAD$^+$-binding domain is more mobile in the oligomer than in the monomer. Oligomerization enables tethering of one end of the enzyme (the oligomeric interface and GLU-binding domain), while allowing the distal end to exhibit increased mobility about the pivot helix. Such mobility, as the MSF comparisons indicate, was not possible when the enzyme was in its monomeric form. Previously, researchers have proposed that the mobility of the NAD$^+$-binding domain can potentially aid in the enzyme's allosteric behavior. If this is true, based on the results presented here it appears reasonable to propose that the enzyme may exhibit diminished allosteric behavior in its monomer form.

The new conformational flexibility introduced upon oligomerization may also be explained in terms of conservation of mobility. For bovine GDH enzyme, an increase in the mobility of the NAD$^+$ domain upon oligomerization is compensated by the stabilization of the GLU-binding domain and at the oligomeric interface, which exhibit a significant reduction in mobility. This explanation is also supported by Figure 2.4 which demonstrates that, upon oligomerization, while half the residues in a protein show no significant change in flexibility, the remaining fraction are almost equally divided between those exhibiting increased and reduced MSF values. Oligomerization could thus be a key contributing factor to the functioning of multi-domain enzymes where one domain is required to be stable and another, mobile. Owing to the observed rigidity in the NAD$^+$ domain, GDH may be also be catalytically less efficient as a monomer. With the newly acquired flexibility in its oligomeric form, the enzyme can now sample new conformations which may not have been accessible to the monomeric form owing to their energetic overhead. Allosteric regulators can exploit this newly introduced conformational flexibility which occurs only in the oligomeric state of the enzyme.

The second part of this study reveals the localization of functionally significant sites in regions having reduced flexibility. The results suggest that residues with reduced flexibility upon oligomerization are more conserved than residues with either increased or no significant changes in flexibility. For the current study, this is true for proteins with oligomeric states ranging from two to six, and we observe similar distributions with varying choices of the FCR cutoff (Figure A.6). From the case studies, these residues can be in regions distant from the oligomeric interface and can present themselves as orthosteric sites where mutations may negatively impact the protein’s function. There are also regions which have no experimentally assigned functional role that exhibit reductions in fluctuations. As can be seen from our four case studies, these residues are present as neighbors to key functional sites. We speculate that these residues could possibly serve as key anchoring sites, whose structural robustness may be critical for the efficient catalytic activity of the enzyme. This however remains to be confirmed experimentally.

The final section of this study investigates the changes in residue communities upon oligomerization and their functional role for triosephosphate isomerase (TIM). The study of TIM shows the critical role of oligomerization in changing the community structure of the active site residues. While the Lys-His-Glu triad remains rigid in the monomer at different levels of hierarchical clustering, oligomerization facilitates a change in this dynamic architecture and promotes the coordination of the Glu165 with loop 6. Previous studies have confirmed a strong correlation between the mobility of loop 6 and Glu165 (Kurkcuoglu, Jernigan, & Doruker, 2006). The mobility of Glu165 has also been proposed to play a key role in placing the substrate into its proper orientation, a requisite step prior to its catalysis. We also observe that the phosphate group of the substrate moves collectively with the same community as loop 6 and Glu165. This is in agreement with previous observations according to which the phosphate forms hydrogen bond with Gly171 in the closed conformation of the loop (Kurkcuoglu et al., 2006). Interestingly, the enzymatic splitting of the substrate is seen only in the oligomeric form. And this separate community analysis reflecting the anti-correlated motions of the active site appears to relate closely to the enzyme mechanism, with these motions assisting the chemical reaction and removal of product. The monomeric form of TIM has all the residues required for its catalytic activity and doesn’t form such a divided active site as seen in its oligomeric form. In principle it might function as a monomer, however it does not. The mixed coarse-grained model used here to investigate the change in communities shows that the coordination of Glu165 with loop 6 is observed only when the enzyme is in its oligomeric form. As stated earlier, the dynamics of these two key elements is critical for the enzyme's function and hence, the results presented here, in the context of residue community changes at the active site elements could explain the inactivity of the monomeric TIM. Oligomerization, as seen in the previous test cases, facilitates the enzyme's access to certain critical conformations that are inaccessible or require high energy for the monomeric form to change the dynamic architecture of the enzyme.

2.5. Conclusion

Our work outlines two key elements of oligomerization. First, it emphasizes the importance of sites whose flexibility is reduced upon oligomerization. Given that the conservation profile of residues follows an extreme value distribution, a large fraction of residues are conserved, making it difficult to identify on this basis alone the potential drug binding sites in a protein. In current practice, residues at the oligomeric interface are often investigated for candidate drug targets (Cukuroglu, Engin, Gursoy, & Keskin, 2014; Kozakov et al., 2011). From this investigation, we conclude that for homooligomeric complexes, regions with reduced fluctuations might also be explored as potential drug targets even though these regions may not always be on the interface. Second, the test case on triosephosphate isomerase states the importance of the residue community changes, providing a possible explanation as to why certain enzymes function only in their oligomeric form. Both these findings can be further explored to better understand oligomeric systems and identify key aspects of their dynamics.

2.6. Acknowledgement

This research was supported by NIH grant R01-GM72014 and NSF grant MCB-1021785, as well as funds from the Carver Trust awarded to the Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.7. References

Ali, M. H., & Imperiali, B. (2005). Protein oligomerization: How and why. *Bioorganic & Medicinal Chemistry*, *13*(17), 5013–5020.

Allen, A., Kwagh, J., Fang, J., Stanley, C. A., & Smith, T. J. (2004). Evolution of glutamate dehydrogenase regulation of insulin homeostasis is an example of molecular exaptation. *Biochemistry*, *43*(45), 14431–14443.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, *80*(1), 505–515.

Bahar, I., Atilgan, A. R., & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design*, *2*(3), 173–181.

Bahar, I., Lezon, T. R., Yang, L.-W., & Eyal, E. (2010). Global Dynamics of Proteins: Bridging Between Structure and Function. *Annual Review of Biophysics and Biomolecular Structure*, *9*(39), 23–42.

Bastien, O., & Maréchal, E. (2008). Evolution of biological sequences implies an extreme value distribution of type I for both global and local pairwise alignment scores. *BMC Bioinformatics*, *9*, 332.

Bordner, A. J., & Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces. *Proteins*, *60*(3), 353–366.

Cama, E., Emig, F. A., Ash, D. E., & Christianson, D. W. (2003). Structural and functional importance of first-shell metal ligands in the binuclear manganese cluster of arginase I. *Biochemistry*, *42*(25), 7748–7758.

Changeux, J.-P., & Edelstein, S. (2011). Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biology Reports*, *3*(September), 19.

Changeux, J.-P., & Edelstein, S. J. (2005). Allosteric mechanisms of signal transduction. *Science (New York, N.Y.)*, *308*(5727), 1424–1428.

Cukuroglu, E., Engin, H. B., Gursoy, A., & Keskin, O. (2014). Hot spots in protein-protein interfaces: Towards drug discovery. *Progress in Biophysics and Molecular Biology*, *116*(2–3), 165–173.

Dalman, M. R., Deeter, A., Nimishakavi, G., & Duan, Z.-H. (2012). Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*, *13 Suppl 2*(Suppl 2), S11.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System. *Schrödinger LLC Wwwpymolorg*, *Version 1.*, http://www.pymol.org.

Dobbins, S. E., Lesk, V. I., & Sternberg, M. J. E. (2008). Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(30), 10390–10395.

Goodman, J. L., Pagel, M. D., & Stone, M. J. (2000). Relationships between protein structure and dynamics from a database of NMR-derived backbone order parameters. *Journal of Molecular Biology*, *295*(4), 963–978.

Goodsell, D. S., & Olson, A. J. (2000). Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure*, *29*, 105–153.

Griffin, M. D. W., & Gerrard, J. A. (2012). The relationship between oligomeric state and protein function. *Advances in Experimental Medicine and Biology*, *747*, 74–90.

Haliloglu, T., & Bahar, I. (2015). Adaptability of protein structures to enable functional interactions and evolutionary implications. *Current Opinion in Structural Biology*, *35*, 17–23.

Healy, E. F. (2015). A model for non-obligate oligomer formation in protein aggregration. *Biochemical and Biophysical Research Communications*, *465*(3), 523–527.

Henzler-Wildman, K., & Kern, D. (2007). Dynamic personalities of proteins. *Nature*, *450*(7172), 964–972.

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680–682.

Kanyo, Z. F., Scolnick, L. R., Ash, D. E., & Christianson, D. W. (1996). Structure of a unique binuclear manganese cluster in arginase. *Nature*.

Katebi, A. R., & Jernigan, R. L. (2014). The critical role of the loops of triosephosphate isomerase for its oligomerization, dynamics, and functionality. *Protein Science*, *23*(2), 213–218.

Kerr, S. J. (1972). Competing methyltransferase systems. *The Journal of Biological Chemistry*, *247*(13), 4248–4252.

Kozakov, D., Hall, D. R., Chuang, G.-Y., Cencic, R., Brenke, R., Grove, L. E., … Vajda, S. (2011). Structural conservation of druggable hot spots in protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13528–13533.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Source Journal of the American Statistical Association*, *4710087*(260), 583–621.

Kurkcuoglu, O., Jernigan, R. L., & Doruker, P. (2006). Loop motions of triosephosphate isomerase observed with elastic networks. *Biochemistry*, *45*(4), 1173–1182.

Lavulo, L. T., Emig, F. A., & Ash, D. E. (2002). Functional Consequences of the G235R Mutation in Liver Arginase Leading to Hyperargininemia. *Archives of Biochemistry and Biophysics*, *399*(1), 49–55.

Li, M., Li, C., Allen, A., Stanley, C. A., & Smith, T. J. (2012). The structure and allosteric regulation of mammalian glutamate dehydrogenase. *Archives of Biochemistry and Biophysics*, *519*(2), 69–80.

Liao, H., Yeh, W., Chiang, D., Jernigan, R. L., & Lustig, B. (2005). Protein sequence entropy is closely related to packing density and hydrophobicity. *Protein Engineering, Design and Selection*, *18*(2), 59–64.

Liu, Y., & Bahar, I. (2012). Sequence evolution correlates with structural dynamics. *Molecular Biology and Evolution*, *29*(9), 2253–2263.

Luka, Z., Pakhomova, S., Loukachevitch, L. V., Egli, M., Newcomer, M. E., & Wagner, C. (2007). 5-Methyltetrahydrofolate is bound in intersubunit areas of rat liver folate-binding protein glycine N-methyltransferase. *Journal of Biological Chemistry*, *282*(6), 4069–4075.

Marcos, E., Crehuet, R., & Bahar, I. (2011). Changes in dynamics upon oligomerization regulate substrate binding and allostery in amino acid kinase family members. *PLoS Computational Biology*, *7*(9), e1002201.

Marianayagam, N. J., Sunde, M., & Matthews, J. M. (2004). The power of two: Protein dimerization in biology. *Trends in Biochemical Sciences*.

Marsh, J. a., & Teichmann, S. a. (2014). Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*, *36*(2), 209–218.

Matthews, J. M., & Sunde, M. (2012). Dimers, oligomers, everywhere. *Advances in Experimental Medicine and Biology*.

Ofran, Y., & Rost, B. (2003). Analysing Six Types of Protein–Protein Interfaces. *Journal of Molecular Biology*, *325*(2), 377–387.

Peterson, P. E., & Smith, T. J. (1999). The structure of bovine glutamate dehydrogenase provides insights into the mechanism of allostery. *Structure*, *7*(7), 769–782.

Pollegioni, L., Diederichs, K., Molla, G., Umhau, S., Welte, W., Ghisla, S., & Pilone, M. S. (2002). Yeast D-amino acid oxidase: Structural basis of its catalytic properties. *Journal of Molecular Biology*, *324*(3), 535–546.

Pollegioni, L., Langkau, B., Tischer, W., Ghisla, S., & Pilone, M. S. (1993). Kinetic mechanism of D-amino acid oxidases from Rhodotorula gracilis and Trigonopsis variabilis. *Journal of Biological Chemistry*, *268*(19), 13850–13857.

Porter, D. J., Voet, J. G., & Bright, H. J. (1977). Mechanistic features of the D-amino acid oxidase reaction studied by double stopped flow spectrophotometry. *Journal of Biological Chemistry* , *252*(13), 4464–4473.

Pupko, T., Bell, R. E., Mayrose, I., & Glaser, F. (2002). Rate4Site-an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, *18*(1), 71–77.

Qaiser Fatmi, M., & Chang, C. E. A. (2010). The role of oligomerization and cooperative regulation in protein function: The case of tryptophan synthase. *PLoS Computational Biology*, *6*(11).

Rokach, L., & Maimon, O. (2010). Chapter 15— Clustering methods. *The Data Mining and Knowledge Discovery Handbook*, 32.

Rother, K., Hildebrand, P. W., Goede, A., Gruening, B., & Preissner, R. (2009). Voronoia: Analyzing packing in protein structures. *Nucleic Acids Research*, *37*(SUPPL. 1).

Sampson, N. S., & Knowles, J. R. (1992). Segmental movement: definition of the structural requirements for loop closure in catalysis by triosephosphate isomerase. *Biochemistry*, *31*(36), 8482–8487.

Sievers, F., & Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology*, *1079*, 105–116.

Smith, T. J., Peterson, P. E., Schmidt, T., Fang, J., & Stanley, C. a. (2001). Structures of bovine glutamate dehydrogenase complexes elucidate the mechanism of purine regulation. *Journal of Molecular Biology*, *307*(2), 707–720.

Song, G., Doruker, P., Jernigan, R. L., Kurkcuoglu, O., & Yang, L. (2008). Elastic network models of coarse-grained proteins are effective for studying the structural control exerted over their dynamics. In G. A. Voth (Ed.), *Coarse-Graining of Condensed Phase and Biomolecular Systems* (pp. 237–254). CRC Press.

Song, Y. H., Shiota, M., Kuroiwa, K., Naito, S., & Oda, Y. (2011). The important role of glycine N-methyltransferase in the carcinogenesis and progression of prostate cancer. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *24*(9), 1272–1280.

Takata, Y., Huang, Y., Komoto, J., Yamada, T., Konishi, K., Ogawa, H., … Takusagawa, F. (2003). Catalytic mechanism of glycine N-methyltransferase. *Biochemistry*, *42*(28), 8394–8402.

Tibshirani, R. (2007). A comparison of fold-change and the t-statistic for microarray data analysis. *Analysis*, *1*, 1–17.

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, *77*(9), 1905–1908.

Walden, H., Bell, G. S., Russell, R. J., Siebers, B., Hensel, R., & Taylor, G. L. (2001). Tiny TIM: a small, tetrameric, hyperthermostable triosephosphate isomerase. *Journal of Molecular Biology*, *306*(4), 745–757.

Zhang, Z., Sugio, S., Komives, E. a, Liu, K. D., Knowles, J. R., Petsko, G. a, & Ringe, D. (1994). Crystal structure of recombinant chicken triosephosphate isomerase-phosphoglycolohydroxamate complex at 1.8-A resolution. *Biochemistry*, *33*(10), 2830–2837.

CHAPTER 3. PROTEIN DYNAMIC COMMUNITIES FROM ELASTIC NETWORK MODELS ALIGN CLOSELY TO THE COMMUNITIES DEFINED BY MOLECULAR DYNAMICS

(*Manuscript submitted to PLoS One*)

Sambit K. Mishra[1,2] and Robert L. Jernigan[1,2]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa 50011, USA

[2]Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011, USA

Abstract

Dynamic communities in proteins comprise the cohesive structural units that individually exhibit rigid body motions. These can correspond to structural domains, but are usually smaller parts that move with respect to one another in a protein's internal motions, key to its functional dynamics. Previous studies emphasized their importance to understand the nature of ligand-induced allosteric regulation. These studies reported that mutations to key community residues can hinder transmission of allosteric signals among the communities. Usually molecular dynamic (MD) simulations (~ 100 ns or longer) have been used to identify the communities - a demanding task for larger proteins. In the present study, we propose that dynamic communities obtained from MD simulations can also be obtained alternatively with simpler models – the elastic network models (ENMs). To verify this premise, we compare the specific communities obtained from MD and ENMs for 44 proteins. We evaluate the correspondence in communities from the two methods and compute the extent of agreement

in the dynamic cross-correlation data used for community detection. Our study reveals a strong correspondence between the communities from MD and ENM and also good agreement for the residue cross-correlations. Importantly, we observe that the dynamic communities from MD can be closely reproduced with ENMs. With ENMs, we also compare the community structures of stable and unstable mutant forms of T4 Lysozyme with its wild-type. We find that communities for unstable mutants show substantially poorer agreement with the wild-type communities than do stable mutants, suggesting such ENM-based community structures can serve as a means to rapidly identify deleterious mutants.

3.1. Introduction

The dynamic nature of globular proteins allows them to sample multiple conformations around their native equilibrium conformation. Such intrinsic dynamics is conferred by their geometry and can be influenced by events such as ligand binding or even binding of a partner enzyme (Nussinov, 2016). Such events typically shift the conformational equilibrium of proteins allowing them to sample new conformations by lowering energy barriers, which were not accessible from the native state (Alberts et al., 2002; Greives & Zhou, 2014). Such dynamic plasticity is characteristic for protein function (Benkovic & Hammes-schiffer, 2003; Daniel, Dunn, Finney, & Smith, 2003; Yon, Perahia, & Ghélis, 1998). It facilitates signal transduction through allosteric regulation as well as allowing bio-molecular machines to undergo large scale conformational changes from their native state essential for their function (Brignole, Smith, & Asturias, 2009; Changeux & Edelstein, 2005; Kern & Zuiderweg, 2003).

Inspecting the conformational ensemble arising due to the dynamic nature of proteins gives immediate insight into how different parts of a protein move with respect to one another. Some regions may exhibit highly correlated motions while others may be anti-correlated in their motions. A map describing the extent of dynamic correlation between residues can then be used to create a graphical representation which portrays the dynamic nature of a protein (McClendon, Kornev, Gilson, & Taylor, 2014). In such a graph, the nodes represent the residues and the edges are weighted by the dynamic correlation for a residue pair. Residue blocks which are highly correlated in their motions and move as a cohesive unit can then be identified from these graphs and are commonly referred to as dynamic communities (Calligari, Gerolin, Abergel, & Polimeno, 2017; Doshi, Holliday, Eisenmesser, & Hamelberg, 2016). These communities may correspond to structural domains in proteins; however, they are often smaller modules whose motions relate to the protein's function.

Previous studies have used both normal mode analysis (NMA) and molecular dynamics (MD) approaches to detect structural domains and dynamic communities in proteins. Hinsen *et al*. (Hinsen, Thomas, & Field, 1999) used normal modes to compute residue-level deformation energy and then, identified dynamically rigid segments using a threshold based on the deformation energy. Kundu and co-workers (Kundu, Sorensen, & Phillips, 2004) used Gaussian Network Model (GNM) to partition protein structures into domains using the eigenvector corresponding the lowest non-zero eigenvalue, also referred to as the Fiedler vector. In another study, Yesylevskyy *et al*. (Yesylevskyy, Kharkyanen, & Demchenko, 2006) used GNM to obtain a dynamic correlation matrix for residue pairs and used it to calculate a “correlation matrix of correlation patterns” which essentially describes the overlap between the correlation patterns for different residues. Then they performed hierarchical clustering on this matrix to obtain rigid communities. A similar study used residue dynamic correlations from normal mode analysis to decompose protein kinases into residue blocks that are dynamically cohesive (Shudler & Niv, 2009).

Other studies where MD simulations were used to identify the rigid domains have also been carried out. Potestio *et al*. (Potestio, Pontiggia, & Micheletti, 2009) used MD simulations to obtain conformational ensemble describing the essential dynamics of proteins and then used dominant eigenvectors from covariance matrix describing the variation in the ensemble to identify rigid domains. McClendon and co-workers (McClendon et al., 2014) performed a thorough investigation of protein kinase A using microsecond-scale MD simulations and then identified communities using residue dynamic correlations from the trajectory with the Girvan-Newman clustering scheme to understand the mechanism of allostery in the enzyme. A similar study on Bruton's tyrosine kinase by Chopra *et al*. (Chopra et al., 2016) revealed that inspecting the community changes for the enzyme's mutant form reveals the changes in the allosteric coupling in the enzyme. In another study, Yao and co-workers (Yao et al., 2016) performed community analysis on G proteins using 80-ns MD simulations to identify residues playing a critical role in the allosteric coupling between functional domain interfaces.

MD simulations do provide a high resolution dynamic image of a protein describing detailed motions of individual atoms at different time points. However, most proteins require energy minimization with respect to an all-atom potential prior to any simulation, a computationally demanding task for larger structures. Moreover, to observe large-scale conformational changes as often seen in the case of multi-domain proteins, simulations need to be performed on the microsecond to millisecond time-scales, which also require considerable computing power. In such cases, coarse-grained approaches like ENM have an upper hand (Atilgan et al., 2001; I Bahar, Atilgan, & Erman, 1997; Tirion, 1996). These models adopt a coarse-grained representation for proteins by representing each residue by only its alpha carbon (C^{α}). They also implement a simplified potential that uses Hookean springs to connect residue pairs within a cutoff distance to calculate the native state dynamics for proteins. In assuming that the crystal structure of a protein corresponds to a local minimum on the energy landscape and considering it as the native state conformation, these models eliminate the necessity for energy minimization. Owing to their reduced nature, these models require minimal computational resources even for large macromolecular structures. Previous studies have shown that theoretical B-factors calculated using ENM correspond well to the experimental temperature factors (Atilgan et al., 2001; I Bahar et al., 1997; Tirion, 1996). In addition, normal modes from ENM show significant overlaps with principal components from both experimental sets of structures as well as with MD ensembles (Yang, Song, Carriquiry, & Jernigan, 2008).

In this study, we have performed a large set of comparisons between the dynamic communities obtained from GNM (I Bahar et al., 1997) (a type of ENM) and from MD for a set of 44 non-redundant proteins. After applying a systematic hierarchical clustering scheme on the dynamic cross-correlation matrices, we observe a close correspondence between the communities from GNM and MD for specific community levels, characterized by a significantly high value of Cohen’s kappa coefficient (Cohen, 1960). Centrality measures for the weighted dynamic network from GNM and MD also reveal a strong correlation for the closeness centrality values. We also verify the extent of agreement for the inter-residue cross-correlations between GNM and MD by investigating the overlaps of the principal eigenvectors calculated from the dynamic cross-correlation matrices and observe a good overlap. A further analysis of the effect of mutations on communities derived using GNM for T4 lysozyme confirms that highly deleterious point mutations significantly alter the community structure when compared to the neutral mutations. The results from our study open up new avenues for mining dynamic communities in macromolecular structures with ENM and using their changes to screen for deleterious mutants.

3.2. Materials and Methods

3.2.1. Dataset

We compile a set of 44 non-redundant proteins from the MODEL database (Meyer et al., 2010) by considering only those proteins with MD trajectories of 100 ns or above. Each protein has a minimum of 50 residues. For each protein, we downloaded the all-atom trajectory from the database and parsed the all-atom trajectory into a C^{α} trajectory, having only the coordinates for residue C^{α} atoms in each frame.

3.2.2. Dynamic Cross-Correlations from MD Trajectory

For each protein, we perform calculations for residue-level dynamic cross-correlations on the respective C^{α} trajectory using the *dccm* function in the Bio3D package(Grant, Rodrigues, Elsawy, Mccammon, & Caves, 2006) with the following equation (Kasahara, Fukuda, & Nakamura, 2014; McCammon, 1984).

$$DCC_{MD}(i,j) = \frac{<\Delta r_i(t).\Delta r_j(t)>_t}{\sqrt{<\left|\left|\Delta r_i(t)\right|\right|^2>_t}\sqrt{<\left|\left|\Delta r_j(t)\right|\right|^2>_t}} \tag{3.1}$$

Here, $r_i(t)$ and $r_j(t)$ refer to the coordinates of the *i*th and *j*th atoms as a function of time *t*, $<\cdot>$ indicates the time ensemble average and $\Delta r_i(t) = r_i(t) - (< r_i(t) >)_t$ and $\Delta r_j(t) = r_j(t) - \left(< r_j(t) >\right)_t$.

3.2.3. Dynamic Cross-Correlations from Gaussian Network Model

We use GNM (I Bahar et al., 1997; Rader, Chennubhotla, Yang, & Bahar, 2006), a form of ENM, to calculate the dynamic cross-correlations between residues. In GNM a protein is usually modeled as a coarse-grained system by representing individual residues by their alpha-carbons, but these points can also be atoms, which we use for the computations on the mutant proteins. Residues within a certain distance cutoff (r_c) are connected by Hookean springs. GNM assumes the protein crystal structure to be of energetic minimum conformation and doesn't require the structure to be energy minimized. It also assumes that residue fluctuations about their mean positions are isotropic and follow a Gaussian distribution in their excursions away from the assumed minimum energy structure. The potential for GNM is given as

$$V = \frac{1}{2}\gamma \sum_{i,j}^{n} \Gamma \ \left[\left(\Delta R_i - \Delta R_j\right)^2\right] \tag{3.2}$$

Here, ΔR_i and ΔR_j are the fluctuation vectors for residue *i* and *j* respectively, γ is the stiffness of the springs connecting residues *i* and *j*. Γ is the Kirchhoff matrix defining node connectivity and is defined as the following.

$$\Gamma = \begin{cases} -1, & if \ \ i \ \neq j \ and \ R_{ij} \leq r_c \\ \ 0, & if \ \ i \ \neq j \ and \ R_{ij} > r_c \\ -\sum_{j,j\neq i} \Gamma_{ij}, & if \ i = j \end{cases} \tag{3.3}$$

Here, R_{ij} is the distance between the alpha carbons of residues *i* and *j* while, r_c is the distance cutoff. Diagonalizing Γ yields *N-1* modes with non-zero eigenvalues. Each mode is a vector that describes the residue fluctuations about its mean position while the eigenvalues correspond to the square of the mode frequency and indicate the relative extent of motion of each point. The slow modes or the low frequency modes describe the most energetically favorable motions of a protein.

The Kirchhoff matrix has a zero determinant and is thus, singular. The pseudo-inverse of this matrix is calculated using the *N-1* or a subset of the *N-1* modes with the following equation.

$$\Gamma^{-1} = \sum_{i=1}^{N-1} \lambda_i{}^{-1} V_i V_i^T \tag{3.4}$$

λ_i is the eigenvalue of the i*th* mode, V_i is i*th* mode and V_i^T is the transpose of V_i. The dynamic correlation between residues *i* and *j* is then calculated as

$$DCC_{GNM}(i,j) = \frac{\Gamma^{-1}(i,j)}{\sqrt{(\Gamma^{-1}(i,i)\, \Gamma^{-1}(j,j)\,)}} \tag{3.5}$$

In the present study, we use a range of different values for the distance cutoff r_c (6, 6.5, 7, 7.5 and 8 Å) and for each value we calculate DCC_{GNM} using 5, 10, 20, 30 and 50 low-frequency modes.

3.2.4. Dynamic Communities from Correlation Matrix

For each protein in our dataset, we convert the dynamic correlation matrices DCC_{MD} and DCC_{GNM} into distance correlation matrices as follows

$$dist\_DCC_{MD} = 1 - DCC_{MD}, \tag{3.6}$$

$$dist\_\, DCC_{GNM} = 1 - DCC_{GNM} \tag{3.7}$$

We then perform hierarchical clustering on the distance correlation matrices with weighted pair-group method with arithmetic mean (WPGMA), which takes into consideration the cluster size when calculating the distance between two clusters (Sokal & Michener, 1958). Hierarchical clustering yields dendrograms that can be pruned at different levels to give the desired number of clusters. The clusters obtained upon pruning a dendrogram at a certain height correspond to the dynamic communities, i.e., the blocks of residues that are highly cohesive and move like a rigid body. We cut the dendrograms at different levels to obtain between 2 and 10 communities. The hierarchical clustering was performed using the MATLAB *linkage (*https://www.mathworks.com/help/stats/*linkage.html)* and *cluster (https://www.mathworks.com/help/stats/cluster.html)* modules.

3.2.5. Comparing Community Assignment between MD and GNM

We use 3 metrics to assess the agreement between the communities from MD and GNM.

1. **Cohen's kappa coefficient.** The Cohen's kappa or simply, kappa is a statistic that is often used to evaluate the extent of agreement between data collectors or raters in their assignments to the same variables, referred to as inter-rater reliability. Kappa coefficient is considered to be more robust than percent agreement as it also takes into consideration random agreement (Cohen, 1960). Like correlation coefficients, the value of the kappa statistic can range from -1 to 1. A kappa of 0 indicates an agreement by chance while kappa of 1 indicates perfect agreement (Cohen, 1960; McHugh, 2012). We calculate the kappa coefficient as follows

$$K = \frac{p_o - p_e}{1 - p_e} \tag{3.8}$$

   Here, p_o is the observed probability of agreement for cluster assignment between MD and GNM while, p_e is the expected probability of agreement.

2. **Network Centrality.** We model each protein as a weighted network in which a node represents a residue and the edge between a pair of nodes is weighted by the distance transformed correlation for the residue pair (Eq. 3.6 and Eq. 3.7). Then, we calculate the node betweenness and node closeness centralities for the networks from MD and GNM. The betweenness centrality of any given node is the number of shortest paths between all pairs of nodes that pass through the given node, while the closeness centrality is the sum of the lengths of the shortest paths to all other nodes from the

given node in the graph. We perform all calculations for network centrality using the MatLab *graph* (*https://www.mathworks.com/help/matlab/ref/graph.html*) and *centrality* (*https://www.mathworks.com/help/matlab/ref/graph.centrality.html*) modules.

3. **Overlap between principal eigen vectors.** We perform singular-value decomposition (SVD) on the DCC_{MD} and DCC_{GNM} matrices and then evaluate the overlaps between the MD and GNM eigenvector spaces for subsets of vectors having largest eigenvalues using the root-mean square inner product (RMSIP) (Amadei, Ceruso, & Di Nola, 1999) as

$$RMSIP = \sqrt{\frac{1}{n}}(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(V_i.U_j\right)^2 \tag{3.9}$$

V and U are the principal eigenvectors obtained from SVD of the DCC_{MD} and DCC_{GNM} matrices respectively, while *n* is the number of vectors to be compared. We consider the same number of principal vectors for the two matrices.

3.2.6. Mutant Dataset

We use PDB structures for the T4 lysozyme mutants crystallized by Mooers et al. (Mooers, Baase, Wray, & Matthews, 2009). In their study, the authors performed circular dichroism assays to estimate stability changes upon specific mutations to the enzyme and calculated the free energy change ($\Delta\Delta G$) for the mutants as $\Delta G_{mutant} - \Delta G_{wildtype}$. The stability changes were performed at pH 5.35 and 3.05. In our study, we consider the $\Delta\Delta G$ values calculated at pH 5.35. Details of the mutant structures used and their free energy changes with respect to the wild-type are given in Table 3.1.

3.2.7. Effect of Mutation on Dynamic Communities

We use all-atom GNM to investigate the community change in the mutant structures with respect to the wild-type. For both the mutant and wild-type forms of the enzyme, we retain all heavy atoms in the PDB and use a distance cutoff of 3.5Å to identify interacting spring locations. Using 5, 10, 20, 30 and 50 modes, we initially calculate the inter-residue dynamic correlations and then, perform hierarchical clustering with weighted average linkage to obtain the desired number of clusters. We trim the dendrograms for each structure at specific heights to obtain 2-10 communities and then compute the agreement between the communities for the wild-type and mutant forms with the kappa coefficient.

3.3. Results

We perform our study on a set of 44 non-redundant proteins (see Table B.1) taken from the MOlecular Dynamics Extended Library (MODEL) database (Meyer et al., 2010). Each protein has a minimum simulation time of 100 ns for its MD trajectory. We consider only the positions of the residue alpha-carbon atoms of each protein from the trajectory file and calculate the inter-residue dynamic correlations from the respective MD trajectory (DCC_{MD}) using equation 3.1. In our procedure we consider only the first frame of the MD trajectory of a given protein as its representative structure to render the protein as a mass-spring system. In such a system, each residue is represented by a point mass (its C^{α} atom) and residue pairs within a given distance cutoff (r_c) are connected by hypothetical Hookean springs. Such a model is commonly referred to as an elastic network model. The Gaussian Network Model is a formulation of ENM that assumes residue fluctuations to be isotropic in nature. Details concerning the implementation of GNM are provided in the Materials and Methods section.

We construct GNM by using a selected set of values for the distance cutoff r_c (6, 6.5, 7, 7.5 and 8 Å) and calculate the inter-residue dynamic correlations (DCC_{GNM}) using a subset of 5, 10, 20, 30 and 50 modes (Eq. 3.5) for each r_c. This is followed by a systematic comparison between the inter-residue dynamic correlations from MD and GNM. Initially, we show how closely the dynamic cross-correlation (DCC) matrices from MD and GNM compare with each other for two randomly selected proteins. Following this, we then perform more thorough comparisons using the following three metrics.

Kappa coefficient. The DCC matrix for a protein describes the extent of correlation between the pairs of its C^{α} atoms. We identify blocks of residues that move cohesively (dynamic communities) by first clustering the DCC matrix hierarchically and then, using a cutoff on the height of the dendrogram obtained to identify the required number of communities (N_c). In the present study, we identify 2-10 communities (N_c = 2, 3, 4 …, 10) for a given protein. Agreement between the communities from MD and GNM is then assessed with kappa coefficient (Cohen, 1960; McHugh, 2012).

Network centrality. We model each protein as a weighted network with the nodes corresponding to residues and edges between pairs of residues weighted by their distance transformed dynamic correlations (Eq 6 and Eq 7). Then, we calculate the residue-level betweenness and closeness centralities and verify the correlations for the centralities obtained from MD and GNM.

Overlap between principal eigenvectors. To assess how well the correlation matrices obtained from MD and GNM compare for a protein, we perform eigen decompositions of the matrices and then use root-mean square inner product (RMSIP) to evaluate the extent of overlap between the principal eigenvectors from the two systems.

In the final section of this paper, we use GNM to delineate the community structure of wild-type and mutant forms of T4 Lysozyme and to show that elastic models can capture the difference in community structures for the wild-type and mutant forms.

3.3.1. DCC Maps from MD and GNM

We perform an initial visual inspection of the dynamic maps obtained from MD and GNM to understand the overall extent of agreement for residue correlations from the two methods. Figure 3.1 describes the dynamic map for two randomly selected proteins from our dataset; *top:* copper transporter domain from copper transporting ATPase (PDB 1fvq), *bottom:* alpha-chymotrypsinogen (PDB 1cgi). The figure shows the distance map between C^{α} atoms (*A*, *D*), DCC maps from MD (*B*, *E*) and GNM (*C*, *F*) for the two molecules. We calculated the DCC map for GNM by setting the distance cutoff r_c to 7Å and then

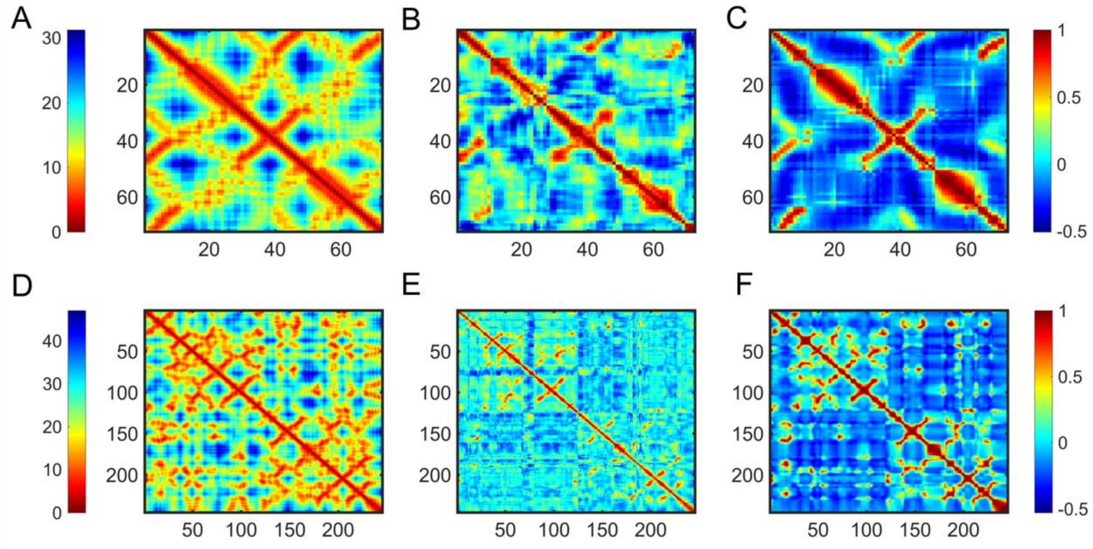

Figure 3.1. Examples of Cα-distance maps and dynamic cross-correlations from MD and GNM for i. Copper transporter domain from copper transporting ATPase (top), and ii. alpha-chymotrypsinogen (bottom). For each protein, the figure shows the distance map for alpha-carbons (A and D), DCC_MD (B and E) and DCC_GNM (C and F). The color scale for the distance matrix has been inverted to agree to the color scale of the DCC matrices (red indicating spatially close residues and blue, distant pairs). PDB IDs of the structures used are 1fvq and 1cgi for i and ii respectively.

considering only the 20 non-zero lowest frequency modes as these have often been shown to circumscribe the most energetically favorable conformation fluctuations in proteins (Haliloglu & Bahar, 2015). The diagonal elements of the correlation maps describe self-correlations while off diagonal elements describe inter-residue correlations or cross-correlations. We note from the outset that there are strong similarities among these representations, corresponding to the secondary structures present in these structures.

The distance map for a protein provides information about the spatial proximity of residues. Spatially close residues are naturally expected to have high correlations in their dynamics. For the two proteins, we observe both MD and GNM showing high dynamic correlations for the spatially close residues. However, it is interesting to notice that correlations for residues in spatial proximity are more strongly indicated with the GNM than by MD. The cross-correlation maps from MD and GNM exhibit good overall agreement. It is also worth noting that for alpha-chymotrypsinogen, the blocks of residues with high dynamic correlation in MD ([1-70], [80-120, 1-70] and [120-220]) are almost closely replicated by GNM. Moreover, the extent of similarity in the correlation profiles of the secondary structure elements (helical regions along the diagonal and anti-parallel beta strands perpendicular to the diagonal, shown in red) for MD and GNM is quite remarkable.

3.3.2. Metric Based Comparisons

i. Kappa coefficient. Our objective is to investigate the level of similarity between the communities obtained from MD and GNM. As we identify a range of communities for a protein (N_c = 2, 3, 4 …, 10), we perform a one-to-one comparison between MD and GNM for a given N_c. To this end, we first calculate for each protein, the dynamic cross-correlation maps for MD (DCC_{MD}) with Eq. 3.1. For each protein, we then construct GNMs by choosing multiple distance cutoffs (r_c) instead of using a single generalized cutoff to address the fact that proteins exist in different geometries and a generalized r_c might not accurately model the dynamics of all different geometries. We construct GNMs for all proteins using r_c= 6, 6.5, 7, 7.5 and 8Å and then, calculate DCC_{GNM} using a subset of the low frequency modes (5, 10, 20, 30 and 50 modes) for each r_c (Eq. 3.4 and Eq.3.5). We thus have 5 correlation matrices for each set of modes and hence, 25 DCC_{GNM} matrices in total for each protein. For a given protein, we then perform hierarchical clustering on the distance transformed DCC_{MD} (Eq. 3.6) and DCC_{GNM} (Eq. 3.7) and then truncate the resulting dendrogram to get 2-10 communities. Using kappa coefficient (Eq. 3.8) (Cohen, 1960; McHugh, 2012), a metric which is used to test inter-rater reliability (extent of agreement between data collectors in assigning same scores to the same variables), we then determine the extent of similarity between the communities from MD and GNM.

For a given protein, we consider the specific combinations of r_c (6, 6.5, 7, 7.5 and 8) and N_c (2, 3, 4 … 10) that yield the maximum kappa coefficient ($Kappa_{max}$) for a chosen subset of modes. For example, if we choose the subset of modes used to calculate DCC_{GNM} as the first 10, then we first calculate the kappa coefficient for all combinations of r_c and N_c (5x9=45 combinations in total) and then choose the particular combination which gives the maximum kappa coefficient. In doing so, we are possibly permitting different r_c for each protein as well as identifying the community level for GNM that shows maximum agreement with MD. Also, for a given protein we assume that communities from MD and GNM best agree for a particular community number and hence, we consider a single value of N_c that satisfies this criterion.

Figure 3.2 shows the median of $Kappa_{max}$ for each subset of modes used. Similar to correlation coefficient, the kappa coefficient can vary from -1 to 1. A value of -1 indicates complete disagreement whereas, 0 indicates the random case. It can be seen that for all subsets of modes used, the median value for $Kappa_{max}$ is at least 0.5, indicating that the agreement is reasonably good and is not just random. It is also seen that using the first 20 low frequency modes yields a median $Kappa_{max}$ of 0.61 with the mode of r_c for $Kappa_{max}$ for 20 modes being 7.5 (Table B.2). We also consider all kappa coefficients for all community levels obtained using the distance cutoff 7.5 and calculate the median kappa for each subset of modes (Fig. B.1). As might be expected, the median kappa when considering all community levels for each subset of modes is smaller than the median of $Kappa_{max}$ ($\approx$ 0.41). Considering the fact that the conformations sampled by MD might be limited, biased by the trajectory time scale whereas ENMs can sample a relatively broader ensemble independent of time, a kappa coefficient of 0.4 indicates fair agreement between the communities but importantly, the agreement is not random.

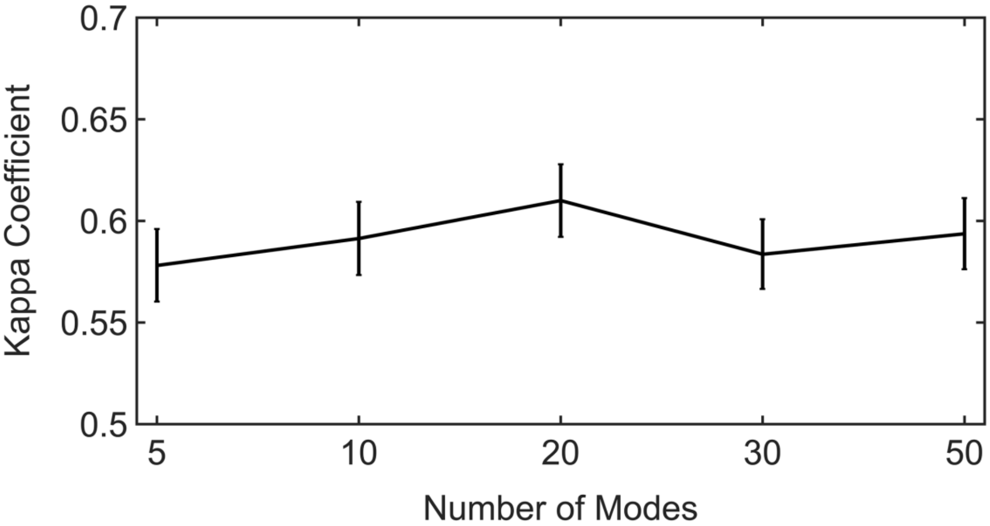

Figure 3.2. Variation of kappa coefficient with the number of modes. The figure shows the median Kappa_max for all proteins in the dataset for subsets including 5, 10, 20, 30 and 50 modes. Vertical bars represent the standard error of Kappa_max.

In Figure 3.3, we show the communities from MD and GNM mapped onto the structures of 5 proteins (A. Copper transporting ATPase, B. Adhesion kinase, C. Guanine nucleotide dissociation inhibitor, D. Hemoglobin, and E. Ubiquitin). For each protein, the figure shows only the community level N_c that provides the best agreement with MD. The figure clearly depicts the close agreement between the communities from GNM and from MD.

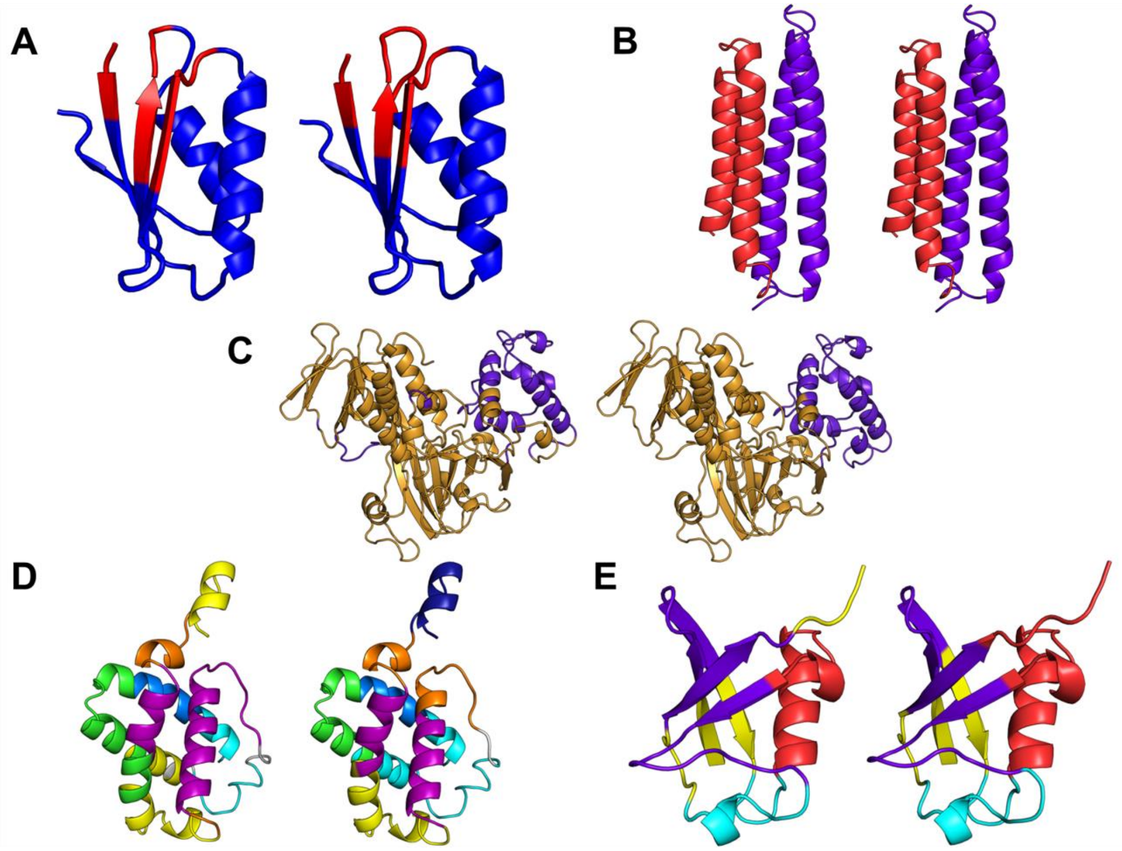

Figure 3.3. Comparison of communities from MD and GNM. Mapped communities for five proteins: (A) Copper transporting ATPase (PDB ID: 1fvq), (B) Focal adhesion targeting domain from adhesion kinase (PDB ID: 1k40), (C) Guanine nucleotide dissociation inhibitor (PDB ID: 1gnd), (D) Hemoglobin (PDB ID: 1idr), (E) Ubiquitin (PDB ID: 1ubq). The number of communities (N_c) shown for each case corresponds to the case of maximum agreement between MD and GNM given by $Kappa_{max}$. DCC_{GNM} calculated with a subset of 20 low frequency modes was used for each protein to perform calculations for communities.

ii. Network centrality. The node centrality is computed by modeling a protein as a network where nodes are the C^{α} atoms and the edges are weighted by the dynamic correlation between a residue pair. Centrality measures tell us the importance of nodes in facilitating the flow of information within the network (Bonacich, 1987; Borgatti, 2005; O'Rourke, Gorman, & Boehr, 2016). The most central nodes act as hubs and can be essential to the transmission of information between nodes at the extreme ends of the network. We compare the extent of correlation for residue centralities between GNM and MD.

We consider two types of node centrality: closeness and betweenness. The closeness centrality of a given residue is the cumulative sum of the lengths of the shortest paths from the residue to all other residues (Bavelas, 1950; Sabidussi, 1966). It is also defined as the reciprocal of farness. The betweenness centrality for a node is the number of shortest paths between all pairs of nodes that pass through the given node (Freeman, 1977). The closeness and betweenness centralities were calculated using the distance transformed DCC_{GNM} and DCC_{MD} (Eq. 3.6 and Eq. 3.7). To verify how well the node centralities from MD and GNM compare, we choose the distance cutoff r_c that maximizes the correlations between MD and GNM for a selected subset of nodes (Table B.4 and B.5). Figure 3.4 shows the correlations for the node closeness (red curve) and node betweenness (blue curve) centralities between MD and GNM. There is a significantly higher correlation for the closeness centrality than for node betweenness. While the maximum correlation for node betweenness is 0.39 (50 modes), the correlation for node closeness is 0.68 (50 modes). It is worth noting that although the maximum correlation is obtained using 50 modes for the two curves, a positive rise in the slope of the two curves is observed only until 20 modes, after which the curves converge. It is also to be noted that the median correlations for 20 modes (closeness = 0.66, betweenness

= 0.35) are not significantly smaller than the values observed for 50 modes (closeness = 0.69, betweenness = 0.39).

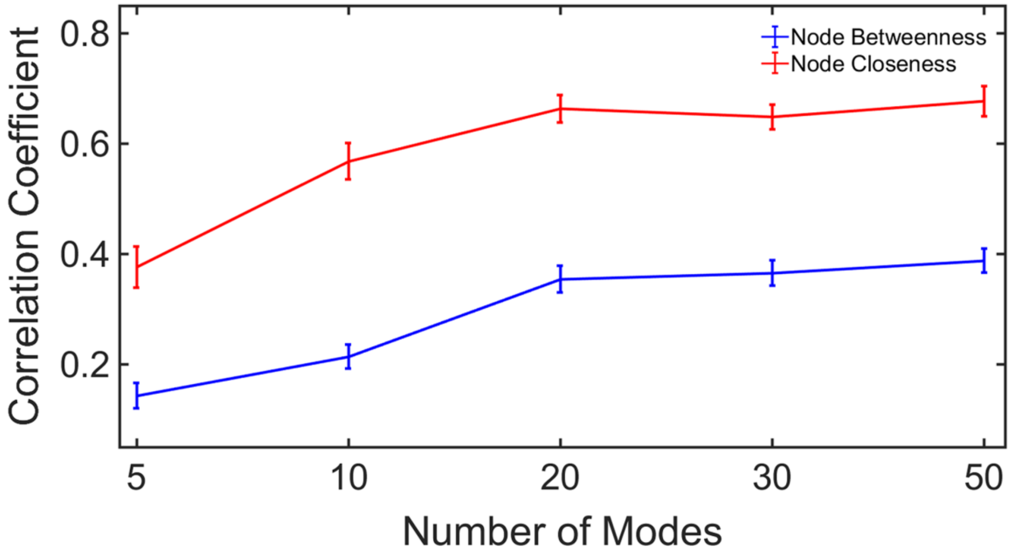

Figure 3.4. Node centrality correlations. The median correlation for closeness centrality (red curve) and betweenness centrality (blue curve) from DCC_{GNM} with DCC_{MD} is shown for different subsets of modes for all proteins. For each protein, the highest correlation and the associated value of r_c is considered for a subset of modes. Vertical bars give values of standard errors.

iii. Overlap between principal eigen vectors. How well do the dominant motions captured from DCC_{GNM} quantitatively compare with DCC_{MD}? How many low frequency GNM modes are required to closely reproduce the correlation pattern from MD? To answer these questions, we investigate the extent of overlap between the principal eigenvectors from DCC_{GNM} and DCC_{MD}.

Let U^N and V^N be the set of N principal eigenvectors obtained upon singular value decomposition (SVD) of DCC_{GNM} and DCC_{MD}. By principal eigenvectors we are referring to the set of eigenvectors with highest eigenvalues. Because the DCC matrix is comparable to a covariance matrix, vectors U_i and V_i are comparable to the principal components of a covariance matrix, capturing the directions of maximum variance from the residue cross-correlation matrix. We inspect the overlap between U and V using root-mean square inner product (RMSIP) (Eq. 3.9) and quantitatively evaluate the extent of similarity between the two matrices. For each set of modes investigated (5, 10, 20, 30 and 50), we consider the value of r_c that maximizes RMSIP. It is also to be noted that we consider the same number of principal eigenvectors each from U^N and V^N as the subset of modes used. Details about the calculation of RMSIP are provided in Materials and Methods. In Figure 3.5, we show that the overlaps between the principal eigenvectors of the DCC_{GNM} and DCC_{MD} matrices are high. The figure also depicts sharp increases in RMSIP and hence, a steep positive gradient as the subset of modes selected increases from 5 to 10 and then from 10 to 20, following which there is convergence.

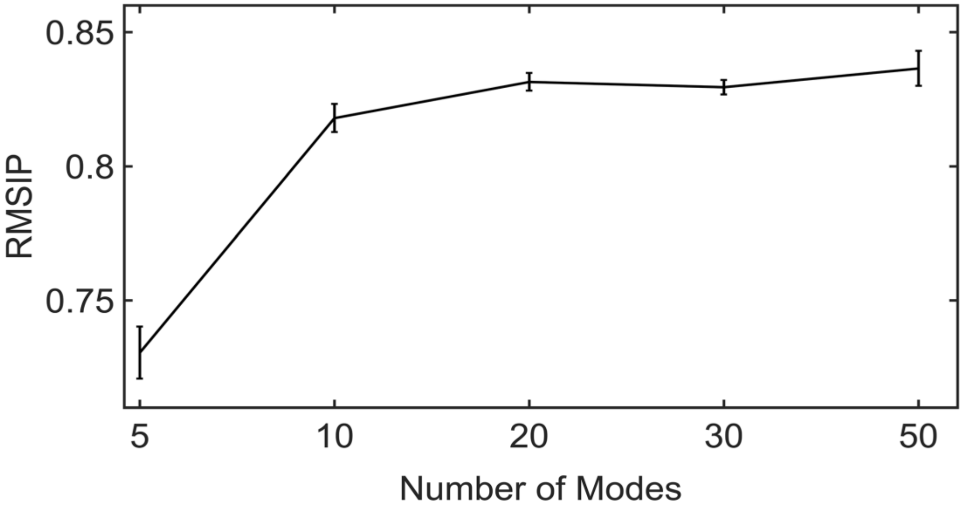

Figure 3.5. Overlap between principal vectors from DCC_{GNM} with DCC_{MD}. The figure shows the extent of agreement between the residue cross-correlation matrices from MD and GNM in terms of the principal eigenvectors. The principal eigenvectors are obtained from singular value decomposition of the DCC_{GNM} with DCC_{MD} matrices, respectively. The median overlap between the vectors from MD and GNM, computed with RMSIP, is shown for subsets of 5, 10, 20, 30 and 50 modes. Vertical bars represent the standard errors in RMSIP.

Table B.6 gives the RMSIP values of individual proteins for different subsets of low-frequency modes.

3.3.3. Changes to Dynamic Communities upon Mutations

Mutations can lead to changes in the structure of dynamic communities (Chopra et al., 2016). We hypothesize that highly unstable mutations tend to change the community structure in a protein more radically than mutations that are less unstable. To test this, we consider 16 mutant structures of T4 Lysozyme crystallized and reported by Mooers *et al* (Mooers et al., 2009). In their study, the authors investigated the effect of mutating Arg96 on the stability of the enzyme. $\Delta\Delta G$ values were reported that indicate changes in the stabilities relative to the wild-type (Table 3.1). We arbitrarily divide the dataset into two groups: the more unstable mutants (rows 1-8) having $\Delta\Delta Gs$ between -4.7 and -2.6 and less unstable mutants (rows 9-16), $\Delta\Delta Gs$ varying between -2.6 and 0. For simplicity, we refer to the more unstable type as *unstable* and the less unstable type as *stable*. We obtain the dynamic communities with GNM using all heavy-atoms from the atomic protein structures and then, with DCC_{GNM} from 10 modes, we verify the community agreement for each of the two mutant types with the wild-type with the kappa coefficients. We consider only 10 modes because this shows the maximum difference in the community structures for the two categories. Results from using other subsets of modes (5, 20, 30 and 50) are also given in Figure B.2.

In Figure 3.6, we show the variation in kappa coefficient for the two mutant categories. For each category, the plot shows the median kappa for individual community levels. It is seen that the s*table* mutants (blue curve) exhibit better agreement with the wild-type than the *unstable* mutants (red curve). Also, it is interesting to note that these differences are manifested in the first 6 communities. At higher community levels, the two mutant types almost come into agreement. To visualize these differences on the protein structures, we consider 3 pairs of *unstable* and *stable* mutants: (PDB IDs: 3c80, 3c81), (PDB IDs: 3c82, 3c81) and (PDB IDs: 3c82, 3c8s). For each pair, we identify the smallest number of communities for which the change is significant. The $\Delta\Delta G$ for each of these mutants can be seen in Table 3.1.

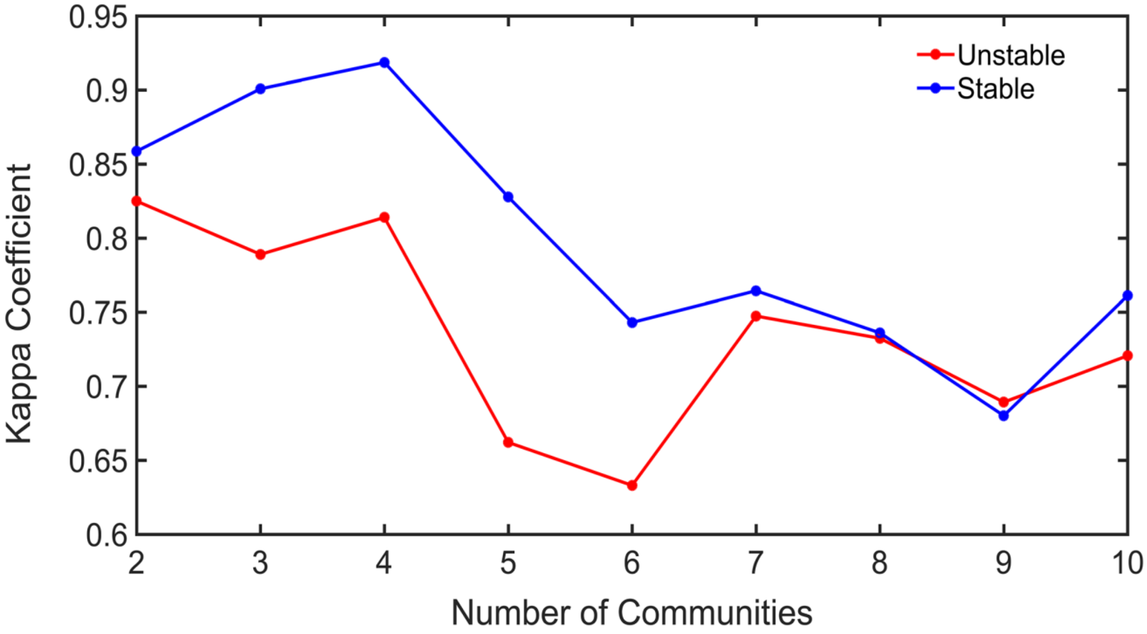

Figure 3.6. Community agreement for *unstable* (*red*) and *stable* (*blue*) mutants of T4 lysozyme with the wild-type. The figure shows the median kappa coefficient (agreement with wild-type) at each community level for the *unstable* and *stable* mutants. The communities were obtained with DCC_{GNM} calculated using 10 low-frequency modes. Calculations using 5, 20, 30 and 50 modes are shown in Figure B.2.

Table 3.1. Mutants for T4 Lysozyme sorted by $\Delta\Delta G$. The set of PDB structures used to compare the community structure of stable and unstable mutants is given below. The Mutation column gives information on the mutation and has the format "xRy", where 'x' is the residue in the wild-type, 'y' the residue in the mutant, and R is the position of mutation in the protein.

PDB Identifier	**Mutation**	**$\Delta\Delta G$ (pH 5.35)**	
3c80	R96Y	-4.7000	
3fi5	R96W	-4.5000	↑
3c7z	D89A, R96H	-3.8000	Decrease (*unstable*)
3c82	K85A, R96H	-3.6000	
3c8q	R96D	-3.5000	
3cdt	R96N	-3.0000	
3cdv	R96M	-2.7000	Stability
3c8r	R96G	-2.6000	
3cdq	R96S	-2.6000	
3c8s	R96E	-2.5000	Increase (*stable*)
3cdo	R96V	-2.4000	
3c7y	R96A	-2.0000	
3c81	K85A	-0.6000	↓
3c83	D89A	-0.5000	
3cdr	R96Q	-0.3000	
3c7w	R96K	0.0000	
4s0w	None (wild-type)	0	

Figure 3.7 (3c80, 3c81), Figure 3.8 (3c82, 3c81) and Figure 3.9 (3c82, 3c8s) show communities for each mutant pair relative to the wild-type (4s0w). In each figure, the wild-type structure with the communities is shown on left, the *stable* mutant in the center and the *unstable* mutant on the right. Side chains of mutation sites are shown as sticks with the same residue side chains displayed in the same color. In Figure 3.7, the difference in community structure for 3c80 (*unstable*) and 3c81 (*stable*) is distinct showing two different communities.

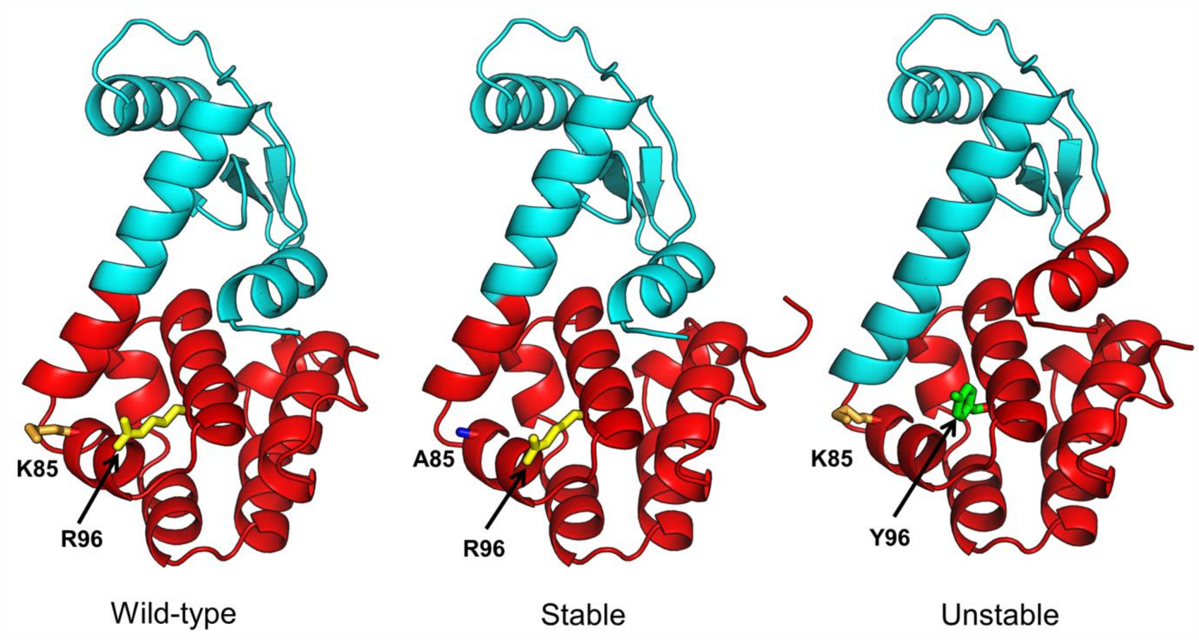

Figure 3.7. Comparison of community structures for wild-type (PDB: 4s0w), stable (PDB: 3c81) and unstable (PDB: 3c80) mutant forms of T4 lysozyme. Two communities (red and cyan) are shown for each structure. We choose $N_c = 2$ because the differences in community structure for the stable and unstable forms are most distinctive at this level. Similarly localized communities are colored alike. Sites of mutations are shown in sticks with the corresponding residue names labelled. Side chains of same amino acids in the sites of mutation are colored alike.

The *stable* and *unstable* forms differ visibly in the dynamic correlation of the N-terminal helix (residues 1-12), which is cohesive with the adjacent N-terminal beta sheets and helices in the wild-type and stable forms, while it moves in coordination with the C-terminal domain in the unstable form. The kappa coefficient for the unstable and stable mutant structures is 0.74 and 0.98, respectively. For 3c82 (*unstable*) and 3c81 (*stable*) (Figure 3.8), the difference is apparent at 3 communities (kappa values of 0.65 and 0.97 respectively).

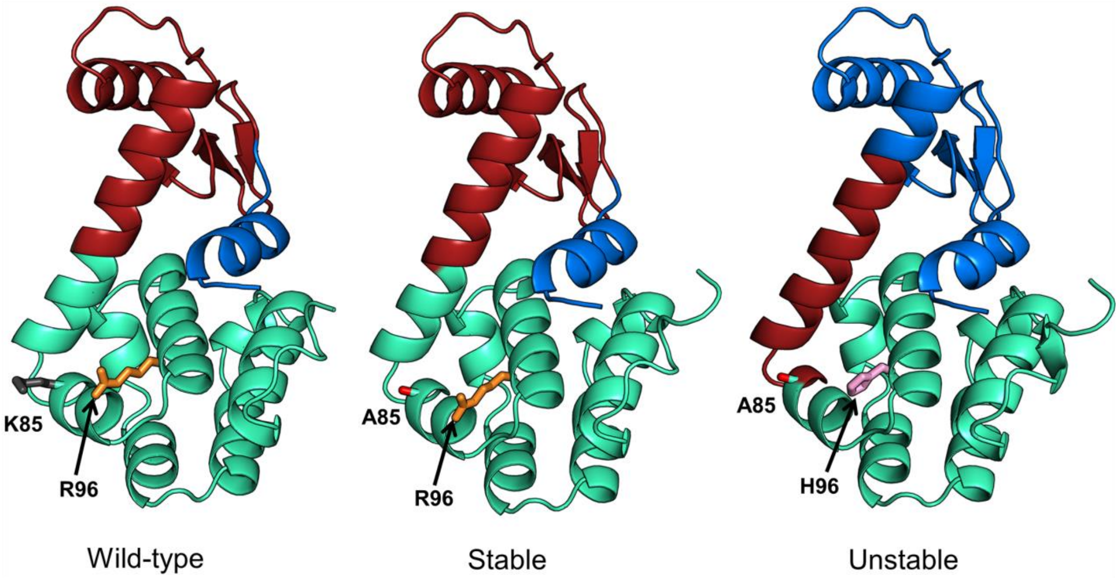

Figure 3.8. Comparison of community structures for wild-type (PDB: 4s0w), stable (PDB: 3c81) and unstable (PDB: 3c82) mutant forms of T4 lysozyme. Three communities (green, brown and blue) are shown for each structure. N_c = 3 shows maximum structural difference between the community structures of mutant and wild-type forms, hence the choice. Coloring scheme is the same as in Figure 3.7.

Again we observe a change in the N-terminal helix that moves as an independent unit in the wild-type and *stable* forms, but shows more coordinated motion with the N-terminal domain in the *unstable* form. In Figure 3.9, we notice the difference at 3 communities and as previously observed, the difference between the *stable* and *unstable* forms becomes visible in the N-terminal helix. The kappa coefficients for the *unstable* (3c82) and *stable* (3c8s) forms at the level of 3 communities are 0.65 and 0.94, respectively.

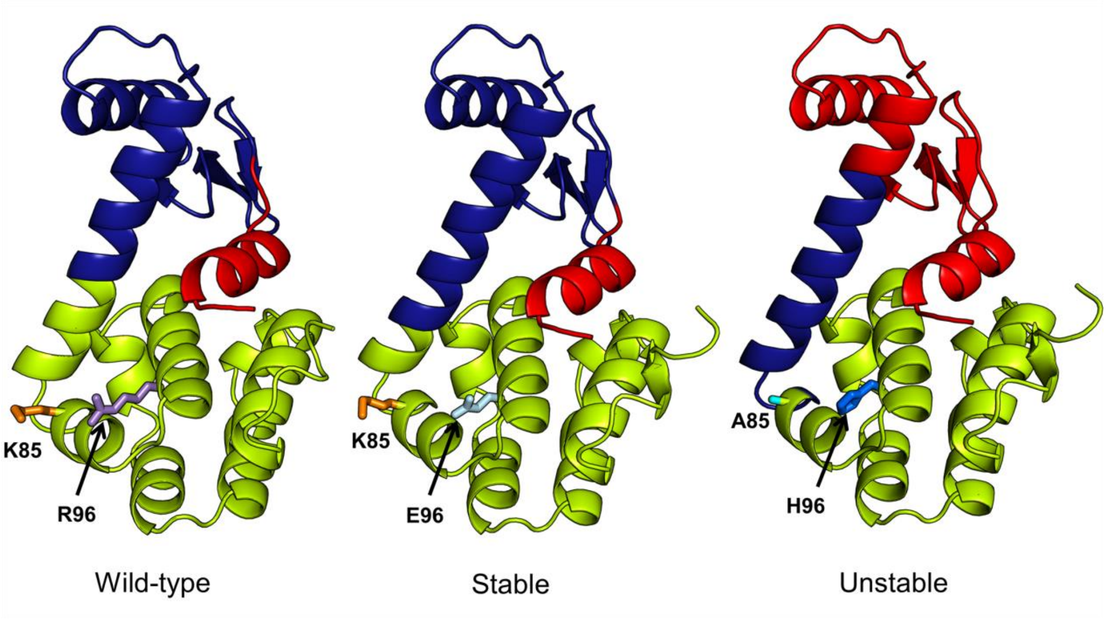

Figure 3.9. Comparison of community structures for wild-type (PDB: 4s0w), stable (PDB: 3c8s) and unstable (PDB: 3c82) mutant forms of T4 lysozyme. Three communities (red, blue and green) are shown for each structure. $N_c = 3$ shows the maximum structural differences for the community structures in the mutant and wild-type forms, hence its choice. The coloring scheme is same as in Figure 3.7 and 3.8.

3.4. Discussion

Previously methods to investigate dynamic communities in proteins have relied on the use of trajectory data from MD simulations. Analyses of dynamic communities stress the importance of identifying the cohesive parts of proteins for their functional dynamics and to understand the mechanisms of protein function and allostery. However, simulations from MD are computationally expensive for large macromolecular structures. There is also a need for long time scale simulations to adequately sample the conformational ensemble for any given protein. This can be demanding in terms of time and often requires use of the highest performance computers. Thus, there is a significant need for a simpler method to aid to capture these dynamic communities, which is computationally less expensive and yet maintains substantially good agreement with the results from MD. This has been accomplished here.

In this present study, we show that communities extracted from GNM exhibit a considerable similarity to the communities from MD. We choose GNM over its anisotropic counterpart ANM (Atilgan et al., 2001) because it is simpler and because previous studies have shown that GNM exhibits better correlations with experimental B-factors than ANM (Kundu, Melton, Sorensen, & Phillips, 2002). Moreover, in preliminary analysis we observe that the communities derived with GNM show better agreement with MD than does ANM. In Figure 3.1, the DCC_{GNM} and DCC_{MD} matrices for two proteins selected randomly from our dataset show considerable agreement for the regions with high dynamic correlation. However, it is surprising to notice a better cohesive behavior, in the case of GNM, showing a close connection between residue dynamic correlation and residue spatial proximity. The dispersion of close contacts suggested by the distance matrix is more closely reproduced with DCC_{GNM} than with DCC_{MD}. This cohesiveness is a hallmark of the elastic network models in general, and is one reason that they can show better agreement with various protein behaviors than MD. It is however to be noted that we use only the first twenty low-frequency modes from GNM to calculate DCC_{GNM}. As we find in other analysis, the agreement between MD and GNM for different metrics converges for the first 20 normal modes, with the addition of more modes not providing any significant gains.

We have considered a range of different distance cutoffs r_c for each protein and then choose the cutoff to maximize the kappa coefficient. In this context, we would like to argue that there is no clear and strict rule for selecting r_c. Previous implementations of ENM have used a range of different r_c and then considered the r_c that best reproduces the experimental

B-factors (Atilgan et al., 2001; I Bahar et al., 1997). Besides, using a generalized distance cutoff fails to take into account the size and variations in the packing density in different proteins and may not accurately represent the protein dynamics. Hence, we presume that the choice for r_c is subjective and proceed by considering a range of values.

Our results from comparing the communities obtained upon clustering the distance transformed DCC_{GNM} and DCC_{MD} matrices hierarchically, suggest that for a certain number of communities N_c, MD and GNM show near-perfect agreement. A median $Kappa_{max}$of 0.61 is observed for 20 modes and convergence of $Kappa_{max}$with 20 modes is quite clear. This also corroborates previous studies that showed that the first few low frequency modes are adequate to reproduce the experimentally observed conformational ensemble of proteins (Ivet Bahar, Lezon, Bakan, & Shrivastava, 2010; Haliloglu & Bahar, 2015). Also, in the case of GNM, though the model assumes isotropic, non-directional residue fluctuations not accounting for the directional preferences of residue mobilities, previous studies suggested that using the first few low frequency modes nonetheless results in good correlations with experimental B-factors. When verifying the median kappa for all modes with r_c = 7.5 Å (Figure B.1), it is interesting to note that the median kappa for each subset of modes at all community levels is almost the same (≈ 0.41). However, when we consider for each protein the r_c giving the highest median kappa over all communities, for the subset of 20 modes, the median kappa increases to 0.49 (Table B.3). While kappa coefficients of 0.41 and 0.49 rule out the possibility of random agreement, at the same time, one must also consider that there could be possible conformational under-sampling depending on the time scale of the MD trajectory that restricts the extent of agreement between MD and GNM.

Convergence with a subset of 20 modes is also consistent for the correlation of node centralities and RMSIP between MD and GNM. Similar to our approach of comparing the communities between MD and GNM, we have used the r_c for which the correlation for node centralities is maximum for a chosen subset of modes. It is quite surprising to observe the relatively higher correlation for node closeness (0.66) than for node betweenness (0.35) centrality. As betweenness for a given node gives an estimate of how many shortest paths between all pairs of nodes pass through the given node, we assume that the networks constructed for the MD and GNM correlation matrices might differ in the shortest paths between two nodes. As Figure 3.1 suggests, DCC_{GNM} and DCC_{MD} do not exhibit 100% agreement with each other. They agree to a large extent in the correlations of secondary structure elements and residues in spatial proximity, however they differ in their scale of inter-residue correlations which could possibly explain the lower correlation for node betweenness.

Singular value decomposition of DCC_{GNM} and DCC_{MD} helps in capturing the directions of maximum variations for inter-residue correlations through its principal eigenvectors. Upon verifying the overlap of the principal eigenvectors between MD and GNM we observe an RMSIP of 0.83 (for 20 modes) followed by convergence. This confirms that the DCC_{GNM} and DCC_{MD} matrices agree to a large extent in terms of the inter-residue fluctuation correlation. It is also interesting to note that when using either a smaller number of modes (5 modes) or too many modes (50 modes) the standard error in RMSIP increases. While using very few modes possibly leads to a loss in information, including more modes in the calculations for DCC_{GNM} possibly adds to the noise, since the most reliable modes of motion for the elastic network models are those at the lower frequency end. Higher frequency modes describe local residue-level dynamics and are less reliable. Hence, including those modes in the calculation of the correlation matrix can potentially reduce the signal to noise ratio, resulting in observed lower agreement of DCC_{GNM} with DCC_{MD}.

The ability of GNM to discriminate stable mutants from unstable ones by evaluating community agreement is notable. The extent of change in community structures in unstable mutants is much greater than for stable mutants. We have used the atomic structures of T4 Lysozyme in the GNM as opposed to the coarse-grained version to account for the mutation changes. Interestingly, we observe that changes to community structures are more distinct in the higher community levels (smaller number of communities) as described by Figure 3.6. One should consider that we have performed this study only for a set of 16 mutant structures of T4 lysozyme, which is really a very small sample. However, we are limited in the availability of experimentally determined mutant structures for a single protein (Ng & Henikoff, 2001; Reva, Antipin, & Sander, 2011). There is some data for the changes in free energy associated with a single point mutation in proteins (Gromiha et al., 2002) however, the crystal structures corresponding to these mutants are not usually available. To use this data, previous methods have considered computational approaches to mutate targeted residues in a given protein and then, used the modeled structure as a representative of the mutant form (Guerois, Nielsen, & Serrano, 2002). However, such computational approaches rely upon the potential function used in the modeling tool and hence, the structure of the modeled mutant (especially the sidechain positions of the mutant site and its neighbors) may be biased by the potential function. The data we have used should be more reliable because these are experimentally reported crystal structures.

In the present study, we focus on a simple approach for detecting dynamic communities in proteins with elastic networks. ENM is simpler to formulate, easier to implement and is computationally less expensive when compared to MD. Our results reveal that this single-parameter model can closely reproduce the results from a complex, multi-parameter model like MD. Owing to its reduced nature, ENM also is superior to MD in terms of execution time and thus, can contribute significantly to the investigation of the dynamic communities in large proteins.

3.5. Acknowledgement

This research was supported by NIH grant R01-GM72014 and NSF grant DBI-1661391, as well as funds from the Carver Trust awarded to the Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology. The funders had no role in study design, data collection or analysis, decision to publish, or preparation of the manuscript.

3.6. References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell. 4th Edition, New York*.

Amadei, A., Ceruso, M. A., & Di Nola, A. (1999). On the convergence of the conformational coordinates basis set obtained by the Essential Dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Structure, Function and Genetics, 36*(4), 419–424.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal, 80*(1), 505–515.

Bahar, I., Atilgan, A. R., & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design, 2*(3), 173–181.

Bahar, I., Lezon, T. R., Bakan, A., & Shrivastava, I. H. (2010). Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical Reviews, 110*(3), 1463–1497.

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America*, *22*(6), 725–730.

Benkovic, S. J., & Hammes-schiffer, S. (2003). R EVIEW A Perspective on Enzyme Catalysis. *Science*, *301*(August), 1196–1202.

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, *92*(5), 1170–1182.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71.

Brignole, E. J., Smith, S., & Asturias, F. J. (2009). Conformational flexibility of metazoan fatty acid synthase enables catalysis. *Nature Structural & Molecular Biology*, *16*(2), 190–197.

Calligari, P., Gerolin, M., Abergel, D., & Polimeno, A. (2017). Decomposition of proteins into dynamic units from atomic cross-correlation functions. *Journal of Chemical Theory and Computation*, *13*(1), 309–319.

Changeux, J.-P., & Edelstein, S. J. (2005). Allosteric mechanisms of signal transduction. *Science (New York, N.Y.)*, *308*(5727), 1424–1428.

Chopra, N., Wales, T. E., Joseph, R. E., Boyken, S. E., Engen, J. R., Jernigan, R. L., & Andreotti, A. H. (2016). Dynamic Allostery Mediated by a Conserved Tryptophan in the Tec Family Kinases. *PLoS Computational Biology*, *12*(3), 1–19.

Cohen, J. (1960). A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES. *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT*, *XX*(1), 37–46.

Daniel, R. M., Dunn, R. V., Finney, J. L., & Smith, J. C. (2003). The Role of Dynamics in Enzyme Activity. *Annual Review of Biophysics and Biomolecular Structure*, *32*(1), 69–92.

Doshi, U., Holliday, M. J., Eisenmesser, E. Z., & Hamelberg, D. (2016). Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proceedings of the National Academy of Sciences*, *113*(17), 4735–4740.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35.

Grant, B. J., Rodrigues, A. P. C., Elsawy, K. M., Mccammon, J. A., & Caves, L. S. D. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England)*, *22*(21), 2695–2696.

Greives, N., & Zhou, H.-X. (2014). Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. *Proceedings of the National Academy of Sciences*, *111*(28), 10197–10202.

Gromiha, M. M., Uedaira, H., An, J., Selvaraj, S., Prabakaran, P., & Sarai, A. (2002). ProTherm , Thermodynamic Database for Proteins and Mutants : developments in version 3 . 0, *30*(1), 301–302.

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology*, *320*(2), 369–387.

Haliloglu, T., & Bahar, I. (2015). Adaptability of protein structures to enable functional interactions and evolutionary implications. *Current Opinion in Structural Biology*, *35*, 17–23.

Hinsen, K., Thomas, A., & Field, M. J. (1999). Analysis of domain motions in large proteins. *Proteins: Structure, Function and Genetics*, *34*(3), 369–382.

Kasahara, K., Fukuda, I., & Nakamura, H. (2014). A novel approach of dynamic cross correlation analysis on molecular dynamics simulations and its application to Ets1 dimer-DNA complex. *PLoS ONE*, *9*(11).

Kern, D., & Zuiderweg, E. R. (2003). The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, *13*(6), 748–757.

Kundu, S., Melton, J. S., Sorensen, D. C., & Phillips, G. N. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophysical Journal*, *83*(2), 723–732.

Kundu, S., Sorensen, D. C., & Phillips, G. N. (2004). Automatic domain decomposition of proteins by a Gaussian Network Model. *Proteins: Structure, Function and Genetics*, *57*(4), 725–733.

McCammon, J. A. (1984). Protein Dynamics. *Reports on Progress in Physics*, *47*(1), 1–46.

McClendon, C. L., Kornev, A. P., Gilson, M. K., & Taylor, S. S. (2014). Dynamic architecture of a protein kinase. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(43), E4623-31.

McHugh, M. (2012). Interrater reliability. *Biochemia Medica, 22*(3), 276–282.

Meyer, T., D’Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., … Orozco, M. (2010). MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure*, *18*(11), 1399–1409.

Mooers, B. H. M., Baase, W. A., Wray, J. W., & Matthews, B. W. (2009). Contributions of all 20 amino acids at site 96 to the stability and structure of T4 lysozyme. *Protein Science*, *18*(5), 871–880.

Ng, P. C., & Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions Predicting Deleterious Amino Acid Substitutions. *Genome Research*, *11*, 863–874.

Nussinov, R. (2016). Introduction to Protein Ensembles and Allostery. *Chemical Reviews*, *116*(11), 6263–6266.

O'Rourke, K. F., Gorman, S. D., & Boehr, D. D. (2016). Biophysical and computational methods to analyze amino acid interaction networks in proteins. *Computational and Structural Biotechnology Journal*, *14*, 245–251.

Potestio, R., Pontiggia, F., & Micheletti, C. (2009). Coarse-grained description of protein internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. *Biophysical Journal*, *96*(12), 4993–5002.

Rader, A. J., Chennubhotla, C., Yang, L.-W., & Bahar, I. (2006). The Gaussian Network Model: theory and applications. *Normal Mode Analysis - Theory and Applications to Biological and Chemical Systems*, *10*(20), 41–64.

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, *39*(17), 37–43.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.

Shudler, M., & Niv, M. Y. (2009). Blockmaster: Partitioning protein kinase structures using normal-mode analysis. *Journal of Physical Chemistry A*, *113*(26), 7528–7534.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.

Tirion, M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, *77*(9), 1905–1908.

Yang, L., Song, G., Carriquiry, A., & Jernigan, R. L. (2008). Close Correspondence between the Motions from Principal Component Analysis of Multiple HIV-1 Protease Structures and Elastic Network Modes. *Structure*, *16*(2), 321–330.

Yao, X. Q., Malik, R. U., Griggs, N. W., Skjærven, L., Traynor, J. R., Sivaramakrishnan, S., & Grant, B. J. (2016). Dynamic coupling and allosteric networks in the α subunit of heterotrimeric G proteins. *Journal of Biological Chemistry*, *291*(9), 4742–4753.

Yesylevskyy, S. O., Kharkyanen, V. N., & Demchenko, A. P. (2006). Hierarchical clustering of the correlation patterns: New method of domain identification in proteins. *Biophysical Chemistry*, *119*(1), 84–93.

Yon, J. M., Perahia, D., & Ghélis, C. (1998). Conformational dynamics and enzyme activity. *Biochimie*, *80*(1), 33–42.

CHAPTER 4. IDENTIFYING PROTEIN REGULATORY AND FUNCTIONAL BINDING SITES BY COUPLING DYNAMICS AND EVOLUTIONARY INFORMATION WITH STRUCTURE

Sambit Kumar Mishra[1,2], *Gaurav Kandoi*[1,3], *Robert L. Jernigan*[1,2]

[1]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa 50011, USA

[2]Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa 50011, USA

[3]Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011, USA

Abstract

Binding interactions of proteins with other molecules is a key determinant of their functional role. Binding sites can be either specifically functional binding sites or regulatory binding sites. Functional binding sites are also referred to as active sites while, regulatory sites are referred to as allosteric sites. The intrinsic dynamics of proteins plays a key role in maintaining their function. Such dynamic behavior also controls binding events and their induced effect on the protein's intrinsic dynamics. This study presents a novel binding site prediction method, AR-Pred (**A**ctive and **R**egulatory site **Pred**iction), by supplementing protein geometry, evolutionary and physicochemical features with information about protein dynamics. We use a common subset of these features to train and test separate models to separately predict allosteric and active site residues. The models are trained and validated on 10 balanced training and validation sets. Our models for active site prediction yield a median

AUC of 91% and MCC of 0.68, whereas the allosteric site prediction models show a median AUC of 80% and MCC of 0.48. When tested on a subset of proteins, our models for active site prediction show comparable performance to two existing methods and gains compared to two others, while the allosteric site predictions show significant gains in performance compared to three existing prediction methods.

4.1. Introduction

Many globular proteins are enzymes that catalyze chemical reactions on bound substrates with the whole protein facilitating the reaction by lowering energy barriers (Alberts et al., 2009). Their catalytic efficiency can be regulated by environmental factors such as temperature and pH and, importantly, often also by the binding of effectors or allosteric modulators. Such interactions with other molecules are a key regulatory aspect of proteins in general, which closely relate to their functions. Consequently, identification of possible binding sites is of vital importance. It is a useful step in the process of annotating proteins for function, and is a widely acknowledged important problem.

Proteins exhibit a broad spectrum of ligand and macromolecule binding sites. Metalloproteins have metal ion cofactor binding sites, molecular chaperones like GROEL can bind to other proteins, DNA binding sites are found in helicases and topoisomerases, while proteases bind to targeted peptides. Specifically, ligand binding sites in most enzymes can be broadly classified into two categories: a functional binding site or active site where the substrate binds in order to undergo chemical modification, and a regulatory binding site or allosteric site where, binding of an effector molecule can regulate and control the activity of the protein. The active site may be further divided into the substrate binding site comprising all residues that interact with the substrate and a catalytic site, consisting of only residues directly taking part in the chemical reaction for substrate modification. In this study, we will use the term active site to refer broadly to include both of these sub-categories.

A protein’s active site is comprised of a group of residues, most frequently located deep in its interior and even sometimes at the interfaces between subunits, and in many cases the site is accessible through a network of channels (Pravda et al., 2014). Proteins also frequently undergo transitions between different conformations which control access to the active site. The structural architecture and the physicochemical nature of the residues in the active site are evolutionarily conserved across different species, to retain the specific function of the protein. Active sites constitute the functional binding sites of enzymes and play a key role in defining an enzyme’s function. Deletion of residues at or near the active site can result in total loss of function. While an enzyme’s active site defines directly its biological activity, allosteric or regulatory sites control such activity remotely. Residues constituting such sites are commonly localized to cavities on protein surfaces and are typically more accessible to ligands than are the residues in active sites. Protein allostery is a fundamental biological mechanism through which binding of a ligand molecule at a site remote to the functional site in an enzyme results in changes to the shape or dynamics of the functional site, either activating or inactivating the enzyme’s activity (Tsai, Del Sol, & Nussinov, 2009). Such allosteric processes facilitate communication between distant sites in proteins. Allostery is key for signal transduction: the receptors on the surface of cells use it to transmit signals from the exterior to the interior of the cell (Motlagh, Wrabl, Li, & Hilser, 2014; Nussinov & Tsai, 2014). Abnormalities in allosteric regulation have also been linked to several human diseases such as cancer and Alzheimer’s (Li et al., 2013). Allosteric drugs currently represent a major effort in pharmaceutical industries in contrast to drugs targeted to active sites.

Because allosteric residues are subject to lower evolutionary pressure compared to orthosteric residues, they are often not conserved across protein families and have the advantage of being highly specific to a given protein. Hence, allosteric drugs have a lower risk of interfering with a host's protein. They also have the potential to activate as well as inhibit the target protein and can be used in combination with drugs that target active site residues.

A number of computational methods exist for the prediction of ligand binding sites in proteins. Based on the distinguishing properties which they use, such computational approaches may either be template-based, utilizing homologous structures with known binding sites or even geometry-based, using structural geometry to detect binding site pockets. Also, some methods are energy-based and rank putative ligand binding sites by their interaction energies with hypothetical ligands (Xie & Hwang, 2015). Specific methods also exist for the prediction of functional sites (active sites). The Fuzzy Oil Drop model developed by Brylinski and co-workers (Brylinski et al., 2007) accounts for irregularities in hydrophobicity distribution of different residues in a protein and assigns functional importance to regions with high irregularities. Ondrechen *et al*. developed a computational method that calculates theoretical microscopic titration curves (THEMATICS) and showed that residues exhibiting anomalies in their predicted titration curves occur at active sites (Murga et al., 2004; Ondrechen, Clifton, & Ringe, 2001). A more sophisticated method POOL was later developed that uses electrostatic and geometric properties derived from protein structures in addition to sequence conservation and features from THEMATICS to assign likelihood estimates for residues being part of the active site (Somarowthu & Ondrechen, 2012; Tong, Wei, Murga, Ondrechen, & Williams, 2009). Thornton and co-workers developed ConCavity which combines evolutionary sequence conservation with geometric features obtained from pocket finding algorithms to predict active site residues (Capra, Laskowski, Thornton, Singh, & Funkhouser, 2009). Another method that predicts active site pockets is AADS (Singh, Biswas, & Jayaram, 2011) that uses geometric information on cavities in addition to physico-chemical properties of residues. Some methods have implemented genetic algorithms which use structural information as well as sequence and network based properties in combination with machine learning to identify active site residues (Izidoro, De Melo-Minardi, & Pappa, 2015; J. Song et al., 2018). More recently, protein dynamics was also used as a predictor for active sites. Glantz-Gashai and co-workers revealed that normal modes can expose active sites and used changes in solvent accessibilities to predict active site residues (Glantz-Gashai, Meirson, & Samson, 2016).

Numerous initiatives have also been taken to identify allosteric sites. The ASD database includes a diverse set of proteins with known allosteric residues. The identifications of allosteric sites for the proteins in this database are based on experimental methods which include disulfide trapping, high-throughput screening and fragment-based screening (Z. Huang et al., 2011). There have also been different approaches taken that use sequence and structural information to make predictions of allosteric sites in proteins. Lockless and Ranganathan used statistical coupling analysis (SCA) to identify networks of coevolving residues for protein families and later, used them to identify potential allosteric sites and pathways (Lockless & Ranganathan, 1999). Allosite is a structure-based machine learning predictor that uses the physicochemical properties of pockets predicted by FPocket as descriptors to train a support vector machine (SVM) model and make predictions of allosteric pockets (W. Huang et al., 2013). AlloPred uses normal mode perturbations on different pockets in a protein to identify the pockets whose perturbation induces maximum flexibility changes for the catalytic residues (Greener & Sternberg, 2015). A similar method that uses normal modes to simulate the effect of ligand binding on protein flexibility is used in the protein allosteric and regulatory sites (PARS) server (Panjkovich & Daura, 2012). This server tags those pockets in a protein as allosteric that induce maximum flexibility changes in the protein upon ligand binding. SPACER is another predictive tool that combines normal modes with dynamics and uses ‘binding leverage’ to locate potential sites in proteins where ligand binding can trigger a population shift affecting the conformation state of the protein (Goncearenco et al., 2013).

The dynamic nature of proteins is a critical element that can control function by transient reorganization of enzyme active sites (Benkovic and Hammes-Schiffer, 2003) and their regulatory behavior by a shift in conformational dynamics upon effector binding (Motlagh et al., 2014). In addition, protein dynamics is thought to play a pivotal role in the evolution of novel function (Campbell et al., 2016). Collectively these studies suggest that supplementing information on protein dynamics with structural and evolutionary features inside a machine learning scheme ought to lead to improved predictions of ligand binding residues, both for active site and allosteric residues, the underlying premise for this work. To test this hypothesis, we use the dataset compiled by Greener and Sternberg (Greener & Sternberg, 2015) for AlloPred since it includes information about both allosteric and active site residues. In our model, we include features that describe the dynamic behavior of residues in a protein molecule by simulating the protein with elastic network models (Atilgan et al., 2001; Bahar, Atilgan, & Erman, 1997). This includes mean-square fluctuations of residues and the resilience of residues to external perturbations given by dynamic flexibility index (Gerek, Kumar, & Ozkan, 2013). For prediction of allosteric residues, we additionally consider the shortest dynamically correlated path between a given residue and the active site residues and the effect of perturbing the active site residues on a given residue. In addition, we also model a protein structure as a network where each node is a residue and the edge between a pair of nodes is weighted by the extent of dynamic correlation between them, following which we calculate network centrality features for each residue. We supplement dynamic features with structure-based features such as solvent accessibilities, amino acid physicochemical properties like hydrophobicity and also evolutionary conservation. Our results suggest that residue-level conservation is the most important determinant for both allosteric and active site residues. In addition, we observe that for the predictive models of both allostery and active site, the dynamic features are one of the top 10 most important features. Of the four methods compared using a compiled test set of proteins, our predictive models for active sites show comparable performance to two (POOL and ConCavity) and outperform two others (Fuzzy Oil Drop and AADS). Our models for allostery however, outperform all three of the other methods compared (AlloPred, AlloSitePro and Spacer). Our study thus, verifies the importance of incorporating residue-level dynamic information into predictive models for ligand binding sites.

4.2. Materials and Methods

4.2.1. Dataset

Since our aim is to develop predictive models for both allosteric and active site residues, we use the dataset of protein structures (PDB files) compiled by Greener and Sternberg (Greener & Sternberg, 2015) for AlloPred that contains information on both allosteric and active site residues. The authors obtained information about allosteric residues from ASBench and used Catalytic Site Atlas and UniProt in addition to ASBench to identify active site residues. The training and testing datasets provided there include a total of 119 proteins.

4.2.2. Dataset Processing

We split the multimeric proteins in our dataset into their individual chains. This resulted in a total of 173 separate protein chains. We then retain those chains which had both allosteric and catalytic residues, leaving 165 protein chains (from 105 proteins). For the same set, we calculate all the features as described next in the *Feature calculations* section. For some structures, we encountered errors during feature calculations. For example, calculations for evolutionary conservation gave errors in the presence of non-standard amino acids and in some cases, solvent accessibility and secondary structure calculations couldn’t be performed for all residues for some proteins. We discarded these structures and our final dataset contains 144 protein monomers taken from 105 proteins.

4.2.3. Feature Calculations

For each protein, we calculate features at a residue-level which are based on amino acid physico-chemical properties, evolutionary conservation, protein structure geometry, and protein dynamics as described below.

Residue type

We classify residues based on their hydrophobicity and charge into three classes similar to the approach taken by Petrova *et al.* (Petrova & Wu, 2006).

Class 1: His, Arg, Lys, Glu and Asp (charged residues)

Class 2: Gln, Thr, Ser, Asn, Cys, Tyr and Trp (polar residues)

Class 3: Gly, Phe, Leu, Met, Ala, Ile, Pro and Val (hydrophobic residues)

Residue identity

We label each of the 20 amino acids (A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y and V) separately.

Solvent accessibility

We perform calculations for solvent accessibility using Naccess (Hubbard & Thornton, 1993) with default parameters. Naccess reports the following absolute and relative accessibilities, all of which we include in our feature set.

i. ASA_ALLATOM (Abs/Rel): Relative and absolute solvent accessible surface area for a given residue based on all atoms in the residue.

ii. ASA_SIDECHAIN (Abs/Rel): Relative and absolute solvent accessible surface based on only the side chain atoms. Naccess considers the C^{α} atoms to be side chain atoms, so that glycine also has side chain accessibility.

iii. ASA_MAINCHAIN (Abs/Rel): The relative and absolute solvent accessibilities for the main chain atoms of a residue, excluding C^{α} atoms.

iv. ASA_NONPOLAR (Abs/Rel): Relative and absolute solvent accessibilities for all non-oxygen and non-nitrogen atoms in the side chains.

v. ASA_POLAR (Abs/Rel): Relative and absolute solvent accessibilities for all oxygen and nitrogen atoms in the side chains.

Secondary structure

We use the DSSP program to assign the secondary structures to the residues. DSSP assigns a single letter code (H, S, G, T, E, B, I, -) to each residue corresponding to the secondary structure type.

Mean square fluctuations

We use the Anisotropic Network Model (ANM), a type of Elastic Network Model (ENM) to calculate the residue-level fluctuations (Atilgan *et al.*, 2001). We model each protein as a coarse-grained elastic network by representing its residues by their respective C^{α} atoms and connecting close pairs of these atoms with harmonic springs. The potential energy of this system under equilibrium is given as

$$V = \frac{1}{2}\, \Delta R^{T} H \Delta R \tag{4.1}$$

Here, ΔR is the vector of change in position for all residues, ΔR^{T} is its transpose and H is the $3N$ by $3N$-dimensional Hessian matrix obtained from the second derivatives of the potential function. We vary the strength of the springs γ between a residue pair by the inverse of their separation distance(d_{ij}), given by the following equation (L. Yang, Song, & Jernigan, 2009).

$$\gamma = \left(\frac{1}{d_{ij}}\right)^{2} \tag{4.2}$$

Upon diagonalization, the Hessian matrix yields 3N-6 normal modes (V) and eigenvalues (λ) corresponding to the non-rigid body fluctuation dynamics of the protein. We calculate the mean square fluctuations (MSF) for residues in a protein using the eigenvalues and eigenvectors obtained with the following equation.

$$< \Delta R_{i}^{2} > = \sum_{j=1}^{3N-6} \frac{1}{\lambda_j} \sum_{i=3k-2}^{3k} V_{ji}^{2} \tag{4.3}$$

Hydropathy index

We use the Kyte-Doolittle hydropathy scale (Kyte & Doolittle, 1982) to calculate residue hydrophobicity.

Dynamic flexibility index (DFI)

From linear response theory, the response vector to an external perturbation in a protein structure such as binding of a ligand can be obtained using the equation

$$\Delta R_{3N\times 1} = H^{-1}_{3N\times 3N}\, F_{3N\,\times 1} \tag{4.4}$$

Here, ΔR is the 3N dimensional response vector giving the positional displacement of each atom in X, Y and Z. H^{-1} is the 3N by 3N dimensional inverse Hessian matrix and F is the 3N dimensional force vector. Based on linear response theory, the metric dynamic flexibility index (DFI) (Gerek *et al.*, 2013; Kumar *et al.*, 2015) estimates the resilience of a given residue position to perturbations at all positions within the 3-D structure of the protein. Sites with low DFI, such as hinges, are more resilient to perturbations and are hence, dynamically more stable than sites having high values of DFI. DFI also measures the significance of each position's contribution to the global functional dynamics of the protein. We perform calculations for DFI for each protein and obtain the indices for each residue using the method described by Gerek *et. al* (Gerek *et al.*, 2013).

Active site perturbation response (only for allosteric classifier)

The active site perturbation response is a measure of the effect of perturbations on the functional binding site or active site on other residues. Residues which show higher fluctuation responses upon perturbation of the active site are often associated with allosteric signal transmission. We calculate the active site perturbation response as described by Kumar et al. (Kumar et al., 2015). For the calculation of this feature, identification of which residues form the active site is essential.

Residue conservation scores

For each protein, we extract the sequence from the PDB file and then search for homologous sequences using BLAST against the non-redundant protein sequence database with an e-value cutoff of 0.01, percentage identity in the range of ≥ 35% and ≤ 95% and query coverage of 80%. To filter duplicates, we use CD-Hit and cluster the initial set of homologs (Y. Huang, Niu, Gao, Fu, & Li, 2010) at 95% sequence identity and then select only the representative sequences from each cluster. We perform multiple sequence alignment (MSA) with Clustal Omega (Sievers & Higgins, 2014) with default parameters on a randomly selected set of 150 representative homologs for each protein. Using Rate4Site (Pupko, Bell, Mayrose, & Glaser, 2002) with its default parameters for the evolutionary model (JTT) and rate inference method (Bayesian), we then calculate the conservation scores for each protein from its respective MSA file. Rate4Site reports the extent of conservation at a position as a z-score, where a lower score indicates higher conservation.

Network centralities

We render each protein structure as a coarse-grained system in which residues are represented by their C^{α} atoms and model the protein as a network in the following different ways.

i. Network based on distance cutoff. A protein is modeled as a coarse-grained system by representing individual residues by their C^{α} atoms and by adduing edges between residue pairs which are within a given distance cutoff. The network is unweighted, i.e. there are no weights on the edges connecting a pair of nodes in the network. In the present study, we explore cutoffs of 10-15 Å and observe, for a subset of features, the predictive performances to be very similar. Thus, we choose the cutoff as 13Å.

ii. Distance weighted network. The edge between a residue pair is weighted by their spatial proximity – in this case the distance between their C^{α} atoms. Such a network can be regarded as an interaction strength network – edges between spatially close residues are given higher weights than edges between distant residues.

iii. Network weighted by the correlation of inter-residue dynamics. We model each protein as a coarse-grained C^{α} system in which a residue pair is connected by a spring with stiffness varying inversely with the distance between the two residues (Eq. 4.2). Following diagonalization of the Hessian of this system, we use the first 20 low frequency normal modes vectors (V) and their corresponding eigenvalues (λ) to obtain the inverse Hessian with the following equation.

$$H^{-1} = \sum_{i=1}^{20} \lambda_i^{-1} V_i V_i^T \tag{4.5}$$

The H^{-1} is a 3N by 3N dimensional matrix, where N corresponds to the number of residues and it gives the correlations between residue fluctuations in the X, Y and Z directions. We then calculate the correlation between the fluctuations of residues *i* and *j* as

$$c_{ij} = \frac{trace\ H_{ij}^{-1}}{\sqrt{trace\ H_{ii}^{-1}\ trace\ H_{jj}^{-1}}} \tag{4.6}$$

In this above equation, H_{ij}^{-1} is a 3 by 3 block element of the inverse Hessian corresponding to residues *i* and *j*. It provides the correlation between the fluctuations of residues *i* and *j* in the X, Y and Z directions. H_{ii}^{-1} and H_{jj}^{-1} are the block elements corresponding to the fluctuations of residues *i* and *j*. The trace is the sum of the diagonal elements of each block matrix. We express the correlation matrix as a dissimilarity (or distance matrix) by subtracting each element from 1.

$$D_{ij} = 1 - c_{ij} \tag{4.7}$$

Thus, residue pairs with perfect positive correlations in dynamics would have a value of 0, while pairs with completely negative correlation would have a value of 2. The network thus created, has edges weighted by the extent of correlation in the fluctuation dynamics between residue pairs, such that residues having high correlations in their dynamics are connected by edges with higher weights than residues with low correlations.

iv. Network weighted by the interaction energy. The edges between residue C^α atoms are weighted by their interaction strengths obtained by using the Betancourt and Thirumalai (BT) potential (Betancourt & Thirumalai, 1999). We convert energies in the BT potential matrix into positive scores by calculating their exponential forms. Thus, more favorable interacting pairs (lower interaction energies) have larger weights.

For each of the above four network formulations, we calculate the following node centralities.

a. Betweenness. Node betweenness of a given node *N* is the number of shortest paths between any two nodes that pass through *N*. Nodes that are more central in a network have high node betweenness values (Freeman, 1977). For weighted graphs, the shortest path is the one with minimum edge weight (least cost). When calculating the node betweenness for the network constructs in *ii* and *iii*, we weigh the edges between residue pairs by their respective Euclidean distances and the distance-transformed correlations. However, for the network constructed using interaction energies in *iv*, we weigh the edges by inverse of the Boltzmann weights obtained from the interaction energies, so that the shortest path is the path with least energy.

b. Closeness. Node closeness provides a measure of the closeness of a given node to all other nodes and is expressed as the inverse of the sum of the lengths of the shortest paths between the given node and all other nodes (Bavelas, 1950; Sabidussi, 1966). Closeness may be considered as a measure of the time taken to transmit the information from a given node to all other nodes, sequentially. Like node betweenness, the shortest path has minimum edge weight. When calculating the node closeness, we take a similar approach to node betweenness in terms of assigning edge weights for the different network constructs.

c. Degree. The node degree is a measure of the number of edges connected to a specific node. For weighted graphs, it provides a measure for the strength of a given node by considering the cumulative sum of the weights of the edges connected to this node (Newman, 2004; Opsahl, Agneessens, & Skvoretz, 2010). The edge weights correspond to their importance. In the calculation of degree centrality for the distance-weighted network (construct *ii*), the importance of an edge for a given node falls off with distance (edges to spatially close residues are of greater importance than edges connecting to spatially distant residues). However, we assign greater importance to correlations in dynamics (network construct *iii*) that is, edge importance increases with increase in the correlation of the inter-residue dynamics. For networks constructed based on interaction energy (construct *iv*), the edge importance is higher for more favorable interactions.

d. Eigenvector. Eigenvector centrality measures the influence of a given node in a network. It is based on the principle that connections to high scoring nodes are more significant than connections to low scoring nodes (Newman, 2007). In assigning edge importance to the different weighted networks, we follow the same rule as in the case of degree centrality.

e. Page rank. The page rank centrality is a variant of the eigenvector centrality and is frequently used by Google to rank webpages rendered by the search engine (Page, Brin, Motwani, & Winograd, 1998). It assigns importance to a node not based only on the number of nodes it is linked to, but also based on the importance of the linked nodes and their centralities. We use the same approach in assigning edge importance to the weighted networks as in degree and eigenvector centralities.

Pocket residues

We use Fpocket (Le Guilloux, Schmidtke, & Tuffery, 2009) to predict cavities or pockets in all-atom protein structures and identify residues that are located in pockets. We consider a residue to be part of a pocket if any of the residue atoms are in contact with the voronoi vertices of the pocket.

Shortest path to catalytic residues (only for allosteric classifier)

Upon binding of effectors, allosteric residues transmit signals to functional binding sites via allosteric signaling pathways – chain of residues which lie in between the regulatory and active site. For identification of residues involved in effector binding, one of the features that we also consider is the shortest dynamically correlated path between a given residue and the active site. Our underlying hypothesis is that potential effector binding residues will have shorter paths that are dynamically more correlated than other residues.

By considering a protein as a system of C^{α} atoms with residues connected by Hookean springs with stiffness varying inversely with the square of distance (Eq. 4.2), we obtain the correlation between inter-residue dynamics (Eq. 4.6) and transform it into a dissimilarity matrix (Eq. 4.7). The protein is then modeled as a network wherein each residue pair within 13 Å is connected by an edge having edge weight equal to the distance-transformed correlation in dynamics obtained using Eq. 4.7. Using such a network formulation, we apply Dijkstra’s algorithm to calculate the shortest path between a given residue and any of the active site residues. In addition, we also consider the median shortest path from a given residue to all active site residues. It is to be noted that for the calculation of this feature, identification of active site residues is required.

4.2.4. Training, Validation and Test Datasets

We investigate the distribution of the number of known allosteric and active site residues in 105 proteins (144 subunits) (Fig C.1 and C.2). For allosteric prediction models, we divide our dataset of 105 distinct proteins into three groups based on the distribution of allosteric residues (Fig. C.1). *Group 1: proteins with 2-10 allosteric residues, group 2: 11-19 allosteric residues and group 3: 20-44 allosteric residues.* From each group, we randomly sample 90 percent of the proteins to create training and 10 percent to create test sets. We thus have 94 proteins (129 protein subunits) for training and the remaining 11 proteins (15 protein subunits) for testing. In terms of the number of allosteric residues, our training set for allostery has a total number of 1288 allosteric and 44,946 non-allosteric residues, while our test set has 167 allosteric and 6.607 non-allosteric residues. We compile the residue-level features calculated for the training set into a consolidated training file. The consolidated training file is imbalanced (not having an equal number of allosteric and non-allosteric residues) and we use it to create 10 pairs of balanced training and validation sets. For each of these pairs, the training set has features corresponding to 1000 allosteric residues sampled randomly without replacement and an equal number of non-allosteric residues, whereas the validation set has the remaining 288 allosteric and an equal number of non-allosteric residues.

For active site residues, we follow a similar protocol as described above for allosteric residues. Based on the distribution of the number of active site residues in all proteins (Fig. C.2), we also divide this dataset into 3 groups: *Group 1: proteins having up to 10 active site residues, group 2: 11-29 active site residues, and group 3: more than 30 active site residues.* For each group, we randomly sample 90 percent of the proteins for training and 10 percent for testing. This gives us 94 proteins (125 subunits) for the training and 11 proteins (19 subunits) for testing, which includes 1,018 labelled active site residues and 47,476 non-active site residues in the training set and 180 labelled active site residues and 4,344 non-active site residues in the test set. We compile the residue-level features calculated for the training dataset into a consolidated training file and then create 10 pairs of balanced training and validation sets – each balanced training set having features corresponding to 800 positive labels randomly sampled without replacement and 800 negative labels, while each balanced validation set having the remaining 218 positive labels and a randomly sampled 218 negative labels.

4.2.5. Machine Learning Models

We use the TreeBagger module (www.mathworks.com/help/stats/treebagger.html) in Matlab, an implementation of the random forest algorithm, to develop separate predictive models for allosteric and active site residues. For each type, we first train the algorithm with each of the 10 balanced training sets and then test it with the corresponding validation set. Thus, we have 10 models each trained and tested using a different dataset. Our random forest implementation uses 100 trees and a minimum of 2 leaves at each node. To optimize the performance of each model we then include misclassification costs (penalty for false negatives and false positives) in our model with a cost matrix. Using a brute force approach, we verify the classification performance using different cost combinations for false positives and false negatives in the range of 0.1 to 1 in steps of 0.1 and select the combination that maximizes the Matthews correlation coefficient (MCC) for a given model. Including such costs in each model slightly improves the performance as shown in Fig. C.3 and C.4.

4.2.6. Feature Selection

We exclude all features having feature importance below 0.3.

4.2.7. Prediction on Test Dataset

We weigh the probability score assigned to each residue by a particular model by its MCC and then obtain a cumulative weighted score for each residue in a protein from an ensemble of 10 models with the following equation.

$$Score_{weighted}^{i} = (\sum_{N=1}^{10} MCC_N\, S_N^i\,) \,/\, \sum_{N=1}^{10} MCC_N \qquad (4.8)$$

Here, MCC_N is the MCC of the *N*th model and S_N^i is the score of *i*th residue assigned by the *N*th model. We use this formulation of weighted scores on the models trained for allostery and also on those trained for active site detection to identify the most probable allosteric and active site residues, respectively.

4.3. Results and Discussion

We use a previously compiled dataset by Greener and Sternberg for the allosteric prediction tool, AlloPred (Greener & Sternberg, 2015). The compiled dataset has information on both allosteric and active site residues and thus provides a basis for a scheme to predict both allosteric and active site residues. In our approach, we compile a diverse set of features based on amino acid physicochemical properties, evolutionary conservation, protein structural geometry and supplement them with features that relate to the dynamic nature of proteins. Since dynamics is pivotal in maintaining the functional and regulatory roles in proteins, we presume that including such information will improve the detection of residues important for regulation or substrate modification.

Our goal is to develop prediction models for active and allosteric site residues (AR-Pred) using a common subset of features. To this end, we first calculate the features described in Methods for all proteins in the dataset and exclude proteins for which any of the features could not be calculated. For multimeric proteins in our dataset, feature calculations were performed on each subunit after splitting the multimer into its respective subunits. Feature calculation renders a feature vector of size M (M is the number of features) for each residue. A single protein having N residues can thus be described by an N by M matrix of features. Next, we divide the dataset of protein structures into distinct training, validation and test sets based on the distribution of the number of active and allosteric residues (Fig. C.1 and Fig. C.2). For each prediction class (allosteric and active site), we create 10 balanced training and validation sets. We train a random forest classification model on each training set and verify its performance on the respective validation set. Consequently, we have 10 models trained and validated for each prediction class. We use this ensemble of 10 models to make predictions on test sets created for each class. Details concerning the creation of training, validation and test datasets are provided in Materials and Methods.

The prediction models for allostery and active sites collectively constitute AR-Pred. First we compare the performances for AR-Pred's allosteric and active site prediction models for their respective validation sets. Second, we focus on the features which were important determinants for the models' performance. Third, we compare the performance of our models with other existing tools. Fourth, we compare the distribution of Euclidean distance of the predicted active and allosteric sites from the known sites with that of sites chosen at random. Fifth, we closely inspect the predictions made by the allosteric models on one of the test proteins to verify the existence of false positives. Finally, we consider one protein common to the test data sets of allostery and active site to verify the localization of predicted allosteric and active site residues and show a connection between the intrinsic dynamics of these sites.

4.3.1. Performance on Validation Sets

Figure 4.1 and Figure 4.2 show the metrics for the average performance of the 10 models on the validation data set for active site and allosteric site respectively. It is seen that the average performance of the models for active site is superior to that of allosteric site. The performance for each of the 10 models for active and allosteric sites is shown in Fig. C.5 and C.6, respectively.

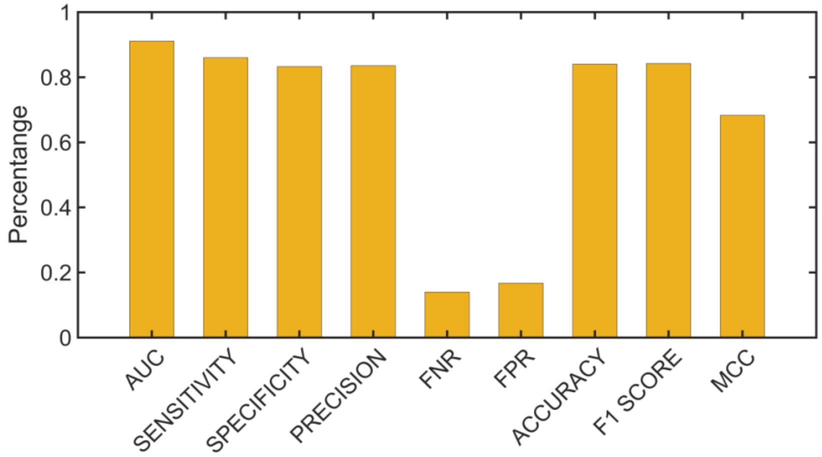

Figure 4.1. Metrics describing the performance of active site models. Median metrics calculated across the ten models for active site prediction. The metrics were calculated on the validation set corresponding to each model.

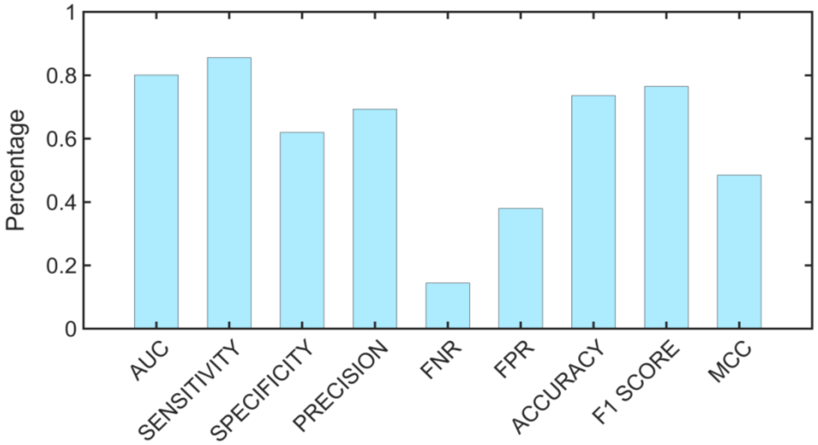

Figure 4.2. Metrics describing the performance of allosteric site models. Median metrics calculated for the ten models for allosteric site prediction. Calculations were performed similar to the metrics of the active site models.

It is interesting to note a greater inter-model variability in sensitivity and specificity for allostery than for active sites. The models for allostery also exhibit higher variance for false positive rate (FPR) than the models for active sites. Both of these indicate that predicting active sites is more reliable than predicting allosteric residues. The receiver operating characteristic (ROC) curves for active site and allosteric models are shown in Fig. C.7 (A and B) and the area under the curve (AUC) for active sites is higher than for allostery.

At first this suggests that the predictive nature of our models for allostery is more random than the models for active site, however one must consider that active sites are substantially better known and have been investigated more exhaustively in comparison with allosteric sites. Active sites have long been exploited as popular drug targets by pharmaceutical industries and thus, their identification is supported by a plethora of experimental evidence. There have been relatively fewer studies on allostery which may indicate that quite a few allosteric sites in a protein remain unknown, explaining the nature of the inter-model variance.

4.3.2. Feature Importance

The feature importance for the two model classes is shown in Fig. 4.3 and Fig. 4.4. For both the models of allosteric and active site predictions, residue conservation scores are the most significant determinants for the models’ predictive performance. We also notice that the residue node betweenness centralities obtained by representing proteins as unweighted networks and adding edges between residues which are within 13 Å is rated as the second most important feature for both allosteric and active site residues. More importantly we observe features related to the residue-level dynamics ranked in the top 10 important features for both model types. It is seen that for both predictors, the resilience of residues to external perturbations described by the dynamic flexibility index (DFI) is also listed as one of the top 10 important features. However, the extent of residue mobility described by mean-square fluctuations (MSF) is a more important factor for allostery than for active site residues.

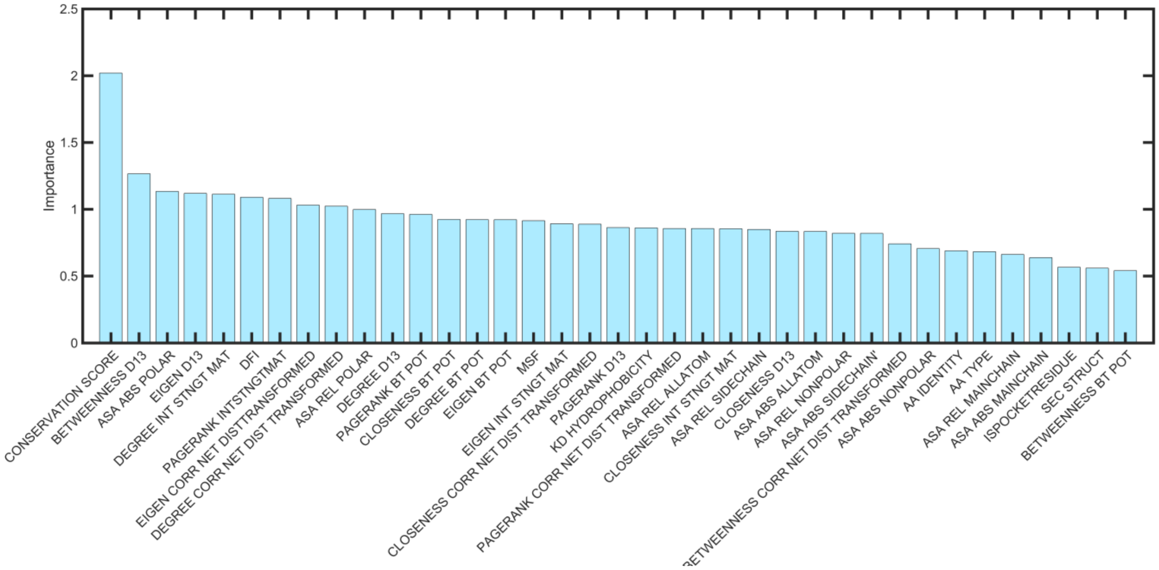

Figure 4.3. Feature importance for active site models. The median feature importance calculated across the 10 models for active site prediction are shown. The features are ordered by their importance.

Besides, features which relate residues closely to the active site such as shortest path to the active site residues and the dynamic response upon perturbing the active sites are important determinants for allostery, as one might expect.

Figures 4.3 and 4.4 also suggest that solvent accessibility is more important for determining active site residues than allosteric residues. Active sites are often concealed in the hydrophobic core, intermittently allowing access to substrates through changes in conformations. However, no strict pattern in terms of solvent accessibility has been observed in the case of allosteric sites. Also, features relating to the physicochemical properties of amino acids such as amino acid hydrophobicity and their secondary structures are important predictors for the active site residues.

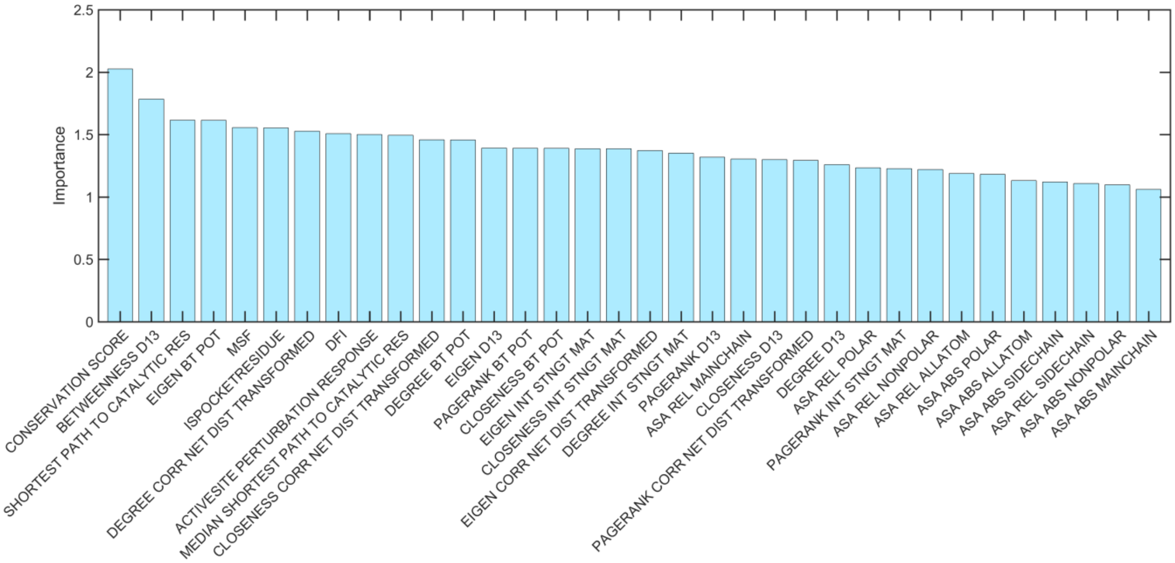

Figure 4.4. Feature importance for allosteric site models. The median feature importance calculated across the 10 models for allosteric site prediction are shown. The features are ordered by their importance.

4.3.3. Predictions on Test Datasets

Active site prediction

We have mapped the predictions for active site residues from AR-Pred and compare it with the known active sites for 6 proteins in the test dataset. We rank residues by their weighted probability scores and for each protein we show only the top 15 residues. In Fig. 4.5, the known active site residues are colored orange, the predicted ones green, and the predicted true positives as red spheres. It is seen from the figure that in the predicted pool of residues, a minimum of 2 residues are true positives in all 6 cases, while in two cases (A and D) there are 3 true positives and 4 true positives in E. For 4 cases, the top 15 predicted residues are tightly clustered around the known active sites (Fig. 4.5 A, B, E and F) while, in two cases (Fig. 4.5 C and D) the predicted residues are more scattered. Some of this might possibly be alternative binding sites for ligands, metal ions or even for co-factors.

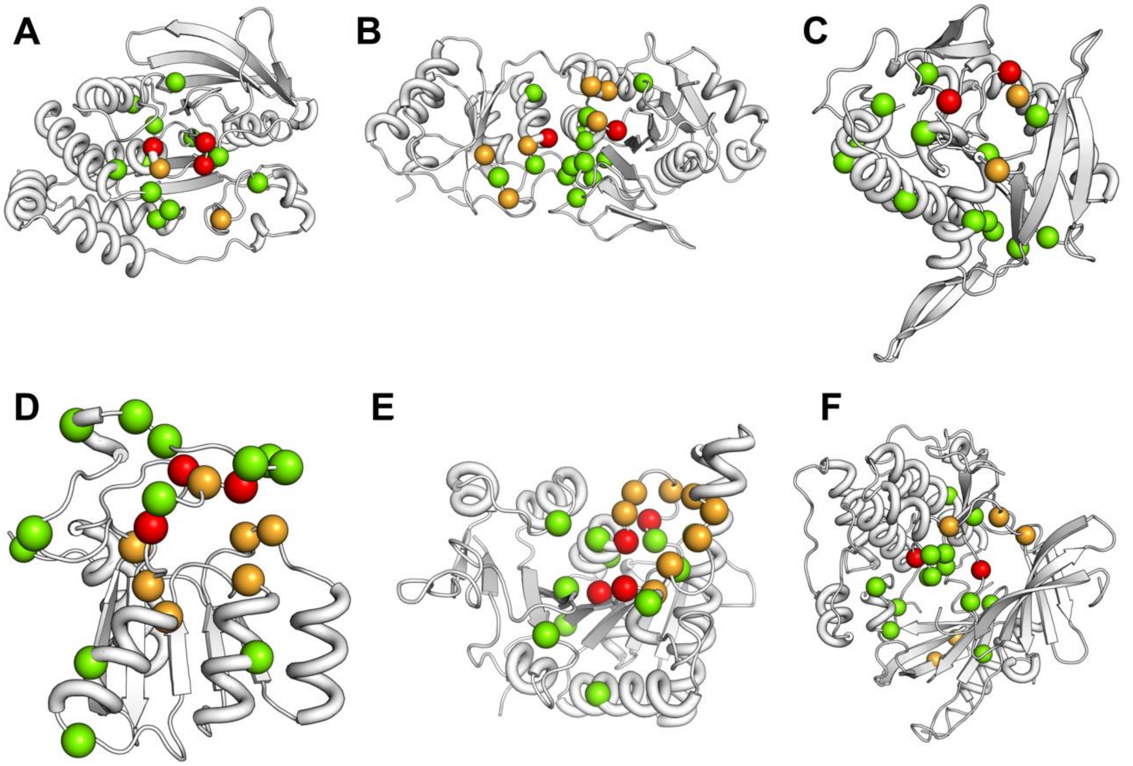

Figure 4.5. Predictions on active site test dataset. Top 15 predicted residues for active sites are shown for the 6 proteins in the test dataset. The reported active sites are colored orange, the predicted true positives as red and the putative active sites predicted are shown as green spheres. The proteins for which these predictions were made are: A. Protein tyrosine phosphatase 1B (PDB 1T49, chain A), B. L-Asparaginase I (PDB 2HIM, chain A), C. Uracil phosphoribosyltransferase (PDB 1JLR, chain A), D. Deoxycytidylate deaminase (PDB 2HVW, chain A), E. UMP Kinase (PDB 2V4Y, chain A), F. AKT 1 (PDB 3O96, chain A).

To verify that the predictions are not random, we perform two tests. First, for all proteins in the test data, we consider the shortest distance between the heavy atoms of the top 15 residues and any of the known active site residues and plot their distribution. Second, we perform 50 iterations of random residue selection by picking 15 residues randomly from each protein and verifying the distribution of the shortest distances between the heavy atoms of these residues and any of the known active site residues, in each iteration. Such an analysis will tell us how closely clustered the predicted active site residues are around the known active site residues. The results are shown in figures C.8 and C.9, respectively. In Fig. C.8, we observe the highest peak near 2.5Å and the distribution has a negative gradient at 5Å suggesting that the predicted residues are in close proximity to the known ones. Fig. C.9 suggests that the predictions are not random since the peaks are much sharper for the predicted residues (red) than for the random ones (blue). It is also worth noting that there is a smaller peak for the predicted residues, around 20Å, suggesting a bimodal distribution of the shortest distances and the presence of alternative functional binding sites for a given protein.

Allosteric site prediction

In Fig. 4.6, we have mapped the predicted allosteric residues by AR-Pred onto the structures of 6 proteins (showing cyan colored spheres for known allosteric residues, green for predicted and red for predicted true positives). It is seen that in five out of the 6 cases (Fig. 4.6 A, C, D, E and F) the predicted residues are tightly clustered around the known ones. We also observe a higher number of true positives for the allosteric predictions: a maximum of 11 residues are true positives out of the top 15 (Fig. 4.6F). One of the six proteins (Fig. 4.6B) shows complete mismatch between the predicted and known allosteric residues. The protein is DAH7PS from *Thermotoga maritima* (PDB 3PG9) which is involved in the shikimate pathway, essential for the synthesis of aromatic amino acids. We further verify the significance of the predicted residues for this protein and investigate whether they constitute potential allosteric pathways.

DAH7PS has two domains – an N-terminal regulatory domain and a C-terminal catalytic domain (Fig. C.10) and catalyzes the condensation between the substrates phosphoenolpyruvate (PEP) and D-erythrose 4-phosphate (E4P) to form 3-deoxy-D-arabinoheptulosonate 7-phosphate (DAH7P). It is known to be regulated by tyrosine which binds to the regulatory domain and reduces affinity for both substrates (Cross, Dobson, Patchett, & Parker, 2011).

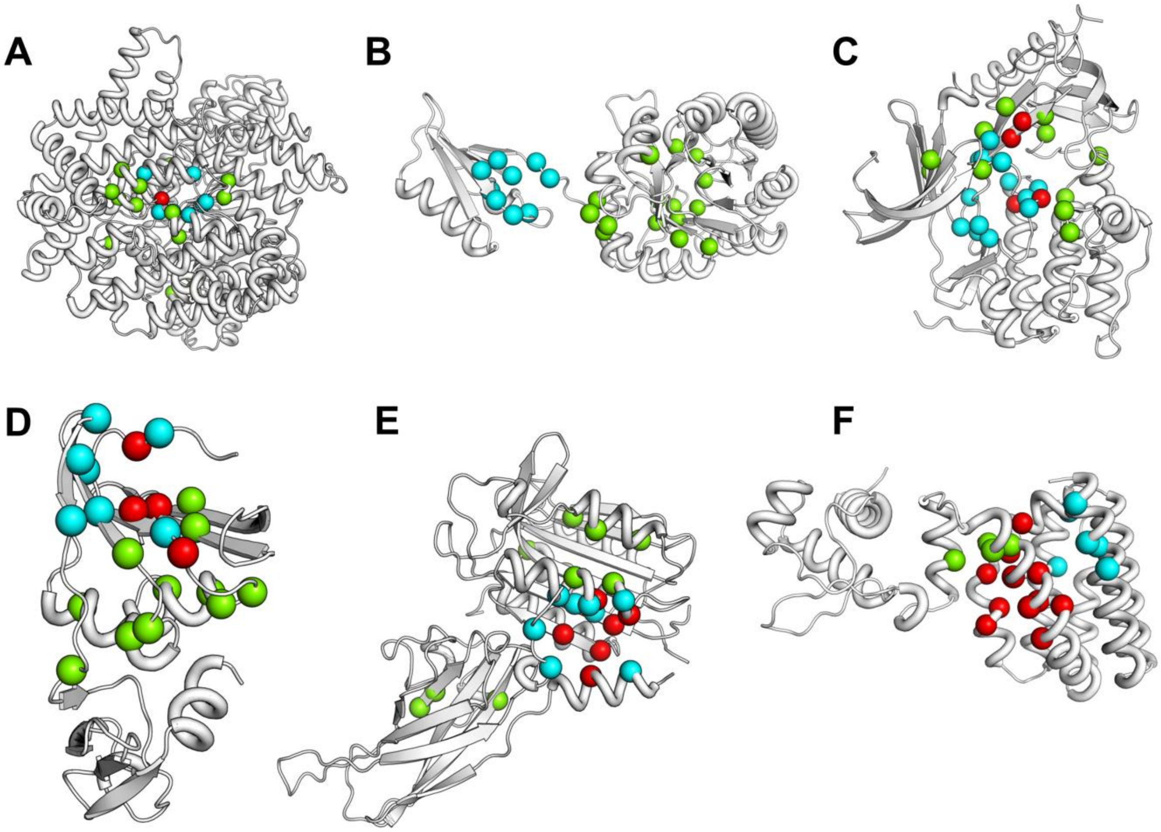

Figure 4.6. Predictions on allosteric site test dataset. The top 15 residues predicted for allosteric sits are shown for the 6 proteins in the corresponding test dataset. Previously reported allosteric sites are shown in cyan, the predicted true positives in red and the putative predicted sites as green spheres. The proteins considered are: A. Phosphoenolpyruvate carboxylase (PDB 1FIY, chain A), B. DAH7P synthase (PDB 3PG9, chain F), C. AKT 1 (PDB 3O96, chain A), D. Aspartate transcarbamoylase (PDB 2BE9, chain B), E. MALT1 (4I1R, chain A), F. FadR (1H9G, chain A).

Upon binding, tyrosine induces a displacement in the position of the β2-α2 loop in the catalytic domain (colored in purple in Fig. C.10). In Fig. C.11 (A, B, C, D, E and F), we have mapped the predictions for the top 5, 10, 15, 20, 30 and 40 allosteric residues. It is worth noting that one of the top 5 predicted residues (Fig. C.11A) is located on the β2-α2 loop of the catalytic domain. We observe that with an increasing number of predicted residues, more residues are predicted on the linker connecting the regulatory and catalytic domains and also on the β2-α2 loop. Also, it can be seen that in the top 40 predicted residues (Fig. C.11F), 3 residues are on the regulatory domain of which, 2 are true positives. Figure C.11F also describes two putative allosteric pathways (red and blue) originating at the regulatory domain and leading to the β2-α2 loop of the catalytic domain. It is seen that the two pathways are on either side of the active site (shown as orange spheres). Interestingly, some of the predicted residues are in vicinity of the active site residues. Since a protein's dynamic nature introduces the possibility of multiple allosteric pathways, these residues may be part of such pathways to control activity or even the dynamics of the catalytic site.

To verify the extent of randomness in our predictions for allosteric residues, we perform an analysis similar to that of the predicted active site residues: a) we probe the distribution of the shortest Euclidean distances between the heavy atoms of the predicted residues and any of the true allosteric residues, and b) we compare the distributions for the predicted residues against a pool of randomly selected residues. In Fig. C.12, we plot the distribution of the shortest distances for the top 15 predicted residues for all proteins in the allosteric test data and it shows that the peak density for the predicted residues distances is close to 6 Å. This suggests that a major fraction of the predicted residues are tightly clustered around the experimentally verified residues. However, as shown for DAH7PS some predicted residues constitute allosteric pathways and are of equal importance. Such residues can be located away from the effector binding site and can skew the distribution plot. Fig. C.13 compares the distribution of the shortest distances for the top 15 predicted allosteric residues in all proteins with 15 randomly chosen residues. The comparison is carried out for 50 iterations. It is clearly seen that in all iterations, predicted residues are associated with sharper peaks at shorter distances than the randomly picked residues, further confirming that the predictions are not random.

4.3.4. Comparisons with Existing Methods

Active site prediction

We compare our AR-Pred's predictions for active site with the results from four other methods: Concavity (Capra et al., 2009), AADS (Singh et al., 2011), POOL (Tong et al., 2009) and FOD (Brylinski et al., 2007). For each method, we rank the predictions by their scores and plot the percentage of true positives predicted (ordinate) for a certain percentage of the ranked predictions (abscissa) referred to here as percentage threshold. Our aim is to then systematically compare the percentage of predicted true positives under a particular percentage threshold from each of these methods with our method. Three out of the four methods (Concavity, POOL and FOD) assign scores to residues in a protein based on their propensity for being active site residues. However, AADS predicts active site pockets, where each pocket contains multiple residues. To make comparisons with such pocket based methods, we rank residues based on the rank of their pocket. Thus, all residues in a given pocket are assigned the same rank. Then, we filter residues which appear in multiple pockets by considering them only as part of the higher ranked pocket and consider the pool of residues in every threshold percent to identify the number of predicted true positives.

For each protein in the test dataset, Figure 4.7 compares the prediction performance of AR-Pred (red curve) with the above mentioned methods. When considering the percentage of true positives in the top 10 percent of the predicted residues AR-Pred outperforms FOD in

11 out of the 19 cases and AADS in 14 out of 19 cases and we observe similar performance for 5 and 3 proteins, respectively. At the same threshold, we perform better than Concavity in 5 cases and show similar performance in 6 cases. In the case of POOL, we have results only for 18 out of the 19 cases (4JAF gave errors). We see similar performance for POOL as that of Concavity, with 6 cases of improved performance and 5 cases of at-par performance. When considering a threshold of the top 30 percent of the predicted residues, our method performs better than Concavity and POOL in 4 and 6 cases, respectively and we observe similar performances in 7 and 8 cases, respectively. Upon comparing with FOD and AADS, at 30 percent threshold, we perform better in 9 and 12 cases and observe similar performance in 8 and 3 cases, respectively. Table C.1 shows the percentage of proteins from the test data for which our method predicts the same or higher numbers of true positives than the four other methods at different threshold percentages. AR-Pred shows at least similar or better performance compared to Concavity, AADS, POOL and FOD for a median 57.9%, 79.0%, 66.7% and 84.2% of the test proteins, respectively for the threshold of 10–50 percent of the predictions. These results clearly indicate that including protein-dynamics information together with the physiochemical, structural and evolutionary features, leads to improved detection of active site residues.

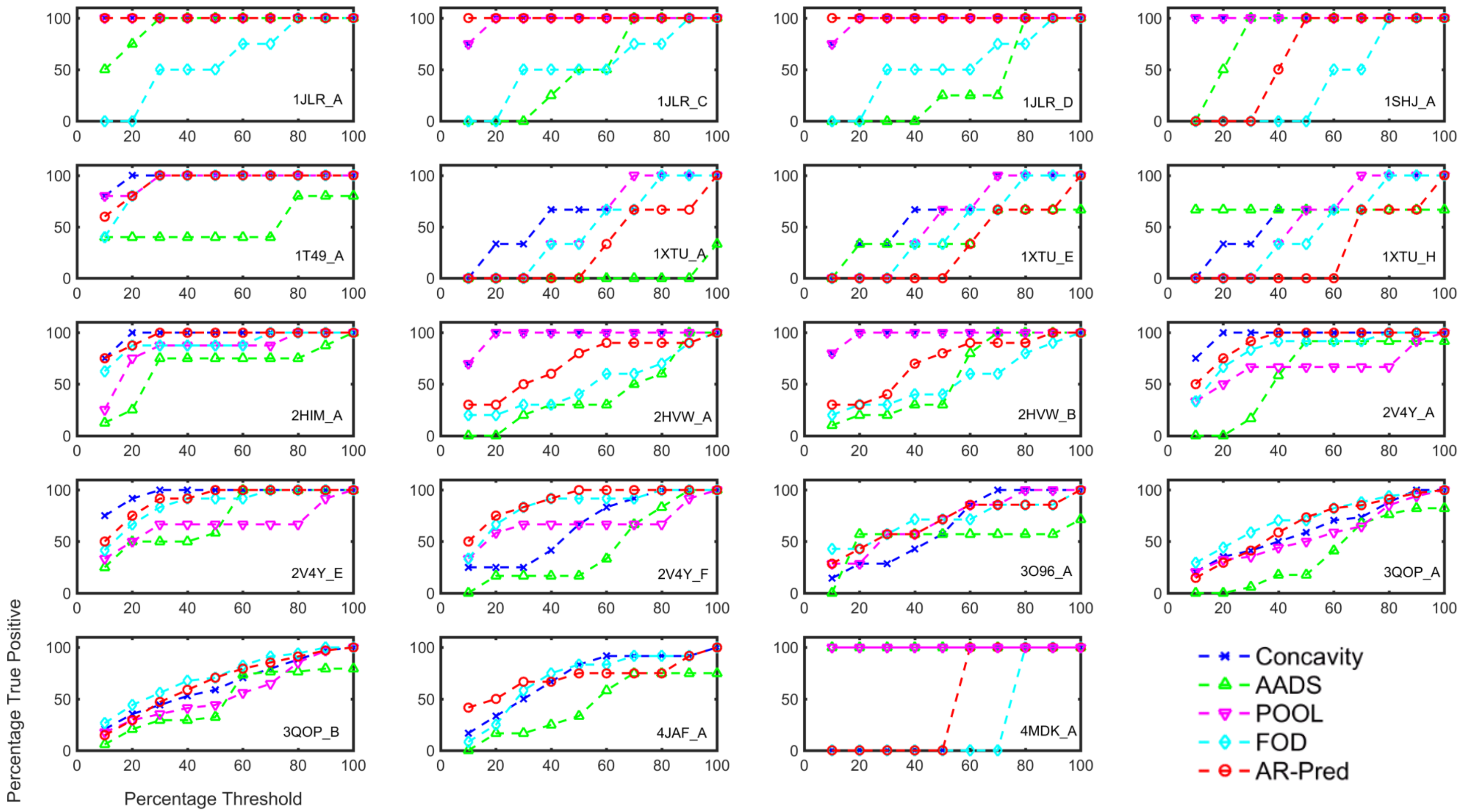

Figure 4.7. Comparison of the AR-Pred's predicted active sites with existing methods. Prediction comparisons are made between the AR-Pred's active site predictive models and four existing methods (Concavity, AADS, POOL and FOD) for each protein in the test data. On the X-axis we have the percentage of predictions considered as a threshold and plot the percentage of true positives predicted under a certain threshold by each method.

Allosteric site prediction

We compare the predictive power of our method with three existing methods: AlloPred (Greener & Sternberg, 2015), AlloSitePro (K. Song et al., 2017) and SPACER (Goncearenco et al., 2013). AlloPred is the source of the dataset we have used to develop our prediction models. AlloSitePro is an upgraded implementation of AlloSite (W. Huang et al., 2013). SPACER uses binding leverage, the ability of a binding site to couple with the intrinsic motions of a protein to identify potential allosteric sites and makes predictions at the residue-level; whereas, both AlloPred and AlloSitePro predict pockets. To perform comparisons, we follow the same procedure as above for the active site prediction models.

With a threshold of 10 percent of the predicted residues, we observe gains in true positives against AlloPred, AlloSite and SPACER for 8, 9 and 9 proteins and similar performances for 4, 5 and 4 proteins, respectively (Fig. 4.8). In 7 cases, our method performs better than all the three methods, at the 10 % threshold. Table C.2 shows the percentage of proteins in the test data for which our method shows better or comparable true positive rates for different threshold percentages. When compared to AlloPred, AlloSitePro and SPACER, our method gives comparable or better predictions for a median 80, 93.3 and 86. 7 percent of the test files, respectively. These results confirm our underlying premise – that including dynamic information with other features leads to improvements in the prediction of allosteric residues.

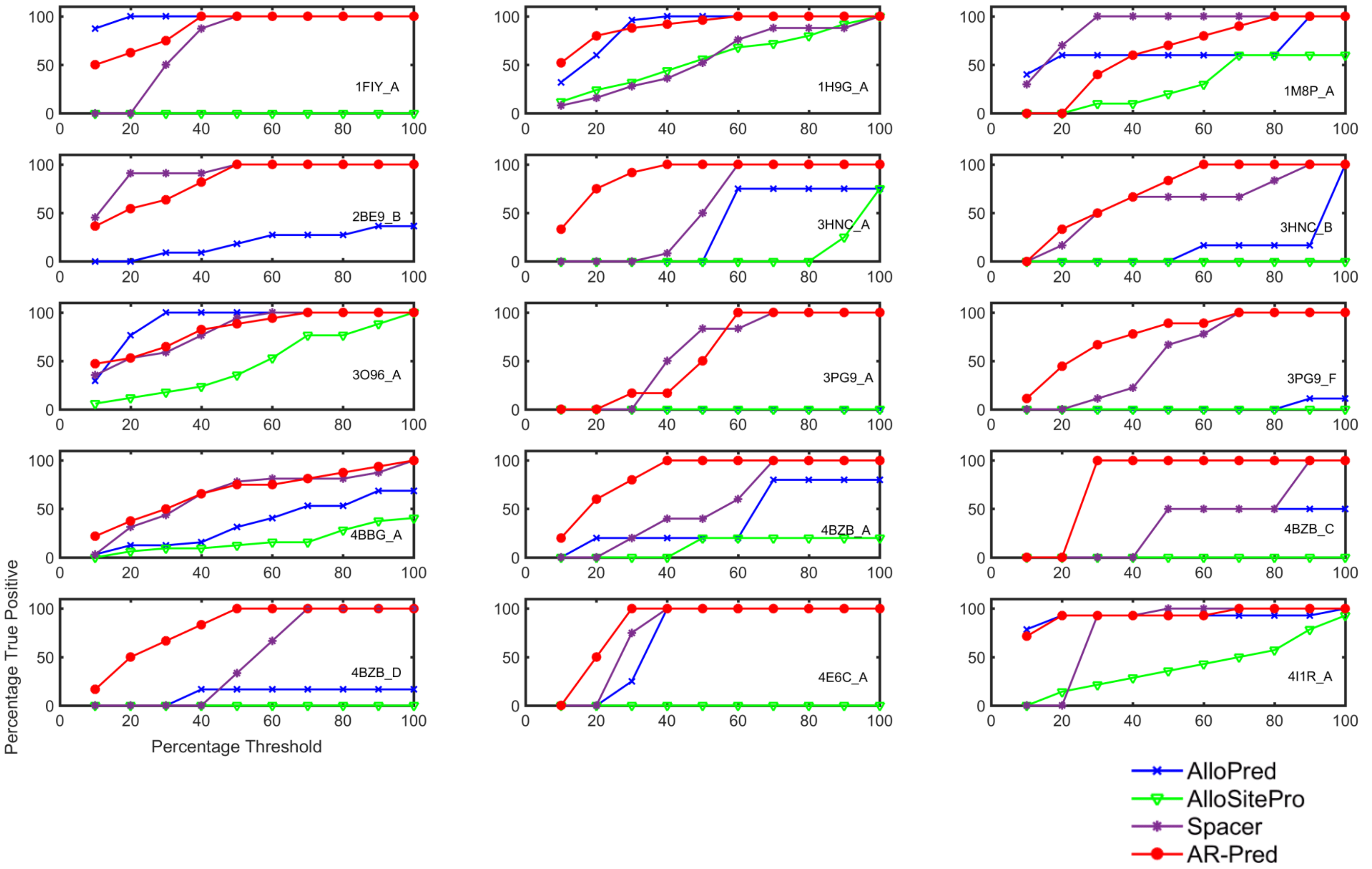

Figure 4.8. Comparison of the AR-Pred's allosteric site predictions with existing methods. We compare the predictive performance for our allosteric prediction against three existing methods (AlloSite, AlloPred and SPACER) for each protein in the test data. The abscissa and ordinates have same descriptions as in Fig. 4.7.

4.3.5. Investigation of False Positives in Allosteric Predictions

The protein aspartate transcarbamoylase (ATCase) from *Sulfolobus acidocaldarius* ATCase plays a vital role in the pyrimidine biosynthesis pathway, catalyzing the carbamoylation of the α-amino group of L-aspartate by carbamoyl phosphate and forming N-carbamoyl-L-aspartate and orthophosphate. It is a heteromeric structure comprised of two

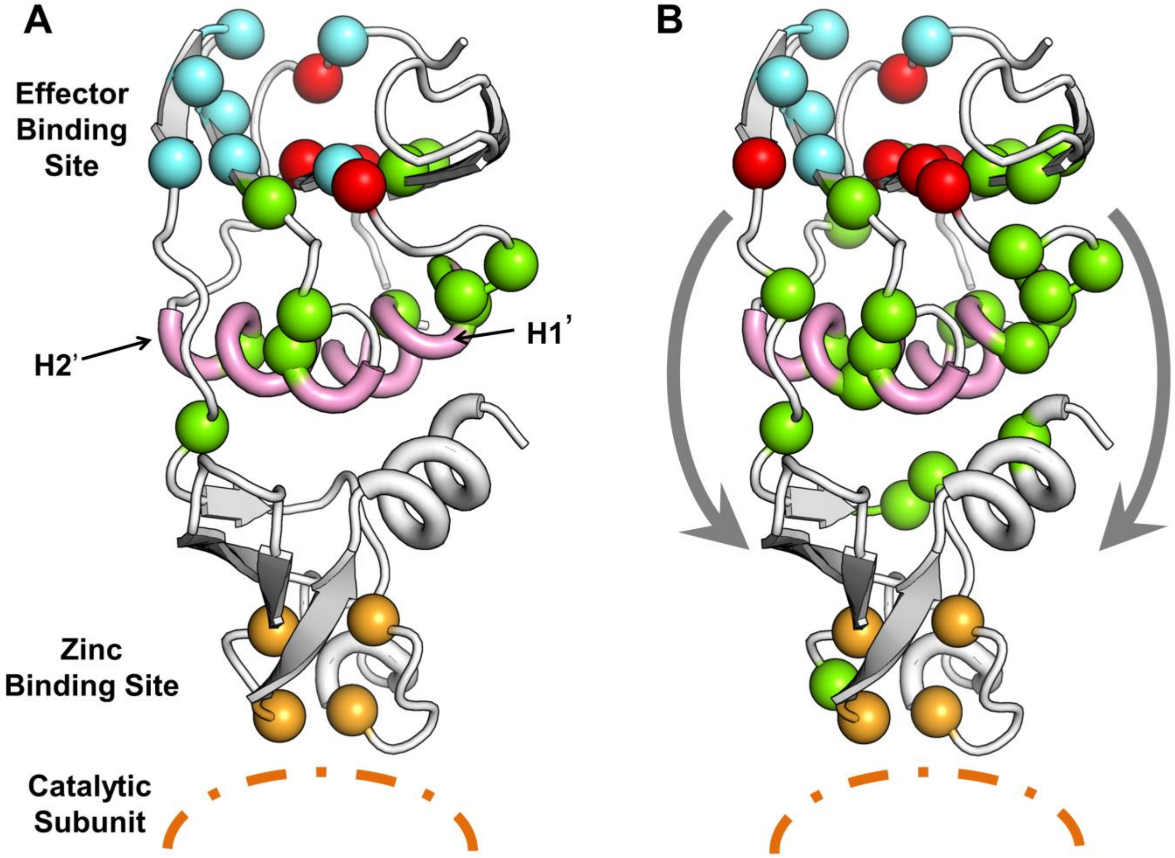

Figure 4.9. Potential allosteric pathways for aspartate transcarbamoylase (PDB 2BE9). Predictions made with AR-Pred's allosteric model for the top 15 allosteric residues (A) and top 30 allosteric residues (B) are shown. The two helices previously proposed to play a key role in transmitting allosteric signal from the effector binding site to the catalytic site are colored in pink. The zinc binding site is shown in orange. The reported allosteric residues are colored in cyan, the predicted true positives in red and the putative allosteric residues predicted by AR-Pred are shown as green spheres. The two proposed pathways are described in (B) by the two arrows.

chains, catalytic and regulatory (Lipscomb & Kantrowitz, 2012). While the catalytic chain comprises aspartate and carbamoyl phosphate binding domains, the allosteric chain has the allosteric domain which binds to regulators and zinc binding domains, which makes contact with the catalytic subunits.

We consider the regulatory chain of the enzyme (PDB 2BE9, chain F) and the predictions made for the allosteric residues. In Fig. 4.9 we show the top 15 (Fig. 4.9A) and top 30 (Fig. 4.9B) allosteric residues predicted for the protein. Previously, Vos *et. al* (De Vos, Xu, Aerts, Van Petegem, & Van Beeumen, 2008) compared the crystal structures for the CTP (allosteric regulator) bound and unbound structures for the *Sulfolobus acidocaldarius* ATCase and observed changes to the conformation of the bound form relative to the unbound form. Based on these observations, the authors proposed two allosteric pathways that transmit the effector binding signal to the catalytic subunits. We have shown the direction of these pathways with arrows (Fig. 4.9B). The H1’ and H2’ helices (shown in pink) show conformational deformations upon effector binding and hence, are considered critical for the allosteric signal transmission. For the top 15 predicted allosteric residues (Fig. 4.9A), we have 4 true positives (red spheres) while, 6 predicted residues lie on the H1’ and H2’ helices. Upon considering the top 30 predicted allosteric residues, we observe an increase in the number of residues on the two helices. It is interesting to note that the predicted residues align closely to the two proposed pathways and one of the residues in the pathways (Fig. 4.9B) is in close proximity to the catalytic subunit. Such residues may be regarded as “sink” or “terminal” residues in an allosteric pathway in which the “source” is the effector binding site.

4.3.6. Overlap between Allosteric and Active Site Residue Predictions

One of the proteins common to our allosteric and active site test structures is AKT1, a serine/threonine AGC protein kinase from human (PDB 3O96) associated with the PI3K/AKT and other signaling pathways. AKT1 contains an N-terminal PH domain, inter-domain linker, a kinase domain and a C-terminal domain often referred to as the C-terminal hydrophobic motif (Fig. 4.10A). PH domain binds phosphatidylinositide and directs the translocation of the protein from cytosol to the plasma membrane. The kinase domain contains the catalytic site responsible for phosphorylation and binds ATP (J. Yang et al., 2002). We use this protein structure to investigate the extent of agreement between the predicted and known allosteric and active site residues. By dividing the proteins into cohesive units that move as rigid bodies (McClendon, Kornev, Gilson, & Taylor, 2014), we also learn about the localization of the predicted residues with respect to these structural blocks.

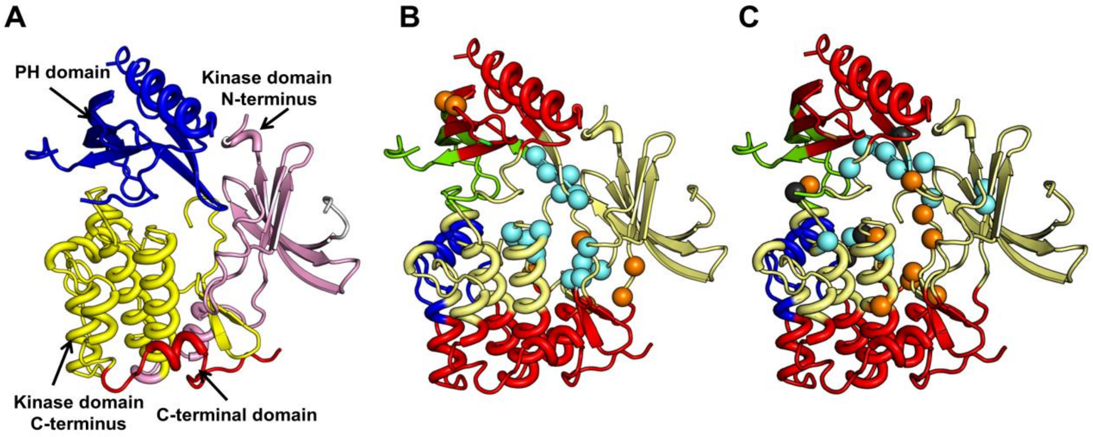

Figure 4.10. AKT 1 (PDB 3O96) domains, communities and predictions. (A) The three domains of AKT 1 are described. The kinase domain is split into its respective N and C-terminus domains. Reported (B) and AR-Pred predicted (C) allosteric and active site residues are shown. The protein backbone is colored based on its division into four dynamic communities. Regions in the same color show highly correlated motions, indicating they are rigid elements in their dynamics. The allosteric residues are in cyan, the active site residues in orange. In the predictions made by AR-Pred, residues which were predicted both as allosteric and active sites are shown in gray.

First, we divide the protein into dynamic cohesive units, also referred to as dynamic communities. To do this, we reduce the protein into a coarse-grained C^{α} representation and calculate the inverse Hessian for the elastic potential of the system using the first 20 low frequency normal modes with Eq. 4.5. Next, we calculate the correlation between residue-dynamics and express the inter-residue correlation matrix as a dissimilarity matrix using equations 4.6 and 4.7, respectively. Then, we identify dynamical structural blocks using the method described by Danon (Danon, Díaz-Guilera, & Arenas, 2006), dividing the protein into four dynamic communities.

Figures 4.10B and 4.10C compare the known active and allosteric site residues (B) with the top 15 predictions made by our models (C). Upon verifying the dynamic communities, it is seen that the kinase domain is divided into two communities - red and yellow. The rigid unit in the C-terminus of the kinase domain (in red at the bottom) shows dynamic coordination with the PH domain at the top, together forming one community. 4 out of the top 15 predicted allosteric residues coincide with the known ones, while we see an overlap of 2 residues in the active site. A strikingly common feature shared between the predicted and true allosteric residues is their location on the same dynamic communities, suggesting that both the predicted and known sites are highly correlated in their dynamics. It is even more interesting to notice that some of the predicted active site residues are actually reported to be allosteric. On closer observation, we find some of these residues are neighbors of residues that form the active site. This could make their feature profile very similar to that of the active site residues, making it hard for our models to distinguish between them. This suggests that a residue's functional classification is strongly influenced by its neighboring residues. Terminal or sink residues in an allosteric pathway, which are proximate to the active site, may not strictly be only allosteric. Their physicochemical, structural and dynamic properties may strongly correlate with active sites, even presenting them as potential functional binding sites. Based on these criteria, a strict classification of residues as allosteric or active site may not always be feasible owing to the influence of neighboring residues. This raises a few intriguing questions: could sharing a similar feature profile with active site residues introduce a constraint on a residue's rate of evolution? It is also interesting to consider whether some of these residues might eventually evolve into active site residues.

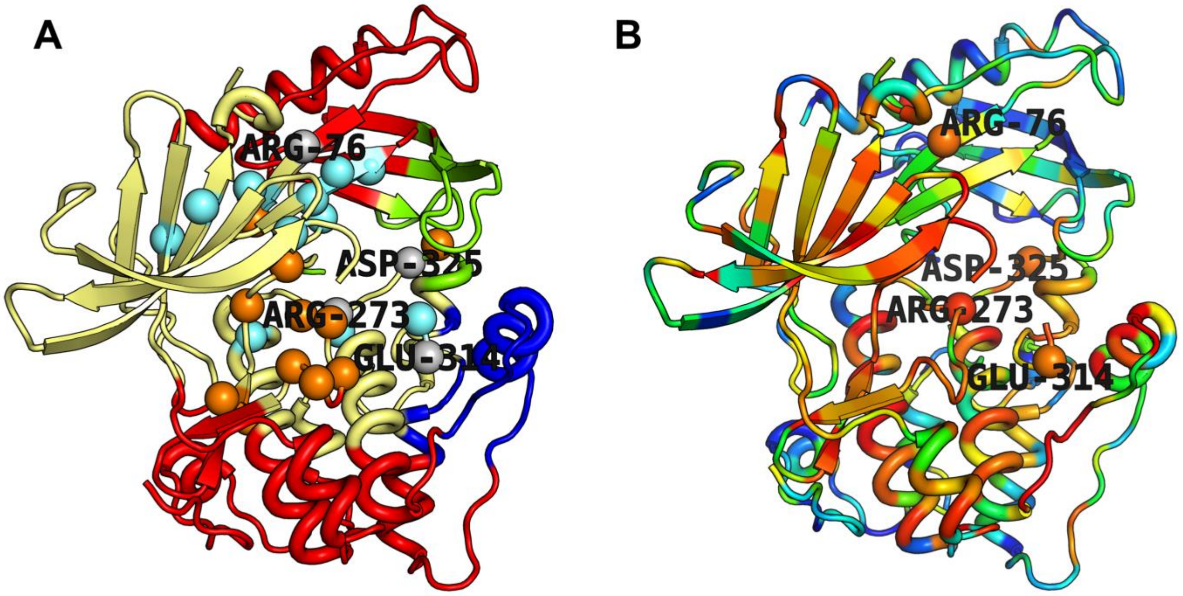

Figure 4.11. Overlap between allosteric and active site residue predictions. (A) Four residues predicted by AR-Pred to be both allosteric and active site are shown in gray and labelled. The coloring scheme is same as in Fig. 4.10. (B) The protein is colored by its evolutionary conservation, with the color scale varying from red to blue – most conserved to least conserved.

Our model predicts four residues (shown in gray) as both allosteric and active site residues (Fig 4.10C and 4.11A). Two of these residues are located on the boundary of a dynamic community pair. We hypothesize that these residues are examples of cases where, a strict classification scheme is not applicable. These residues may be classified into either category. Previous studies have shown that active sites of the proteasome can allosterically regulate each other's activity (Kisselev, Akopian, Castillo, & Goldberg, 1999). Other studies have indicated the presence of intrasteric active sites to which a short peptide, mimicking the substrate in vicinity of the active site, binds and regulates the activity of the active site (Kobe & Kemp, 1999). Such studies suggest that active sites could self-regulate their activity which, in a sense, is closely related to allostery. The residues which our model predicts to be both functional and regulatory sites could then be such self-regulating residues. Owing to their location at the boundaries of dynamic communities, they could also play the key role of allosteric signal transmission between communities. We further confirm the functional importance of these residues by investigating their evolutionary conservation. Fig 4.11B confirms that these residues have strong conservation. More importantly, three of these residues, Arg76, Asp325 and Glu314 have not been reported earlier as either active site or allosteric. Our method is thus, capable of predicting novel putative binding sites which, in principle, should be functionally significant owing to their strong conservation patterns.

4.4. Conclusion

We have developed discrete machine learning models using the random forest scheme to predict allosteric and active site residues. Our prediction models for allostery and active site detection use a common subset of features, which broadly include amino acid physicochemical properties, protein structure geometry, residue conservation and intrinsic dynamics of the protein structure. Instead of making predictions from a single model, we have used an ensemble approach to make predictions. In such an approach, we make multiple models for each prediction class, each model trained and validated on a separate training-validation set and make predictions using each model. Residue-level scores assigned by each model are weighted by the model's MCC and from this we calculate a weighted-ensemble score for each residue that relates to its probability of being an allosteric or active site residue. When compared to existing methods, our implementation makes predictions at a residue level by assigning them weighted probability scores. Such an implementation is useful, especially in the field of protein engineering by providing candidate residues whose mutations could possibly alter a protein's activity.

When assessed on the test dataset, our models for active site detection show comparable performance against two existing methods and gains against two other. Our models for allostery however, show superior performance over three of the existing methods. It is worth noting that including information on residue dynamics in addition to other properties appears to be the origin of the significant gain in performance. It is concerning, that our test datasets for allostery and active site prediction have only a small number of proteins, 15 and 19 respectively. In this context, we present two arguments which favor our selection criteria. First, since our models make predictions at the residue-level, having a larger set of residues in the test dataset is a more important consideration than the number of proteins. A number of existing methods identify pockets and then, rank them based on their propensity of being active or allosteric binding pockets (Akbar & Helms, 2018; Dundas et al., 2006; Greener & Sternberg, 2015; Hendlich, Rippmann, & Barnickel, 1997; Panjkovich & Daura, 2012; Singh et al., 2011). As proteins have fewer pockets than residues, these methods test on datasets having a diverse number of proteins. On the contrary, our models consider the total number of residues in the allosteric and active site test data sets: 167 allosteric and 6607 non-allosteric, 180 active site and 4344 non-active site residues. Second, since our aim is to develop separate models for the predictions of active and allosteric site residues, our required dataset needs to have labels for both allosteric and active site residues; however, we are limited by the availability of such previously compiled datasets.

Our study emphasizes that there can be considerable overlap between the feature profiles of active and allosteric site residues and hence, our models predict certain allosteric residues as active site residues and vice-versa. Residues that are terminal in an allosteric pathway often lie in close spatial proximity to the active site. Hence, their physicochemical, structural and dynamic properties can closely resemble those of the active site residues. Besides, previous studies have also suggested active sites may be allosterically coupled with one another. Based on these observations, a rigid classification of residues into allosteric and functional classes would, in some cases, be inappropriate. This also brings up another important aspect of protein structures and their dynamics – cooperativity. In its classical context, cooperativity is defined for proteins with multiple binding sites as an increase in substrate binding affinity of one site based on a binding event at another site. In the context of this study, we assume that residues sharing similar feature profiles, in principle, may be highly cooperative in nature. In contrast to its classical sense, however, cooperativity may not just be limited to increasing binding affinities, but more generally related to the energetic coupling in correlated dynamics among residues, key to maintaining a protein's functional dynamics.

4.5. Acknowledgements

This research was supported by NIH grant R01-GM72014 and NSF grant DBI-1661391, as well as funds from the Carver Trust awarded to the Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology. The funders had no role in study design, data collection or analysis, decision to publish, or preparation of the manuscript. The authors would like to thank Dr. Mary Ondrechen's group at Northeastern University and her graduate student Lydia for their help in running predictions with POOL.

4.6. References

Akbar, R., & Helms, V. (2018). ALLO: A tool to discriminate and prioritize allosteric pockets. *Chemical Biology and Drug Design*.

Alberts, B., Bray, D., Hopkins, K., Johnson, A., Lewis, J., Raff, M., … And Walter, P. (2009). *Essential Cell Biology*. *Archives of Biochemistry and Biophysics* (Vol. 231).

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, *80*(1), 505–515.

Bahar, I., Atilgan, A. R., & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design*, *2*(3), 173–181.

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America*, *22*(6), 725–730.

Betancourt, M. R., & Thirumalai, D. (1999). Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science : A Publication of the Protein Society*, *8*(2), 361–369.

Brylinski, M., Prymula, K., Jurkowski, W., Kochańczyk, M., Stawowczyk, E., Konieczny, L., & Roterman, I. (2007). Prediction of functional sites based on the fuzzy oil drop model. *PLoS Computational Biology, 3*(5), 0909–0923.

Campbell, E., Kaltenbach, M., Correy, G. J., Carr, P. D., Porebski, B. T., Livingstone, E. K., … Jackson, C. J. (2016). The role of protein dynamics in the evolution of new enzyme function. *Nature Chemical Biology*, *12*(11), 944–950.

Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M., & Funkhouser, T. A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Computational Biology*, *5*(12).

Cross, P. J., Dobson, R. C. J., Patchett, M. L., & Parker, E. J. (2011). Tyrosine latching of a regulatory gate affords allosteric control of aromatic amino acid biosynthesis. *Journal of Biological Chemistry*, *286*(12), 10216–10224.

Danon, L., Díaz-Guilera, A., & Arenas, A. (2006). The effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, (11).

De Vos, D., Xu, Y., Aerts, T., Van Petegem, F., & Van Beeumen, J. J. (2008). Crystal structure of Sulfolobus acidocaldarius aspartate carbamoyltransferase in complex with its allosteric activator CTP. *Biochemical and Biophysical Research Communications*, *372*(1), 40–44.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., & Liang, J. (2006). CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.).

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35.

Gerek, Z. N., Kumar, S., & Ozkan, S. B. (2013). Structural dynamics flexibility informs function and evolution at a proteome scale. *Evolutionary Applications*, *6*(3), 423–433.

Glantz-Gashai, Y., Meirson, T., & Samson, A. O. (2016). Normal Modes Expose Active Sites in Enzymes. *PLoS Computational Biology*, *12*(12), 1–17.

Goncearenco, A., Mitternacht, S., Yong, T., Eisenhaber, B., Eisenhaber, F., & Berezovsky, I. N. (2013). SPACER: Server for predicting allosteric communication and effects of regulation. *Nucleic Acids Research*, *41*(Web Server issue), 266–272.

Greener, J. G., & Sternberg, M. J. (2015). AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinformatics*, *16*(1), 335.

Hendlich, M., Rippmann, F., & Barnickel, G. (1997). LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, *15*(6), 359–363.

Huang, W., Lu, S., Huang, Z., Liu, X., Mou, L., Luo, Y., … Zhang, J. (2013). Allosite: A method for predicting allosteric sites. *Bioinformatics*, *29*(18), 2357–2359.

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680–682.

Huang, Z., Zhu, L., Cao, Y., Wu, G., Liu, X., Chen, Y., … Zhang, J. (2011). ASD: A comprehensive database of allosteric proteins and modulators. *Nucleic Acids Research*, *39*(SUPPL. 1).

Hubbard, S., & Thornton, J. (1993). NACCESS. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*.

Izidoro, S. C., De Melo-Minardi, R. C., & Pappa, G. L. (2015). GASS: Identifying enzyme active sites with genetic algorithms. *Bioinformatics*, *31*(6), 864–870.

Kisselev, A. F., Akopian, T. N., Castillo, V., & Goldberg, A. L. (1999). Proteasome active sites allosterically regulate each other, suggesting a cyclical bite-chew mechanism for protein breakdown. *Molecular Cell*, *4*(3), 395–402.

Kobe, B., & Kemp, B. E. (1999). Active site-directed protein regulation, 373–376.

Kumar, A., Glembo, T. J., & Ozkan, S. B. (2015). The Role of Conformational Dynamics and Allostery in the Disease Development of Human Ferritin. *Biophysical Journal*, *109*(6), 1273–1281.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105–132.

Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, *10*.

Li, X., Chen, Y., Lu, S., Huang, Z., Liu, X., Wang, Q., … Zhang, J. (2013). Toward an understanding of the sequence and structural basis of allosteric proteins. *Journal of Molecular Graphics & Modelling*, *40*, 30–39.

Lipscomb, W. N., & Kantrowitz, E. R. (2012). Structure and Mechanisms of Escherichia coli Aspartate Transcarbamoylase. *Accounts of Chemical Research*, *45*(3), 444–453.

Lockless, S. W., & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, *286*(5438), 295–299.

McClendon, C. L., Kornev, A. P., Gilson, M. K., & Taylor, S. S. (2014). Dynamic architecture of a protein kinase. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(43), E4623-31.

Motlagh, H. N., Wrabl, J. O., Li, J., & Hilser, V. J. (2014). The ensemble nature of allostery. *Nature*, *508*(7496), 331–339.

Murga, L. F., Wei, Y., Andre, P., Clifton, J. G., Ringe, D., & Ondrechen, M. J. (2004). Physicochemical Methods for Prediction of Functional Information for Proteins. *Israel Journal of Chemistry*, *44*, 299–308.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *70*(5), 9.

Newman, M. E. J. (2007). The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, *2*, 1–12.

Nussinov, R., & Tsai, C. J. (2014). Unraveling structural mechanisms of allosteric drug action. *Trends in Pharmacological Sciences*, *35*(5), 256–264.

Ondrechen, M. J., Clifton, J. G., & Ringe, D. (2001). THEMATICS: A simple computational predictor of enzyme function from structure. *Proceedings of the National Academy of Sciences*, *98*(22), 12473–12478.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems*, *54*(1999–66), 1–17.

Panjkovich, A., & Daura, X. (2012). Exploiting protein flexibility to predict the location of allosteric sites. *BMC Bioinformatics*, *13*(1), 273.

Petrova, N. V, & Wu, C. H. (2006). Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, *7*, 312.

Pravda, L., Berka, K., Svobodová Vařeková, R., Sehnal, D., Banáš, P., Laskowski, R. A., … Otyepka, M. (2014). Anatomy of enzyme channels. *BMC Bioinformatics*, *15*(1).

Pupko, T., Bell, R. E., Mayrose, I., & Glaser, F. (2002). Rate4Site-an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, *18*(1), 71–77.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.

Sievers, F., & Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology*, *1079*, 105–116.

Singh, T., Biswas, D., & Jayaram, B. (2011). AADS - An automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. *Journal of Chemical Information and Modeling*, *51*(10), 2515–2527.

Somarowthu, S., & Ondrechen, M. J. (2012). POOL server: Machine learning application for functional site prediction in proteins. *Bioinformatics*, *28*(15), 2078–2079.

Song, J., Li, F., Takemoto, K., Haffari, G., Akutsu, T., Chou, K.-C., & Webb, G. I. (2018). PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. *Journal of Theoretical Biology*, *443*, 125–137.

Song, K., Liu, X., Huang, W., Lu, S., Shen, Q., Zhang, L., & Zhang, J. (2017). Improved Method for the Identification and Validation of Allosteric Sites. *Journal of Chemical Information and Modeling*, *57*(9), 2358–2363.

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, *77*(9), 1905–1908.

Tong, W., Wei, Y., Murga, L. F., Ondrechen, M. J., & Williams, R. J. (2009). Partial Order Optimum Likelihood (POOL): Maximum likelihood prediction of protein active site residues using 3D structure and sequence properties. *PLoS Computational Biology*, *5*(1).

Tsai, C.-J., Del Sol, A., & Nussinov, R. (2009). Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Molecular bioSystems*, *5*(3), 207–216.

Xie, Z.-R., & Hwang, M.-J. (2015). Methods for Predicting Protein–Ligand Binding Sites. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 1215, pp. 383–398).

Yang, J., Cron, P., Thompson, V., Good, V. M., Hess, D., Hemmings, B. A., & Barford, D. (2002). Molecular mechanism for the regulation of protein kinase B/Akt by hydrophobic motif phosphorylation. *Molecular Cell*, *9*(6), 1227–1240.

Yang, L., Song, G., & Jernigan, R. L. (2009). Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(30), 12347–12352.

CHAPTER 5. COMPARISONS AMONG PROTEIN DYNAMICS FROM EXPERIMENTAL STRUCTURE ENSEMBLES, MOLECULAR DYNAMICS ENSEMBLES AND COARSE-GRAINED ELASTIC NETWORK MODELS

Paper published in J. Phys. Chem. B, 2018 (acs.jpcb.7b11668)

Kannan Sankar[1,2]*, Sambit K. Mishra*[1,2] *and Robert L. Jernigan*[1,2]

[1]Bioinformatics and Computational Biology Interdepartmental Graduate Program, Iowa State University, Ames, IA 50011-1178, USA

[2]Roy J. Carver Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA 50011-1178, USA

Abstract

Predicting protein motions is important for bridging the gap between protein structure and function. With growing numbers of structures of the same, or closely related proteins becoming available, it is now possible to understand more about the intrinsic dynamics of a protein with principal component analysis (PCA) of the motions apparent within ensembles of experimental structures. In this paper, we compare the motions extracted from experimental ensembles of 50 different proteins with the modes of motion predicted by several types of coarse-grained elastic network models (ENMs) which additionally take into account more details of either the protein geometry or the amino acid specificity. We further compare the structural variations in the experimental ensembles with the motions sampled in molecular dynamics (MD) simulations for a smaller subset of 17 proteins with available trajectories. We find that the correlations between the motions extracted from MD

trajectories and experimental structure ensembles are slightly different than for the ENMs, possibly reflecting potential sampling biases. We find that there are small gains in the predictive power of the ENMs in reproducing motions present in either experimental or MD ensembles by accounting for the protein geometry rather than the amino acid specificity of the interactions.

5.1. Introduction

Predicting conformational changes in proteins has long been a topic of interest to many who aim to understand protein function and mechanism. Multiple structures of the same protein, or closely related proteins, have been solved by different experimental methods - X-ray crystallography (Kohn, Afonine, Ruscio, Adams, & Head-Gordon, 2010), NMR spectroscopy (Fenwick, van den Bedem, Fraser, & Wright, 2014) and more recently by cryo-electron microscopy (Fernandez-Leiro & Scheres, 2016) under different conditions, in the presence of different ligands, or of mutated protein. These techniques reveal information about the intrinsic protein dynamics. The set of essential motions accessible to a protein can be readily obtained by applying principal component analysis (PCA) (Pearson, 1901) to the position coordinates of the aligned set of multiple experimental structures (Amadei, Linssen, & Berendsen, 1993; Amadei, Linssen, de Groot, van Aalten, & Berendsen, 1996; Howe, 2001; Teodoro, Phillips, & Kavraki, 2002; Teodoro, Phillips Jr., & Kavraki, 2003; van Aalten et al., 1997).

Information on protein motions can also be obtained from computational simulations such as molecular dynamics (MD) or Monte Carlo (MC). However, these applications require significant computer resources, and do not always fully sample the entire conformational space accessible to a protein. Coarse-grained elastic network models (ENMs)

(Chennubhotla, Rader, Yang, & Bahar, 2005; Jernigan, Yang, Song, Kurkckuoglu, & Doruker, 2009; Sanejouand, 2011) on the other hand, offer a faster and cheaper alternative to MD or MC simulations for sampling the intrinsic motions accessible to a protein. By modeling the protein as a string of beads (usually the C^{α} atoms) connected by harmonic springs (interactions), they are often able to capture the most important global motions. ENMs have been used extensively to study the intrinsic dynamics of a variety of biomolecules ranging from small globular and membrane proteins (Bahar, Lezon, Bakan, & Shrivastava, 2010) to nucleic acids (Setny & Zacharias, 2013) and even large biomolecular assemblies such as the ribosome (Burton, Zimmermann, Jernigan, & Wang, 2012; Wang & Jernigan, 2005; Wang, Rader, Bahar, & Jernigan, 2004) and GroEL (Keskin, Bahar, Flatow, Covell, & Jernigan, 2002; Z. Yang, Májek, & Bahar, 2009) They have been shown to accurately predict the crystallographic B-factors of diverse proteins (Soheilifard, Makarov, & Rodin, 2008; L. Yang, Song, & Jernigan, 2009a) as well as to capture conformational changes between pairs of structures of the same protein (Tama & Sanejouand, 2001; L. Yang, Song, & Jernigan, 2007). The normal modes from ENMs have also been shown to capture structural variations extracted from multiple experimental structures of the same protein (Skjaerven, Martinez, & Reuter, 2011; L.-W. Yang, Eyal, Bahar, & Kitao, 2009; L. Yang, Song, Carriquiry, & Jernigan, 2008) or RNA (Zimmermann & Jernigan, 2014).

Specifically, here we focus on ENMs that provide the changes in the geometry, the Anisotropic Network Models (ANM) (Atilgan et al., 2001). A subject of some importance has been how to improve ENMs by accounting for either more specific details of protein geometry or the chemical nature of amino acids (Frappier & Najmanovich, 2014; Kim et al., 2013). Hamacher and McCammon have shown that an extended ANM (**eANM**) (Hamacher

& McCammon, 2006) with spring constants based on the values of the Miyazawa-Jernigan (MJ) potential amino acid interaction energies(Miyazawa & Jernigan, 1996) to account for the amino acid specificity of fluctuations performs better in reproducing crystallographic B-factors. We have also shown that the ANM can be significantly improved by weighting the spring constants between residues by the inverse powers of the distance of separation between them (L. Yang, Song, & Jernigan, 2009b), a model referred to as the parameter-free ANM (**pfANM**) (pf means that there is no cutoff parameter as in the traditional ANM). Other ways of adjusting the springs in ENMs are to use information from the variance-covariance matrix of position coordinates (Moritsugu & Smith, 2007) or the mean square distance fluctuations (Lyman, Pfaendtner, & Voth, 2008) between residues from MD trajectory ensembles of the protein. We and others have also shown that using spring constants based on the variance of internal distance changes between residues also provides significant gains in the ability to reproduce experimentally observed conformational changes (Katebi, Sankar, Jia, & Jernigan, 2015; Skjærven, Yao, Scarabelli, & Grant, 2014).

In this work, we also introduce a modified version of Hamacher and McCammon’s extended ANM (called **ccANM**) in which the spring constants between residues are based on the relative entropies of amino acid pairs rather than the relative energies of the pairs. This is based on our recent work, where we extracted a scale of relative entropies between amino acid pairs (Sankar, Jia, & Jernigan, 2017) based on the frequencies of contact changes between amino acid types during conformational changes within a dataset of proteins. This entropy measure yields significant gains in identifying native structures among decoy sets. Since these entropies measure the tendency for amino acid contacts to change, we hypothesize that information on relative entropies of the amino acid pairs might be more useful than their relative energies for differentiating among springs representing the interactions.

First, we systematically test the effectiveness of the classical coarse-grained ANM and four different variants of the ANM (that incorporate additional information either regarding protein geometry or amino acid specificity) in capturing the motions present in experimental structure ensembles of 50 different proteins. In addition, for a smaller subset of 17 proteins where MD trajectories are available, we also compare the motions present in the experimental ensembles to those in the MD ensembles. Our results suggest that the protein motions as extracted from experimental ensembles can differ significantly from those obtained through MD simulations. Whether this reflects the difference between the crystal environments and the simulation conditions, or a failure of simulations to fully capture the characteristic dynamics remains an open question. In addition, we also investigate how well the motions present in either the experimental or MD ensembles are captured by a variety of simple coarse-grained elastic network models.

5.2. Methods

5.2.1. Experimental Structure Ensemble Data

A set of experimental structure ensembles for 50 different proteins (Table D.1) were collected in our previous work (Sankar, Liu, Wang, & Jernigan, 2015), which we are utilizing here. We refer the reader to this previous work for the list of structures in each ensemble set. These structures were obtained by a clustering of the Protein Data Bank (PDB) (Berman et al., n.d.) at 95% sequence identity level. Only clusters corresponding to monomeric proteins were retained. The structures in each cluster were aligned using the multiple structure alignment program MUSTANG (Konagurthu, Whisstock, Stuckey, &

Lesk, 2006), and the corresponding structure-based sequence alignment was used as a guide to remove any residues and/or structures that introduced significant gaps in the middle of the alignment (relatively few such cases). The final set of aligned structures from our previous work has been used for the experimental protein ensembles. For construction of ANMs, the structure with the lowest average root mean square deviation (RMSD) from all other structures is chosen as the representative structure for each ensemble (see Table D.1 for the list of these representative structures). The distributions of the average RMSDs in each ensemble can be found in our previous work (Sankar et al., 2015).

5.2.2. Molecular Dynamics Trajectories

For each experimental protein set, we have searched for homologous entries in the MoDEL database (Meyer et al., 2010) a repository of publicly available MD trajectories. Since the set of proteins in each cluster have a high sequence identity (≥ 95%), we choose a protein randomly from each cluster and search for its homologs. We set a threshold on the sequence identity of 35% for this selection. For clusters with multiple available homologs, we only choose the one with the highest sequence identity. We then download the C^{α} atom trajectories for the selected homologs for each cluster from the MoDEL database. A list of the proteins whose trajectories were used is given in Table D.2.

In order to obtain a common reference frame, we transform the coordinates of the MD trajectories from their native frame to the frame of their experimental homologs. We do this by superimposing the first frame from each MD trajectory onto the representative structure from the corresponding experimental ensemble set; and then superimposing all the other frames onto the first frame. In order to identify a common subset of residues between the experimental and MD datasets, we then align the sequence of each MD homolog to the profile alignment of its respective experimental set (with *ClustalOmega)* (Sievers & Higgins, 2014) and retain only the subset of residues from the PDB structure in common with the MD homolog and the experimental ensemble. For generating ANMs, the starting PDB structure of each MD dataset is used as the representative of the ensemble (see Table D.2).

5.2.3. Principal Component Analysis of Structural Ensembles

Information about protein dynamics is extracted from either the experimental ensemble or the MD trajectory ensemble by using PCA of the aligned set of structures (to remove rigid body motions). In each case, the dataset for PCA is a matrix $X_{n\times 3N}$ consisting of the X-, Y- and Z-coordinates of the C^{α} atoms of each of N residues in the aligned set of n structures. The variance-covariance matrix $\boldsymbol{C}_{3N\times 3N}$ of the position coordinates is constructed with its elements obtained as

$$C_{ij} = \langle X_{ij} - \langle X_i \rangle \rangle \langle X_{ij} - \langle X_j \rangle \rangle\,; \tag{5.1}$$

where the brackets refer to averages across all n structures. Eigen-decomposition of the matrix $\boldsymbol{C}$ results in the eigenvectors, which are a set of orthogonal directions of the variations present in the dataset having corresponding eigenvalues denoting the variance along the corresponding directions. The principal component (PC) scores are obtained directly from the projections of the mean centered data points along these eigenvectors. The PCs are sorted in decreasing order of the corresponding eigenvalues and referred to as PC1, PC2, PC3 and so on, with PC1 capturing the most significant part of the structural variations.

5.2.4. Coarse-grained Elastic Network Models

Next, we describe the various coarse-grained bead-spring models that we have used in our comparisons. Collectively these are all termed elastic network models (ENMs).

Anisotropic Network Model (ANM). The ANM (Atilgan et al., 2001) is an elastic-network (bead-spring) model in which the C$^{\alpha}$ atoms of each residue in the protein are represented as beads and all interactions between residues are modeled as harmonic springs. Interactions between beads are usually restricted to physically close residues within a fixed distance cutoff R_c. There are two parameters in ANM: the distance cutoff R_c and the spring constant γ_{ij} between every pair of residues i and j. Throughout this study, the value of R_c has been set to 13 Å. In a classical ANM, all springs are assigned uniform values. In other words, for a protein with N residues,

$$\gamma_{ij} = \gamma \; \forall \; i, j \; \epsilon \; \{1, \dots, N\} \tag{5.2}$$

All the springs are assumed to be in equilibrium in the starting structure and the potential energy V of the system is computed as

$$V = \frac{1}{2} \sum_{i,j=1}^{N} \gamma_{ij} \left(R_{ij} - R_{ij}^{0} \right)^2, \tag{5.3}$$

where R_{ij} refers to the instantaneous displacement between atoms i and j and R_{ij}^{0} refers to their equilibrium displacement. The Hessian matrix $\boldsymbol{H}$ of the system, with $N \times N$ superelements H_{ij} is calculated as the matrix of second derivatives of the potential with respect to the Cartesian coordinate positions of the residues as

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 V}{\partial X_i \partial X_j} & \frac{\partial^2 V}{\partial X_i \partial Y_j} & \frac{\partial^2 V}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Y_i \partial X_j} & \frac{\partial^2 V}{\partial Y_i \partial Y_j} & \frac{\partial^2 V}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Z_i \partial X_j} & \frac{\partial^2 V}{\partial Z_i \partial Y_j} & \frac{\partial^2 V}{\partial Z_i \partial Z_j} \end{bmatrix} \tag{5.4}$$

The normal modes of motion from ANM are obtained as eigenvectors of the matrix $\boldsymbol{H}$; with the corresponding eigenvalues representing the square of frequencies of the modes. The correlations in motion between the residues along the X, Y and Z directions can be obtained from the corresponding super-elements of $\boldsymbol{H}^{-1}$ and the mean square fluctuations of each residue i from the diagonal elements of the corresponding super element H_{ii}^{-1} as follows:

$$\langle \Delta R_i^2 \rangle = \frac{k_B T}{\gamma} trace(H_{ii}^{-1}) \tag{5.5}$$

The theoretical B-factors from the ANM can be conveniently calculated from the mean square fluctuations as

$$B_i^{calc} = 8\pi^2 \langle \Delta R_i^2 \rangle / 3 \tag{5.6}$$

In addition to the classical ANM, we also explore some different variants of the ANM. The basic idea of each of the modified ANMs is the same, with the only change being that the spring constants are modified somehow.

Parameter-free ANM (pfANM). In the pfANM (L. Yang et al., 2009b), one of the parameters, the R_c is eliminated by allowing all residues to be connected, but instead of uniform springs the spring constants are taken to be proportional to a given inverse power p of the distance r_{ij} between them as in Eq. 5.7. Previously we found that $p = 6$ gave the best representation of the collective motions; whereas $p = 2$ best fit the experimental B-factors (L. Yang et al., 2009b).

$$\gamma_{ij} = \frac{1}{r_{ij}^p} \ \forall \ i, j \ \epsilon \ \{1, \dots, N\} \tag{5.7}$$

Extended ANM (eANM). We use a simplified version of a modified ANM introduced by Hamacher and McCammon (Hamacher & McCammon, 2006) in which the spring constants between a pair of non-adjacent contacting residues (as identified by R_c) is weighted by the absolute value of the Miyazawa-Jernigan (MJ) potential (Miyazawa & Jernigan, 1996) energy $|\kappa_{ij}|$ between them. The spring stiffness between adjacent residues is set to a much larger value, K = 82 RT/Å^2 in accordance with the values found for peptide bonds. That is,

$$\gamma_{ij} = \begin{cases} K & \text{if } |i-j| = 1 \\ 2|\kappa_{ij}| & \text{if } |i-j| \neq 1 \text{ and } r_{ij} \leq R_c \end{cases} \forall\, i,j \,\epsilon\, \{1, \dots, N\} \tag{5.8}$$

***C**ontact-change based ANM (ccANM).* This is a model similar to the eANM; except that the springs between non-adjacent contacting residues falling within the cutoff distance R_c are weighted by the inverse of the contact-change based entropies (Sankar et al., 2017) s_{ij} between the amino acid pair. That is,

$$\gamma_{ij} = \begin{cases} K & if \;\; |i-j| = 1 \\ \frac{1}{s_{ij}} & if \;\; |i-j| \neq 1 \text{ and } r_{ij} \leq R_c \end{cases} \forall\; i,j \,\epsilon\, \{1, \dots, N\} \tag{5.9}$$

Distance change based ANM (dcANM). This model captures internal distance-changes as observed within an ensemble of structures. For this variant of the ANM, the spring constants between each pair of residues is taken as the inverse of the variance of internal distances ($\sigma^2_{r_{ij}}$) between the residue pair over the set of structures (these spring constant values were further normalized such that they range between 0 and 1). In other words,

$$\gamma_{ij} = \frac{1}{\sigma^2_{r_{ij}}} \;\; \forall\; i,j \,\epsilon\, \{1, \dots, N\} \tag{5.10}$$

5.2.5. Performance Evaluation of the ENMs

We measure the performance of each ENM in terms of how well it can reproduce the protein structural variations present within an ensemble. The directions of motions from the ENM are obtained directly from the ENM modes and the structural variations present in an ensemble (experimental/MD) are obtained with PCA. Similarity comparisons between a PC and a mode are evaluated by three measures defined by Tama and Sanejouand (Tama & Sanejouand, 2001).

Overlap (O). This is a measure of how similar the direction of a given mode of motion M_j from an ENM is in comparison with the PC eigenvector P_i and is calculated as

$$O_{ij} = \frac{|P_i \cdot M_j|}{\|P_i\|\|M_j\|} \tag{5.11}$$

where $|P_i . M_j|$ refers to the absolute value of the dot product of P_i and M_j and $\|P_i\|$ and $\|M_j\|$ refer to the length of the PC and mode vectors, respectively. The sign of the dot product is not considered since the modes are harmonic in nature. The maximum overlap between any of the first k modes of motion with the PC eigenvector P_i is obtained as

$$O_i^{max} = \max_{j=1\ to\ k} O_{ij} \tag{5.12}$$

Cumulative Overlap (CO). This is a measure of how well a set of the first k modes from an ENM capture the motion sampled by a single PC eigenvector P_i and is calculated as

$$CO_i^k = \sqrt{\sum_{j=1}^{k} O_{ij}^2} \tag{5.13}$$

Root Mean Square Inner Product (RMSIP). This quantity measures the similarity in directions between the set of first k modes from an ENM and the first l PC eigenvectors from a structural ensemble as

$$RMSIP_l^k = \sqrt{\frac{1}{l} \sum_{i=1}^{l} \sum_{j=1}^{k} \left(P_i \cdot M_j\right)} \tag{5.14}$$

Based on the above three measures, we use ten different performance metrics to evaluate the performance of elastic network models in comparison to PCs from an ensemble as follows: the maximum overlap between the first 20 modes from the ENM and each of PC1 (O_1^{max}), PC2 (O_2^{max}) and PC3 (O_3^{max}); the cumulative overlap between the first 20 modes from the ENM and PC1 (CO_1^{20}), PC2 (CO_2^{20}) and PC3 (CO_3^{20}); and the RMSIP between the first 20 ANM modes and sets of the first 3 ($RMSIP_3^{20}$), 6 ($RMSIP_6^{20}$), 10 ($RMSIP_{10}^{20}$) and 20 PCs ($RMSIP_{20}^{20}$).

In addition, Pearson's correlation coefficient is reported between the calculated B-factors (B^{calc}) from the ENM and the crystallographic temperature factors (B^{exp}) from the representative structure in the experimental ensemble as

$$\rho^{exp,calc} = \frac{\boldsymbol{B}^{exp}-\langle\boldsymbol{B}^{exp}\rangle}{\|\boldsymbol{B}^{exp}-\langle\boldsymbol{B}^{exp}\rangle\|} \frac{\boldsymbol{B}^{calc}-\langle\boldsymbol{B}^{calc}\rangle}{\|\boldsymbol{B}^{calc}-\langle\boldsymbol{B}^{calc}\rangle\|} \tag{5.15}$$

5.3. Results and Discussion

5.3.1. Comparison of ENM Modes with the Motions Present within Experimental Structure Ensembles

We have previously shown for HIV-1 protease that the modes of motion from the classical ANM of a single structure correspond closely to the motions extracted from a set of experimental structures (L. Yang et al., 2008). Several other studies have also demonstrated the power of ANMs in capturing the structural variations within experimental ensembles for a variety of proteins (Skjaerven et al., 2011; L.-W. Yang et al., 2009). Here, we compare the motions predicted by the classical ANM and four other variants of ENMs with the motions present in experimental structure ensembles for a much larger dataset of 50 different proteins (Sankar et al., 2015) (see Table D.1).

In addition to the classical ANM, we use the four other types of modified ENMs (refer to Methods above for more details): (1) pfANM (L. Yang et al., 2009b) with the spring constants between every residue pair weighted by the inverse of the sixth power of the distance between them; (2) eANM (Hamacher & McCammon, 2006) where the spring constants are weighted by the absolute values of the MJ potential energies between amino acid pair; (3) ccANM, in which the spring constants are weighted by the inverse of the contact-change based entropy value for each amino acid pair (based on our previous work); (Sankar et al., 2017) and (4) dcANM (Katebi et al., 2015) with the spring constant between every pair of residues weighted by the inverse of the variance of the internal distances between them (over all the structures in the experimental ensemble). The performance of each ANM is evaluated for the ten different metrics described in Methods.

We compute the motions for the ENMs of the representative structure from each protein ensemble (identified as the structure having the lowest RMSD from all other structures). Table 5.1 shows the average values (over the 50 proteins) of the 10 metrics for each type of ENM investigated. As expected, the dcANM naturally outperforms all of the other kinds of ANM in almost all the metrics. This is because the springs of the dcANM have been chosen directly from the internal distance changes between every pair of residues within the dataset for each protein; and hence it is naturally able to better reproduce the structural variations present in the dataset since it is built directly on the data being compared. The performance assessment of the other ENMs against one another is more relevant to understanding the behavior of the ENMs. Based on the number of metrics for which the ENM is best, the ranking of the models is as follows: pfANM > ccANM > ANM > eANM. It is clear from Table 5.1 that the pfANM outperforms the other types of ENMs. Also, the ccANM performs essentially at the same level as the ANM on all 10 metrics.

Table 5.1. Performance of different types of ENMs for the dataset of 50 proteins in comparison with the motions present in the experimental ensembles.

Model	$\boldsymbol{O_1^{max}}$	$\boldsymbol{O_2^{max}}$	$\boldsymbol{O_3^{max}}$	$\boldsymbol{CO_1^{20}}$	$\boldsymbol{CO_2^{20}}$	$\boldsymbol{CO_3^{20}}$	$\boldsymbol{RMSIP_3^{20}}$	$\boldsymbol{RMSIP_6^{20}}$	$\boldsymbol{RMSIP_{10}^{20}}$	$\boldsymbol{RMSIP_{20}^{20}}$
ANM	**0.41** ± 0.20	0.42 ± 0.18	0.44 ± 0.14	0.67 ± 0.21	0.70 ± 0.18	0.71 ± 0.13	0.70 ± 0.13	0.67 ± 0.09	0.63 ± 0.07	0.56 ± 0.06
pfANM	0.39 ± 0.19	**0.44** ± 0.17	**0.45** ± 0.13	**0.68** ± 0.21	**0.73** ± 0.17	**0.74** ± 0.12	**0.73** ± 0.12	**0.70** ± 0.09	**0.66** ± 0.07	**0.59** ± 0.06
eANM	0.39 ± 0.20	0.42 ± 0.18	0.44 ± 0.14	0.66 ± 0.22	0.70 ± 0.18	0.72 ± 0.13	0.70 ± 0.13	0.66 ± 0.10	0.63 ± 0.08	0.56 ± 0.06
ccANM	**0.41** ± 0.20	0.43 ± 0.18	0.44 ± 0.13	0.67 ± 0.22	0.71 ± 0.17	0.72 ± 0.12	0.71 ± 0.13	0.67 ± 0.10	0.64 ± 0.07	0.57 ± 0.06
dcANM*	*0.56* *± 0.19*	*0.49* *± 0.15*	*0.50* *± 0.13*	*0.83* *± 0.14*	*0.82* *± 0.12*	*0.82* *± 0.10*	*0.83* *± 0.08*	*0.78* *± 0.07*	*0.73* *± 0.06*	*0.64* *± 0.06*

Values for each metric (as defined in Methods) are averaged over the 50 proteins. Values for the best performing model for each metric are shown in bold.
*dcANM is trained using the variances of the internal distance changes between residues in each experimental ensemble, and results are shown in italics.

5.3.2. Comparison with Protein Motions from MD and Experimental Datasets

Often only one structure of a protein or its close homolog is available. In such cases, a conformational sampling of the protein is often obtained using various computational techniques such as MD or Monte Carlo simulations. Once the simulation is run, the set of resulting structures are aligned to the starting structure and the 'essential motions' (Amadei et al., 1993) extracted from the trajectory using PCA as described in the Methods section.

We perform a sequence-based search on the MODEL database (Meyer et al., 2010), an online repository of MD simulations for available MD trajectories of the proteins or their homologs present in the dataset of 50 proteins. We identify 17 proteins for which MD simulation data were available for the protein or a substantial part of it (Table D.2). We then compare how well the motions sampled by MD simulations for the set of 17 proteins compare against the variations present in sets of experimental structures of the same protein.

Table 5.2 shows this comparison of the PCs extracted from the experimental dataset vs MD dataset for the 17 proteins.

Table 5.2. Comparison of MD and experimental motions for the set of 17 proteins.

Metric	O_1^{max}	O_2^{max}	O_3^{max}	CO_1^{20}	CO_2^{20}	CO_3^{20}	$RMSIP_3^{20}$	$RMSIP_6^{20}$	$RMSIP_{10}^{20}$	$RMSIP_{20}^{20}$
Value	0.37 ± 0.18	0.37 ± 0.13	0.37 ± 0.10	0.63 ± 0.18	0.65 ± 0.17	0.70 ± 0.10	0.67 ± 0.12	0.65 ± 0.09	0.62 ± 0.07	0.56 ± 0.05
Values for each metric (as defined in Methods) are averaged over the 17 proteins.										

The average maximum overlap between the first twenty PC directions from the MD ensemble with the PC1, PC2 and PC3 of the experimental ensemble is 0.37; which is comparatively smaller than the average values obtained for the classical ANM or any of the variants of the ANM. This difference is small and thus probably not significant. Several factors can affect the set of structures sampled in the MD trajectory; including the force field used, the simulation time, etc. It is also possible that the overlap between the conformational space sampled by MD and experiments is relatively small. As a result, a dcANM trained on the MD dataset could not reproduce well the set of motions in the experimental (MD) ensemble (Table D.3). The fact that the ENMs reproduce the experimental ensemble better is noteworthy.

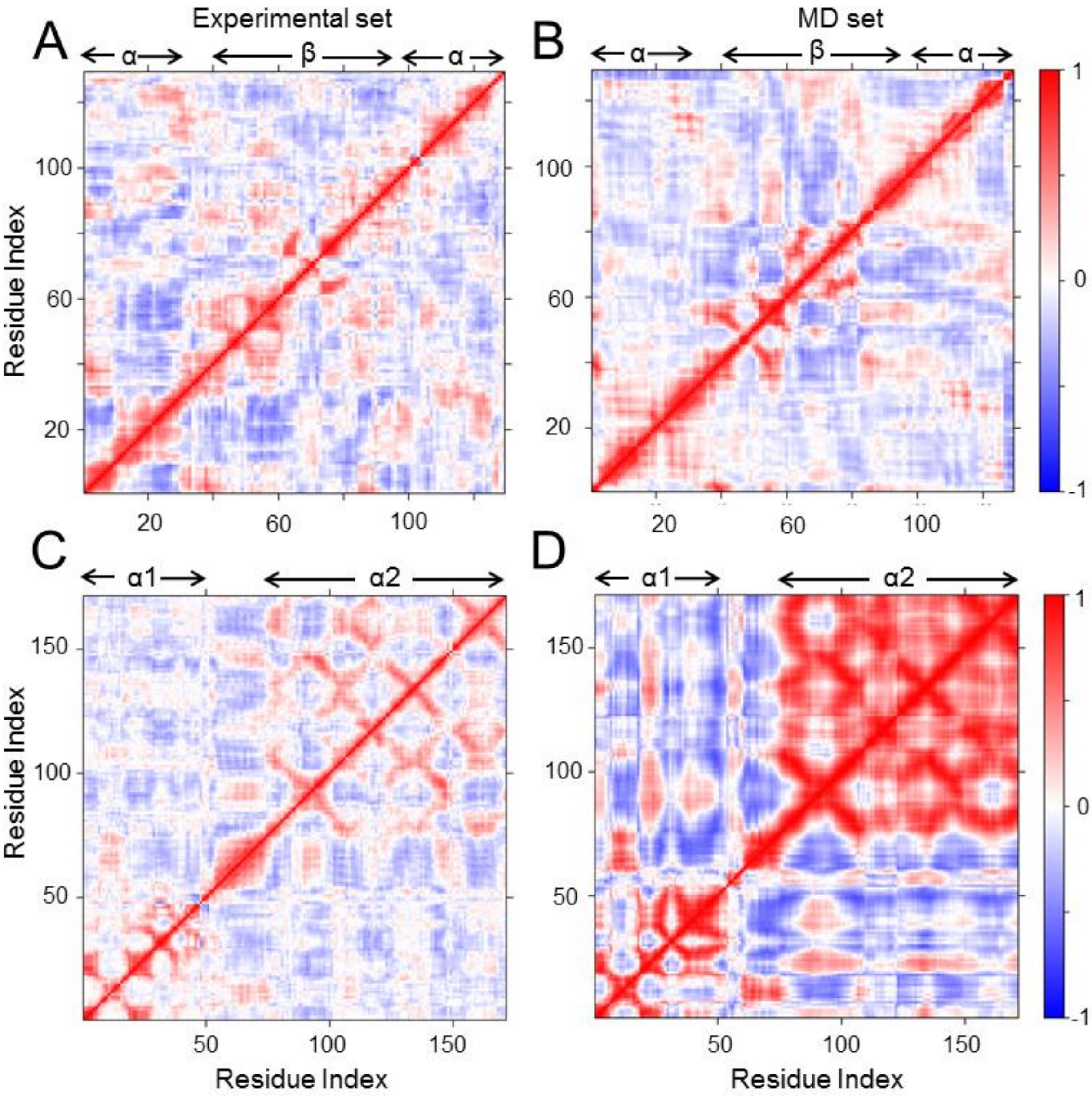

Figure 5.1. Comparison of dynamical cross-correlation matrices (DCCMs) between experimental and MD datasets for lysozyme C (A, B) and HLA-DRA (C, D). Positive correlations between residues are shown red and negative correlations in blue. The two

In order to further demonstrate that the motions sampled by MD and the experimental ensembles are often different, we provide two examples of dynamical cross-correlations (DCCMs) (Ichiye & Karplus, 1991) of the residues from experimental and MD datasets for two different proteins in the dataset, lysozyme C (Figure 5.1A and B) and human leukocyte antigen (HLA) class II histocompatibility antigen alpha chain (HLA-DRA) (Figure 5.1C and D). These were chosen to demonstrate outliers in terms of being most similar and most different. In the case of lysozyme C, the two DCCMs are similar but with intricate differences, whereas in the case of the HLA-DRA, there is major differences between the correlations shown.

A closer inspection of the plots for HLA-DRA reveals that in the MD dataset, there are stronger correlations among the residues within each of its two domains (α1 and α2), particularly for α2, suggesting that the domains move almost as if they were rigid bodies. On the other hand, within the experimental ensemble, the higher correlations mostly correspond to residues within the same secondary structure, which can be easily identified from the plots. In other words, higher variabilities are observed in the relative orientations of the secondary structures within each domain. Previous studies have also shown that the DCCMs of the same protein from distinct simulations over different time-scales in MD simulations can be different (Hünenberger, Mark, & van Gunsteren, 1995). Our results further support these observations in addition to suggesting that the dynamical cross correlations observed in MD often do not correspond to those observed in a set of experimental structures.

Since the length of each MD simulation is not the same, one probable reason for the low level of correspondences between motions from experiments and the simulations is the simulation time. In order to ascertain whether this is the case, we divide the dataset into two sets: short (< 80ns) and long (≥ 80 ns) simulations (see Table D.4). We then perform hypothesis testing to see whether the average values for each of the ten metrics for the short simulations are lesser than those for the long simulations. Our analysis suggests that the observed differences are not significant (the p-values for all metrics are > 0.4), at least for the current dataset (Table D.4). More detailed studies on larger datasets will need to be performed to establish the importance of simulation times or the force-field used in reproducing experimental motions.

5.3.3. Comparison of ENM Modes with Motions Present in MD Structural Ensembles

It is also interesting to test whether the motions predicted by ENMs correlate with the set of motions sampled by MD simulations of the same protein. Starting from the representative structure for each protein, we construct the different types of ENMs and investigate how well the modes compare with the motions extracted by PCA from the MD structural ensemble for each of the 17 proteins. Table 5.3 shows the average values for each of the ten performance metrics for the different types of ENMs.

Again, as expected the dcANM performs the best in all metrics reflecting the fact that it was trained on the dataset itself. The other different ENMs rank in the following order for the ten performance metrics: pfANM > ccANM > ANM > eANM. And, this is the same order as seen in Table 5.1. It can be seen that the pfANM systematically outperforms the other types of ANM in reproducing the protein motions in the MD dataset, even though by a small margin. Taken together with the results from the performance on the experimental dataset, this seems to suggest that the overall intrinsic dynamics of the protein is dictated primarily by its geometry, i.e., the distances of separation between all pairs of different residues. The specific amino acid interactions of course allow the protein perform its specific functions; and will account for the differences in behaviors of various mutants of the protein; however, they do not much affect its global motions.

Table 5.3. Performance of different types of ENMs in comparison with the motions present in the MD dataset of 17 proteins.

Model	O_1^{max}	O_2^{max}	O_3^{max}	CO_1^{20}	CO_2^{20}	CO_3^{20}	$RMSIP_3^{20}$	$RMSIP_6^{20}$	$RMSIP_{10}^{20}$	$RMSIP_{20}^{20}$
ANM	0.42 ± 0.14	0.46 ± 0.22	0.40 ± 0.12	0.68 ± 0.17	0.68 ± 0.20	0.68 ± 0.14	0.68 ± 0.15	0.70 ± 0.11	0.70 ± 0.07	0.65 ± 0.05
pfANM	**0.44** ± 0.14	**0.46** ± 0.21	**0.42** ± 0.09	**0.71** ± 0.15	**0.69** ± 0.20	**0.71** ± 0.15	**0.71** ± 0.15	**0.73** ± 0.10	**0.74** ± 0.06	**0.70** ± 0.04
eANM	0.41 ± 0.15	0.45 ± 0.22	0.39 ± 0.12	0.67 ± 0.18	0.67 ± 0.21	0.66 ± 0.15	0.67 ± 0.16	0.69 ± 0.12	0.69 ± 0.08	0.65 ± 0.06
ccANM	0.43 ± 0.14	0.44 ± 0.21	0.40 ± 0.11	0.70 ± 0.16	0.68 ± 0.20	0.68 ± 0.14	0.69 ± 0.15	0.71 ± 0.10	0.71 ± 0.07	0.66 ± 0.05
dcANM*	*0.51 ± 0.15*	*0.44 ± 0.13*	*0.44 ± 0.13*	*0.80 ± 0.16*	*0.77 ± 0.16*	*0.75 ± 0.16*	*0.78 ± 0.15*	*0.74 ± 0.10*	*0.71 ± 0.07*	*0.63 ± 0.05*

Values for each metric (as defined in Methods) are averaged over the 17 proteins. Values for the best performing model are shown in bold and the next best in italics.
*dcANM is trained using the variances of the internal distance changes between residues in each MD ensemble, and results are shown in italics.

A comparison between the results in Table 5.1 and Table 5.3 shows a remarkable similarity in the abilities of the various ENMs to reproduce the motions in the ensembles of both the experimental sets of structures and the MD ensembles.

5.3.4. Performance of ENMs in Reproducing Crystallographic B-factors

In addition to being able to reproduce intrinsic protein motions, another strength of the ENMs is in their being able to reproduce crystallographic temperature factors of the residues in the protein. Here we generate different types of ENMs using the representative structure for each of the 17 proteins with MD trajectory data and compute B-factors from the models (see Methods). The dcANM models are generated by adjusting the spring constants using the internal distance changes present in the experimental and MD ensembles as described before. We then compute the Pearson's correlations between the predicted B-factors and the crystallographic B-factors of the representative structure in the experimental ensemble (Table 5.4).

Table 5.4. Correlation between experimental temperature factors and predicted B-factors from various types of ANMs on the experimental and MD datasets.

Model	Correlation (MD dataset)*	Correlation (experimental dataset)#
ANM	0.50 ± 0.14	0.53 ± 0.14
pfANM	0.52 ± 0.17	**0.56** ± 0.14
eANM	0.51 ± 0.13	0.53 ± 0.12
ccANM	0.48 ± 0.14	0.50 ± 0.14
dcANM	**0.53** ± 0.20	0.51 ± 0.18

Values are averaged over the 17 proteins. Value for the best performing model is shown in bold.
*dcANM is trained using internal distance changes between residues in the MD dataset;
#dcANM is trained using internal distance changes between residues in the experimental dataset;
Correlation values are with the crystallographic B-factors of the experimental representative structure.

As can be seen in Table 5.4, the pfANM gives the highest correlation with crystallographic B-factors. The dcANM model based on the MD dataset gives only a slightly better correlation with B-factors than the pfANM and is probably not a significant difference. Our results also confirm the observation by Hamacher and McCammon (Hamacher & McCammon, 2006) that the eANM provides slight gains over the ANM in its being able to predict crystallographic B-factors (at least for the cases in the MD dataset). However, the values in Table 5.4 are all very similar. Interestingly, the eANM is slightly worse than the classical ANM or the ccANM at predicting motions present in the experimental ensembles as seen above (Tables 5.1 and 5.3). On the other hand, it is slightly better than the ccANM at reproducing crystallographic B-factors. This is in close agreement with observations by Fuglebakk and others (Fuglebakk, Reuter, & Hinsen, 2013) that a higher correlation with B-factors usually comes at the expense of the ability to predict collective protein motions.

5.4. Conclusions

In this study, we have systematically compared the motions extracted from experimental structure ensembles of 50 different proteins with the motions predicted using several different variants of ENMs. In addition to the classic ANM, we study several modified ANMs which account more specifically for the geometry of the protein (pfANM and dcANM) or for the amino acid specificity of the residues, either in energy (eANM) or in entropy (ccANM). The ccANM is a new model introduced in this paper, which accounts for the relative entropies of amino acid pairs; which were derived from the relative frequencies of contact changes within a set of experimental protein conformational changes. Our results show that pfANMs (taking into account all distances between residues in a protein structure) are best in capturing the structural variations present within an experimental ensemble of the same protein. The ccANMs do perform better than eANMs and the classic ANMs suggesting that the pair-wise entropies are important for conformational changes. The main conclusion is that the distances of separation between residues (i.e. the geometry in pfANM) plays a larger role than the chemical nature of the interactions (as in eANM or ccANM) for the overall intrinsic dynamics of proteins. Interestingly this is consistent with the strong dependence on geometry (shape) for the slowest motions (Doruker & Jernigan, 2003; Ma, 2004) supporting the overall viewpoint implicit in the elastic network models that geometry alone is important for the important protein dynamics.

In addition, we also have collected large scale molecular dynamics simulation data available for 17 proteins in the dataset and compared their structural changes with the structural variations present in the experimental set and those predicted by different types of ANM. The correspondences observed between the MD and experimental datasets is relatively poor when compared to the ANMs, highlighting some of the possible sampling problems in MD datasets, such as the force-field used, and simulation times. We also observe that training ANMs based on internal distance changes between residues observed in an MD simulation (dcANM) does not necessarily improve the correspondence with experimental motions, at least for the dataset of 17 proteins investigated in this study.

We find that some ANMs, specifically the pfANM or ccANM give better agreement with experimental motions extracted from experimental or MD ensembles. On the other hand, they provide only relatively small improvements in terms of the correlation with experimental B-factors, in agreement with previous studies. However, as observed by others (Fuglebakk et al., 2013), we also find that agreement with B-factors and the ability to reproduce collective motions do not necessarily go together.

5.5. Acknowledgement

This work was supported by National Institutes of Health Grant R01-GM72014 and National Science Foundation Grant MCB-1021785, as well as funds from the Carver Trust awarded to the Roy J. Carver Department of Biochemistry, Biophysics, and Molecular Biology. KS was also supported by fellowship funds from the Office of Biotechnology, Iowa State University. The authors thank all members of the Jernigan Lab and the Guang Song Lab at Iowa State University for frequent helpful discussions.

5.6. References

Amadei, A., Linssen, A. B., & Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins*, *17*(4), 412–425.

Amadei, A., Linssen, A. B., de Groot, B. L., van Aalten, D. M., & Berendsen, H. J. (1996). An efficient method for sampling the essential subspace of proteins. *Journal of Biomolecular Structure & Dynamics*, *13*(4), 615–625.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, *80*(1), 505–515.

Bahar, I., Lezon, T. R., Bakan, A., & Shrivastava, I. H. (2010). Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical Reviews*, *110*(3), 1463–1497.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (n.d.). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Burton, B., Zimmermann, M. T., Jernigan, R. L., & Wang, Y. (2012). A computational investigation on the connection between dynamics properties of ribosomal proteins and ribosome assembly. *PLoS Computational Biology*, *8*(5), e1002530.

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

Doruker, P., & Jernigan, R. L. (2003). Functional motions can be extracted from on-lattice construction of protein structures. *Proteins: Structure, Function and Genetics*, *53*(2), 174–181.

Fenwick, R. B., van den Bedem, H., Fraser, J. S., & Wright, P. E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(4), E445-54.

Fernandez-Leiro, R., & Scheres, S. H. W. (2016). Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, *537*(7620), 339–346.

Frappier, V., & Najmanovich, R. J. (2014). A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. *PLoS Computational Biology*, *10*(4).

Fuglebakk, E., Reuter, N., & Hinsen, K. (2013). Evaluation of protein elastic network models based on an analysis of collective motions. *Journal of Chemical Theory and Computation*, *9*(12), 5618–5628.

Hamacher, K., & McCammon, J. A. (2006). Computing the amino acid specificity of fluctuations in biomolecular systems. *Journal of Chemical Theory and Computation*, *2*(3), 873–878.

Howe, P. W. (2001). Principal components analysis of protein structure ensembles calculated using NMR data. *Journal of Biomolecular NMR*, *20*(1), 61–70.

Hünenberger, P. H., Mark, A. E., & van Gunsteren, W. F. (1995). Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *Journal of Molecular Biology*, *252*(4), 492–503.

Ichiye, T., & Karplus, M. (1991). Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, *11*(3), 205–217.

Jernigan, R. L., Yang, L., Song, G., Kurkckuoglu, O., & Doruker, P. (2009). Elastic Network Models of Coarse-Grained Proteins Are Effective for Studying the Structural Control Exerted over Their Dynamics. In G. A. Voth (Ed.), *Coarse-Graining of Condensed Phase and Biomolecular Systems* (pp. 237–254). Boca Raton, FL: CRC Press.

Katebi, A. R., Sankar, K., Jia, K., & Jernigan, R. L. (2015). The use of experimental structures to model protein dynamics. In *Molecular Modeling of Proteins* (Vol. 1215, pp. 213–236).

Keskin, O., Bahar, I., Flatow, D., Covell, D. G., & Jernigan, R. L. (2002). Molecular mechanisms of chaperonin GroEL-GroES function. *Biochemistry*, *41*(2), 491–501.

Kim, M. H., Seo, S., Jeong, J. Il, Kim, B. J., Liu, W. K., Lim, B. S., … Kim, M. K. (2013). A mass weighted chemical elastic network model elucidates closed form domain motions in proteins. *Protein Science*, *22*(5), 605–613.

Kohn, J. E., Afonine, P. V, Ruscio, J. Z., Adams, P. D., & Head-Gordon, T. (2010). Evidence of functional protein dynamics from X-ray crystallographic ensembles. *PLoS Computational Biology*, *6*(8), 1–5.

Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins*, *64*(3), 559–574.

Lyman, E., Pfaendtner, J., & Voth, G. A. (2008). Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophysical Journal*, *95*(9), 4183–4192.

Ma, J. (2004). New advances in normal mode analysis of supermolecular complexes and applications to structural refinement. *Current Protein & Peptide Science*, *5*(2), 119–123.

Meyer, T., D’Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., … Orozco, M. (2010). MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure*, *18*(11), 1399–1409.

Miyazawa, S., & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, *256*(3), 623–644.

Moritsugu, K., & Smith, J. C. (2007). Coarse-grained biomolecular simulation with REACH: realistic extension algorithm via covariance Hessian. *Biophysical Journal*, *93*(10), 3460–3469.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(11), 559–572.

Sanejouand, Y.-H. (2011). Elastic Network Models: Theoretical and Empirical Foundations. *Biomolecular Simulations*, 26.

Sankar, K., Jia, K., & Jernigan, R. L. (2017). Knowledge-based entropies improve the identification of native protein structures. *Proceedings of the National Academy of Sciences*, *114*(11), 2928–2933.

Sankar, K., Liu, J., Wang, Y., & Jernigan, R. L. (2015). Distributions of experimental protein structures on coarse-grained free energy landscapes. *Journal of Chemical Physics*, *143*(24), 243153.

Setny, P., & Zacharias, M. (2013). Elastic network models of nucleic acids flexibility. *Journal of Chemical Theory and Computation*, *9*(12), 5460–5470.

Sievers, F., & Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology*, *1079*, 105–116.

Skjaerven, L., Martinez, A., & Reuter, N. (2011). Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins*, *79*(1), 232–243.

Skjærven, L., Yao, X.-Q., Scarabelli, G., & Grant, B. J. (2014). Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics*, *15*(1), 399.

Soheilifard, R., Makarov, D. E., & Rodin, G. J. (2008). Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. *Physical Biology*, *5*(2), 26008.

Tama, F., & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Engineering*, *14*(1), 1–6.

Teodoro, M. L., Phillips, G. N., & Kavraki, L. E. (2002). A dimensionality reduction approach to modeling protein flexibility. *Proceedings of the Sixth Annual International Conference on Computational Biology RECOMB 02*, 299–308.

Teodoro, M. L., Phillips Jr., G. N., & Kavraki, L. E. (2003). Understanding Protein Flexibility through Dimensionality Reduction. *Journal of Computational Biology*, *10*, 617–634.

van Aalten, D. M., Conn, D. A., de Groot, B. L., Berendsen, H. J., Findlay, J. B., & Amadei, A. (1997). Protein dynamics derived from clusters of crystal structures. *Biophysical Journal*, *73*(6), 2891–2896.

Wang, Y., & Jernigan, R. L. (2005). Comparison of tRNA motions in the free and ribosomal bound structures. *Biophysical Journal, 89*(5), 3399–3409.

Wang, Y., Rader, a J., Bahar, I., & Jernigan, R. L. (2004). Global ribosome motions revealed with elastic network model. *Journal of Structural Biology*, *147*(3), 302–314.

Yang, L.-W., Eyal, E., Bahar, I., & Kitao, A. (2009). Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. *Bioinformatics (Oxford, England)*, *25*(5), 606–614.

Yang, L., Song, G., Carriquiry, A., & Jernigan, R. L. (2008). Close Correspondence between the Motions from Principal Component Analysis of Multiple HIV-1 Protease Structures and Elastic Network Modes. *Structure*, *16*(2), 321–330.

Yang, L., Song, G., & Jernigan, R. L. (2007). How well can we understand large-scale protein motions using normal modes of elastic network models? *Biophysical Journal*, *93*(3), 920–929.

Yang, L., Song, G., & Jernigan, R. L. (2009a). Comparisons of experimental and computed protein anisotropic temperature factors. *Proteins*, *76*(1), 164–175.

Yang, L., Song, G., & Jernigan, R. L. (2009b). Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(30), 12347–12352.

Yang, Z., Májek, P., & Bahar, I. (2009). Allosteric Transitions of Supramolecular Systems Explored by Network Models: Application to Chaperonin GroEL. *PLoS Computational Biology*, *5*(4), e1000360.

Zimmermann, M. T., & Jernigan, R. L. (2014). Elastic network models capture the motions apparent within ensembles of RNA structures. *RNA*, *20*(6), 792–804.

CHAPTER 6. CONCLUSION

"...if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

— Richard Feynman
The Feynman Lectures on Physics

In this dissertation, we present research that exploits the incredible volume of structure and sequence information available for proteins and supplement it with molecular simulations that inform about protein dynamics. The protein data bank serves as an excellent resource for information on protein structure data and is complemented by cross-references to protein sequence repositories such as UniProt and other fold-based classification databases such as CATH, SCOP and Pfam. An important aspect of the PDB database that can also be exploited for the ensemble nature of proteins is its redundancy in terms of structures (multiple structures for the same protein, under different conditions, or mutants). As sequence, structure and dynamics are collectively linked to protein function, combining information for these three aspects is imperative to truly understand protein function. The challenge, however, is in integrating the available sequence and structure information with dynamics and correlating with protein function. The research reported in this dissertation addresses this concern by developing computational methods that collectively use sequence, structure and dynamic information to interpret protein function. In this section, I will first highlight some of the important research findings in this dissertation. Then, I will present some extensions with the potential to improve and extend this current work in future directions.

6.1. Important Research Outcomes

6.1.1. Global Changes to Protein Dynamics upon Oligomerization

While oligomerization is often perceived as a means of stabilizing proteins, our work suggests that certain regions in proteins, especially those residues on the surface, may exhibit increases in their mobilities, while residues at the interfaces tend to be more stabilized. It is also seen that some residues with lower packing densities in the interface behave in the opposite way and become more mobile upon oligomerization. Our findings corroborate previous outcomes that suggested some increases in mobilities upon binding to a partner ligand (Kay, Muhandiram, Farrow, Aubin, & Forman-Kay, 1996) or even another macromolecule, like DNA (Yu, Zhu, Tse-Dinh, & Fesik, 1996). It was previously suggested that such losses in mobility upon binding at the interface closely correlate to a loss of entropy and is often compensated by an increase in mobility at another site and hence, an increase in entropy (Forman-Kay, 1999). We have shown similar results in the context of oligomerization for a diverse set of oligomeric proteins.

6.1.2. Changes to Dynamic Communities upon Oligomerization

Considering triosephosphate isomerase, we have revealed that oligomerization can lead to changes in its dynamic communities. In this particular case, we observe changes to the community structure of the active site core; while the active site in the monomer remains rigid, the dimer active site is split into two communities that are nearly anti-correlated in their dynamics. In this case this community division is essential for the optimal orientation of the substrate and execution of the enzyme’s catalytic activity.

6.1.3. Significance of Oligomerization for Key Functional Residues

One of the key findings from investigating the altered dynamics of protein subunits upon oligomerization is the increased stability of functional binding residues. Interestingly, these residues could be localized distant from the interface, often considered as a hot spot for binding residues. These residues could be further investigated as potential drug targets or for site-directed mutagenesis experiments to validate this finding.

6.1.4. Correspondence between Dynamic Communities from MD and ENM and Screening Deleterious Mutants

We have shown that ENMs can be used to mine dynamic protein communities and that there is a significant correspondence between the communities derived from MD and from ENM. Comparison of the inter-residue cross-correlation matrices (essential for mining communities) shows significant agreement between MD and ENM; there is a high correlation for node centrality and significant overlap for the root mean-square inner product for these matrices. As MD simulations are expensive in terms of time and also demand considerable computing power for larger proteins, a simple alternative such as ENM should be widely useful as it directly overcomes these computational limitations. Also, we have shown that atomic formulations of ENM capture the community differences for wild type, stable and unstable mutants of T4 lysozyme, with the deleterious mutants showing substantial differences in the distribution of their communities compared to the wild type.

6.1.5. Role of Structural Dynamics in Predicting Allosteric and Active Site Residues

Including information on protein dynamics along with structural, physico-chemical and evolutionary features leads to considerable improvements in predictions for functional and regulatory binding residues. In addition, we see that there can be considerable overlap between the features of residues in the active and allosteric sites. Also, it is noteworthy that a common subset of features can be used for predictions of both allosteric and active site residues. A key finding in this work is that our method predicted some residues as both active site and allosteric site residues. At initial glance, this may hint at the possibility of false positives, however upon verifying their evolutionary conservation, these residues are found to be conserved, suggesting that these are functionally important residues. Previous studies established that active sites can be allosterically coupled with each other, so that the predicted sites with dual character may be sites capable of serving as both functional ligand binding and regulation sites.

Combining the dynamic communities with allosteric pathways immediately brings out the suggestion that allosteric pathways that connect between the individual communities are particularly important, but the pathways through the communities themselves are likely to be redundant and unimportant as long as they connect between allosteric residues on the surfaces of the communities. There is a long-standing disagreement between allosteric pathways predicted in different ways. These differences in the pathways might be resolved, if the connected residues inside of the communities are the only points that differ, a point that can be based upon the concept of the communities being quite rigid.

6.1.6. Comparisons between Different Formulations of ENM

We have revealed that the intrinsic dynamics of proteins is more closely replicated by modeling the stiffness of springs between residue pairs as interactions that weaken with distance than by taking into consideration the actual interaction type or by assigning the same stiffness between all residue interactions. This type of distance-based interaction modeling is utilized in the parameter free ANM (pfANM) and it performs better than four of the other types of ENMs studied. Furthermore, it is seen that deriving the stiffness of springs from calculations of internal distance changes from MD trajectories does not lead to improvements in performance. Instead, it is observed that including information on residue-residue entropies into the stiffness of the springs leads to improved performance compared to two other ENM formulations.

6.2. Scope for Improvement and Future Directions

6.2.1. Investigating the Dynamics of Interface Residues and Prediction of True Oligomeric State

We have shown that most interface residues show considerable reduction in their fluctuations upon oligomerization. This observation is consistent with the rationale that upon assembly, the degrees of freedom for residues in the interface decreases, thus reducing their mobilities. However, it is also seen that certain residues in the interface exhibit increases in mobility following oligomerization. Our work suggests that these residues have lower packing densities compared to other interface residues. It would be interesting to consider specific cases of proteins showing increased mobilities of interface residues to investigate the functional roles of these residues. A study which considers proteins with and without these residues and their binding efficiency to partner subunits might shed light if these residues have any role in facilitating oligomerization. Also, machine learning schemes may be implemented that consider the mobility changes of residues in order to predict their functional oligomeric forms.

6.2.2. Changes to Community Architecture of Enzyme Catalytic Residues

We showed in the case of triosephosphate isomerase that oligomerization changes the dynamically cohesive nature of residues in the catalytic core. It will be interesting to further elaborate on this. For example, information for a diverse set of enzymes that function only in their oligomeric forms could be collected. These enzymes should be further divided into two sets: those having all the catalytic residues within a single subunit and those that form shared active sites between subunits. Then it would be interesting to compare the effect of assembly on the community structures of these two sets. We hypothesize that those enzymes not having an active site shared between subunits might show a similar splitting of the active site as we observed for TIM while the same may not be true for the other set. This would be an appropriate extension of our study presented in this thesis.

6.2.3. Effect of Oligomerization on Heteromers

The study of the effect of oligomerization may also be extended to heteromers, i.e. assemblies whose subunits differ from one another. It would be particularly interesting to investigate cases that form multimers with several different other subunits.

6.2.4. Using Residue Frustration as a Feature for Predicting Functional Residues

Frustration is inherent to those residues whose rotameric states are not that of the minimum energy configuration (Ferreiro, Komives, & Wolynes, 2014). Previous studies suggest that frustration is introduced on purpose at certain residues of functional importance. The residue level frustration could be calculated using the Frustratometer (Jenik et al., 2012) and it would be interesting to see whether the prediction of functional residues is improved by including residue frustration as an additional feature.

REFERENCES

Amadei, A., Linssen, A. B., & Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins*, *17*(4), 412–425.

ANFINSEN, C. B., & HABER, E. (1961). Studies on the reduction and re-formation of protein disulfide bonds. *The Journal of Biological Chemistry*, *236*, 1361–1363.

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, *80*(1), 505–515.

Bahar, I., Atilgan, A. R., & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design*, *2*(3), 173–181.

Bahar, I., & Jernigan, R. L. (1999). Cooperative fluctuations and subunit communication in tryptophan synthase. *Biochemistry*, *38*(12), 3478–3490.

Bahar, I., Lezon, T. R., Bakan, A., & Shrivastava, I. H. (2010). Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chemical Reviews*, *110*(3), 1463–1497.

Behr, J. P. (2007). *The Lock-and-Key Principle, The State of the Art--100 Years On*. *The Lock-and-Key Principle, The State of the Art--100 Years On* (Vol. 1).

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Databank. *Nucleic Acids Research*, *28*, 235–242.

Chang, C. A., McLaughlin, W. A., Baron, R., Wang, W., & McCammon, J. A. (2008). Entropic contributions and the influence of the hydrophobic environment in promiscuous protein-protein association. *Proceedings of the National Academy of Sciences*, *105*(21), 7456–7461.

Changeux, J.-P., & Edelstein, S. J. (2005). Allosteric mechanisms of signal transduction. *Science (New York, N.Y.)*, *308*(5727), 1424–1428.

Cheng, J., Tegge, A. N., & Baldi, P. (2008). Machine Learning Methods for Protein Structure Prediction. *IEEE Reviews in Biomedical Engineering*, *1*, 41–49.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–826.

Copley, S. D. (2003). Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology*, *7*(2), 265–272.

Coureux, P. D., Sweeney, H. L., & Houdusse, A. (2004). Three myosin V structures delineate essential features of chemo-mechanical transduction. *EMBO Journal*, *23*(23), 4527–4537.

Dale, J. M., Popescu, L., & Karp, P. D. (2010). Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, *11*.

Damm, K. L., & Carlson, H. A. (2007). Exploring experimental sources of multiple protein conformations in structure-based drug design. *Journal of the American Chemical Society*, *129*(26), 8225–8235.

Demerdash, O. N. A., Daily, M. D., & Mitchell, J. C. (2009). Structure-based predictive models for allosteric hot spots. *PLoS Computational Biology*, *5*(10), 14–19.

Dill, K. a. (1999). Polymer principles and protein folding. *Protein Science : A Publication of the Protein Society*, *8*(6), 1166–1180.

Doruker, P., Jernigan, R. L., & Bahar, I. (2002). Dynamics of large proteins through hierarchical levels of coarse-grained structures. *Journal of Computational Chemistry*, *23*(1), 119–127.

Durrant, J., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology*, *9*(71), 1–9.

Fermi, E., Pasta, J., & Ulam, S. (1955). Studies of nonlinear problems. *LASL Report LA-1940*.

Ferreiro, D. U., Komives, E. A., & Wolynes, P. G. (2014). Frustration in biomolecules. *Quarterly Reviews of Biophysics*, *47*(4), 285–363.

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.*, *27*, 2985–2993.

Forman-Kay, J. D. (1999). The “dynamics” in the thermodynamics of binding. *Nature Structural Biology*, *6*(12), 1086–1087.

General, I. J., Liu, Y., Blackburn, M. E., Mao, W., Gierasch, L. M., & Bahar, I. (2014). ATPase Subdomain IA Is a Mediator of Interdomain Allostery in Hsp70 Molecular Chaperones. *PLoS Computational Biology*, *10*(5).

Goldstein, H., Poole, P. C., & Safko, J. L. (2002). *Pearson Education - Classical Mechanics*. *Pearson*.

Grant, B. J., Gorfe, A. a, & Mccammon, J. A. (2010). Large conformational changes in proteins: signaling and other functions. *Curr Opin Struct Biol.*, *20*(2), 142–147.

Haliloglu, T., & Bahar, I. (2015). Adaptability of protein structures to enable functional interactions and evolutionary implications. *Current Opinion in Structural Biology*, *35*, 17–23.

Hastie, T., Tibshirani, R., Friedman, J., & others. (2009). *The elements of statistical learning*. *Springer New York USA HuberW* (Vol. 18).

Hensen, U., Meyer, T., Haas, J., Rex, R., Vriend, G., & Grubmüller, H. (2012). Exploring protein dynamics space: The dynasome as the missing link between protein structure and function. *PLoS ONE*, *7*(5).

Henzler-Wildman, K. A., Lei, M., Thai, V., Kerns, S. J., Karplus, M., & Kern, D. (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, *450*(7171), 913–916.

Henzler-Wildman, K., & Kern, D. (2007). Dynamic personalities of proteins. *Nature*, *450*(7172), 964–972.

Howe, P. W. A. (2001). Principal components analysis of protein structure ensembles calculated using NMR data. *Jounal of Biomolecular NMR*, *20*, 61–70.

Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G., & Chothia, C. (1999). SCOP: A structural classification of proteins database. *Nucleic Acids Research*.

James, L. C., & Tawfik, D. S. (2009). The specificity of cross-reactivity: Promiscuous antibody binding involves specific hydrogen bonds rather than nonspecific hydrophobic stickiness. *Protein Science*, *12*(10), 2183–2193.

Jenik, M., Parra, R. G., Radusky, L. G., Turjanski, A., Wolynes, P. G., & Ferreiro, D. U. (2012). Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Research*, *40*(Web Server issue), W348-51.

Jeong, J. I., Jang, Y., & Kim, M. K. (2006). A connection rule for α-carbon coarse-grained elastic network models using chemical bond information. *Journal of Molecular Graphics and Modelling*, *24*(4), 296–306.

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, *9*(9), 646–652.

Kay, L. E., Muhandiram, D. R., Farrow, N. A., Aubin, Y., & Forman-Kay, J. D. (1996). Correlation between dynamics and high affinity binding in an SH2 domain interaction. *Biochemistry*, *35*(2), 361–368.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*(4610), 662–666.

Kim, M. H., Seo, S., Jeong, J. Il, Kim, B. J., Liu, W. K., Lim, B. S., … Kim, M. K. (2013). A mass weighted chemical elastic network model elucidates closed form domain motions in proteins. *Protein Science*, *22*(5), 605–613.

Kohavi, R., & Provost, F. (1998). Glossary of Terms. *Machine Learning.*, *30*(2–3), 271–274.

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences*, *44*(2), 98–104.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268.

Kurkcuoglu, O., Jernigan, R. L., & Doruker, P. (2004). Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions. *Polymer*, *45*(2), 649–657.

Kurkcuoglu, O., Kurkcuoglu, Z., Doruker, P., & Jernigan, R. L. (2009). Collective dynamics of the ribosomal tunnel revealed by elastic network modeling. *Proteins: Structure, Function and Bioinformatics*, *75*(4), 837–845.

Kurkcuoglu, O., Turgut, O. T., Cansu, S., Jernigan, R. L., & Doruker, P. (2009). Focused functional dynamics of supramolecules by use of a mixed-resolution elastic network model. *Biophysical Journal*, *97*(4), 1178–1187.

Kurkcuoglu, Z., Bakan, A., Kocaman, D., Bahar, I., & Doruker, P. (2012). Coupling between Catalytic Loop Motions and Enzyme Global Dynamics. *PLoS Computational Biology*, *8*(9).

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., … Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112.

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., & Ortiz, A. R. (2005). An analysis of core deformations in protein superfamilies. *Biophysical Journal*, *88*(2), 1291–1299.

Levitt, M., & Warshel, A. (1975). Computer simulation of protein folding. *Nature*, *253*(5494), 694–698.

Linderstrøm-Lang, K. U. (1952). PROTEINS and ENZYMES. In *Lane Medical Lectures* (Vol. 6). Redwood City, CA, USA: Stanford University Press.

Lindorff-Larsen, K., Best, R. B., DePristo, M. A., Dobson, C. M., & Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature, 433*(7022), 128–132.

López-Blanco, J. R., & Chacón, P. (2016). New generation of elastic network models. *Current Opinion in Structural Biology*, *37*, 46–53.

Lu, S., Huang, W., & Zhang, J. (2014). Recent computational advances in the identification of allosteric sites in proteins. *Drug Discovery Today*, *19*(10), 1595–1600.

Maisuradze, G. G., & Leitner, D. M. (2006). Principal component analysis of fast-folding λ-repressor mutants. *Chemical Physics Letters*, *421*(1–3), 5–10.

Mannige, R. (2014). Dynamic New World: Refining Our View of Protein Structure, Function and Evolution. *Proteomes*, *2*(1), 128–153.

Marcos, E., Crehuet, R., & Bahar, I. (2010). On the conservation of the slow conformational dynamics within the amino acid kinase family: NAGK the paradigm. *PLoS Computational Biology*, *6*(4).

McCammon, J. A. (1984). Protein Dynamics. *Reports on Progress in Physics*, *47*(1), 1–46.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature*, *267*(5612), 585–590.

McClendon, C. L., Kornev, A. P., Gilson, M. K., & Taylor, S. S. (2014). Dynamic architecture of a protein kinase. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(43), E4623-31.

Ng, A. (2012). 1. Supervised learning. *Machine Learning*, 1–30.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., & Thornton, J. (1997). CATH – a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093–1109.

Petrova, N. V, & Wu, C. H. (2006). Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, *7*, 312.

Rader, A. J., Chennubhotla, C., Yang, L.-W., & Bahar, I. (2006). The Gaussian Network Model: theory and applications. *Normal Mode Analysis - Theory and Applications to Biological and Chemical Systems*, *10*(20), 41–64.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94.

Sander, C., & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, *9*(1), 56–68.

Sankararaman, S., Sha, F., Kirsch, J. F., Jordan, M. I., & Sjölander, K. (2010). Active site prediction using evolutionary and structural information. *Bioinformatics*, *26*(5), 617–624.

Scheraga, H. A., Khalili, M., & Liwo, A. (2007). Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annual Review of Physical Chemistry*, *58*(1), 57–83.

Singh, T., Biswas, D., & Jayaram, B. (2011). AADS - An automated active site identification, docking, and scoring protocol for protein targets based on physicochemical descriptors. *Journal of Chemical Information and Modeling*, *51*(10), 2515–2527.

Sinitskiy, A. V, Saunders, M. G., & Voth, G. a. (2012). Optimal number of coarse-grained sites in different components of large biomolecular complexes. *The Journal of Physical Chemistry. B*, *116*(29), 8363–8374.

Skjaerven, L., Martinez, A., & Reuter, N. (2011). Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins*, *79*(1), 232–243.

Sundberg, E. J., & Mariuzza, R. A. (2000). Luxury accommodations: the expanding role of structural plasticity in protein–protein interactions. *Structure*, *8*(7), R137–R142.

Tama, F., & Brooks, C. L. (2006). SYMMETRY, FORM, AND SHAPE: Guiding Principles for Robustness in Macromolecular Machines. *Annual Review of Biophysics and Biomolecular Structure*, *35*(1), 115–133.

Tavernelli, I., Cotesta, S., & Di Iorio, E. E. (2003). Protein dynamics, thermal stability, and free-energy landscapes: a molecular dynamics investigation. *Biophysical Journal*, *85*(4), 2641–2649.

Teilum, K., Olsen, J. G., & Kragelund, B. B. (2009). Functional aspects of protein flexibility. *Cellular and Molecular Life Sciences*, *66*(14), 2231–2247.

Timothy R. Lezon, Indira H. Shrivastava, Z. Y. and I. B. (2009). Elastic Network Models For Biomolecular Dynamics: Theory and Application to Membrane Proteins and Viruses. *Handbook on Biological Networks*, 129–158.

Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, *77*(9), 1905–1908.

Tozzini, V. (2005). Coarse-grained models for proteins. *Current Opinion in Structural Biology*, *15*(2), 144–150.

Vértessy, B. G., & Orosz, F. (2011). From “fluctuation fit” to “conformational selection”: Evolution, rediscovery, and integration of a concept. *BioEssays*.

Wang, Y., Rader, a J., Bahar, I., & Jernigan, R. L. (2004). Global ribosome motions revealed with elastic network model. *Journal of Structural Biology*, *147*(3), 302–314.

Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavlou, G., Eisenmesser, E. Z., & Kern, D. (2004). Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nature Structural and Molecular Biology*, *11*(10), 945–949.

Yang, L., Song, G., Carriquiry, A., & Jernigan, R. L. (2008). Close Correspondence between the Motions from Principal Component Analysis of Multiple HIV-1 Protease Structures and Elastic Network Modes. *Structure*, *16*(2), 321–330.

Yang, L., Song, G., & Jernigan, R. L. (2009). Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(30), 12347–12352.

Yang, L. W., & Bahar, I. (2005). Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. *Structure*, *13*(6), 893–904.

Yu, L., Zhu, C. X., Tse-Dinh, Y. C., & Fesik, S. W. (1996). Backbone dynamics of the C-terminal domain of Escherichia coli topoisomerase I in the absence and presence of single-stranded DNA. *Biochemistry*, *35*, 9661–9666.

Zhang, Z., Lu, L., Noid, W. G., Krishna, V., Pfaendtner, J., & Voth, G. a. (2008). A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophysical Journal*, *95*(11), 5073–5083.

APPENDIX A. SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1. Counts of the number of residues in interface and non-interface regions having increased, decreased and unchanged MSF values for glutamate dehydrogenase (bovine).

PDB ID	Residues	Function	Number of Residues with MSF Increased	Number of Residues with MSF Unchanged	Number of Residues with MSF Decreased	Number of Non-Interface Residues	Fraction of Functional Non-Interface Residues with Reduced Fluctuation
3mw9	K90,K114,K126, R211,S381	Substrate binding, catalytic activity	0	1	4	5	4/5
3mw9	H209,R217,R261,R265	GTP binding, allosterically regulates the protein	1	3	0	4	0/5

Table A.2. Counts of the number residues in interface and non-interface regions having increased, decreased and unchanged MSF values for arginase 1 (rat).

PDB ID	Residues	Function	Number of Residues with MSF Increased	Number of Residues with MSF Unchanged	Number of Residues with MSF Decreased	Number of Non-Interface Residues	Fraction of Functional Non-Interface Residues with Reduced Fluctuation
1rla	H101,D124,H126,D128, D232,D234	Metal binding	0	0	6	6	6/6
1rla	H101,D128,H141,D232, D234,G235	Deleterious mutation sites	0	0	6	6	6/6
1rla	H141,E277	Substrate binding residues (involved in the catalytic activity of the enzyme)	0	0	2	2	2/2

Table A.3. Counts of the number residues in interface and non-interface regions having increased, decreased and unchanged MSF values for glycine N-methyltransferase (rat).

PDB ID	Residues	Function	Number of Residues with MSF Increased	Number of Residues with MSF Unchanged	Number of Residues with MSF Decreased	Number of Non-Interface Residues	Fraction of Functional Non-Interface Residues with Reduced Fluctuation
1bhj	Y21,W30,R40,A64, D85, N116,W117, L136,H142	S-adenosyl methionine binding res idues	0	1	8	2	1/2
1bhj	Y33,G137,N138, R175, Y194, Y220,Y242	Glycine-binding residues	0	1	6	7	6/7
1bhj	Y21,Y33, Y194,Y220	Mutation sites which reduce the catalytic efficiency of the enzyme	0	1	3	3	2/3

Table A.4. Counts of the number residues in interface and non-interface regions having increased, decreased and unchanged MSF values for D-aminoacid oxidase (yeast) at 1.5 fold change cutoff.

PDB ID	Residues	Function	Number of Residues with MSF Increased	Number of Residues with MSF Unchanged	Number of Residues with MSF Decreased	Number of Non-Interface Residues	Fraction of Functional Non-Interface Residues with Reduced Fluctuation
1c0k	S1012,S1015,A1034, R1035,A1047,S1048, G1052,N1054,V1162, S1334,S1335,G1337, Y1338,Q1339	FAD-binding residues	0	10	4	14	4/14
1c0k	Y1223,Y1238, R1285,S1335	Active site residues	0	3	1	4	1/4

Table A.5. Counts of the number residues in interface and non-interface regions having increased, decreased and unchanged MSF values for d-aminoacid oxidase (yeast) at 1.25 fold change cutoff.

PDB ID	Residues	Function	Number of Residues with MSF Increased	Number of Residues with MSF Unchanged	Number of Residues with MSF Decreased	Number of Non-Interface Residues	Fraction of Functional Non-Interface Residues with Reduced Fluctuation
1c0k	S1012,S1015,A1034, R1035,A1047,S1048, G1052,N1054,V1162, S1334,S1335,G1337, Y1338,Q1339	FAD-binding residues	0	6	8	14	8/14
1c0k	Y1223,Y1238,R1285, S1335,	Active site residues	0	3	1	4	1/4

Table A.6. Clusters from dendrograms. We perform hierarchical clustering of the matrix of correlated fluctuations using the MATLAB clustering module (http://www.mathworks.com /help/stats/hierarchical-clustering.html) and then truncate the dendrograms for the unbound (8tim) and substrate bound (1tph) form of TIM in their monomeric and oligomeric forms at the above levels to obtain the desired number of clusters.

PDB	State	Truncation level (%age of max tree height)	Number of clusters
1tph	Monomer	90%	2
1tph	Oligomer	90%	2
8tim	Monomer	90%	2
8tim	Oligomer	90%	2
1tph	Monomer	77%	3
1tph	Oligomer	77%	3
8tim	Monomer	83%	3
8tim	Oligomer	83%	3
1tph	Monomer	74%	4

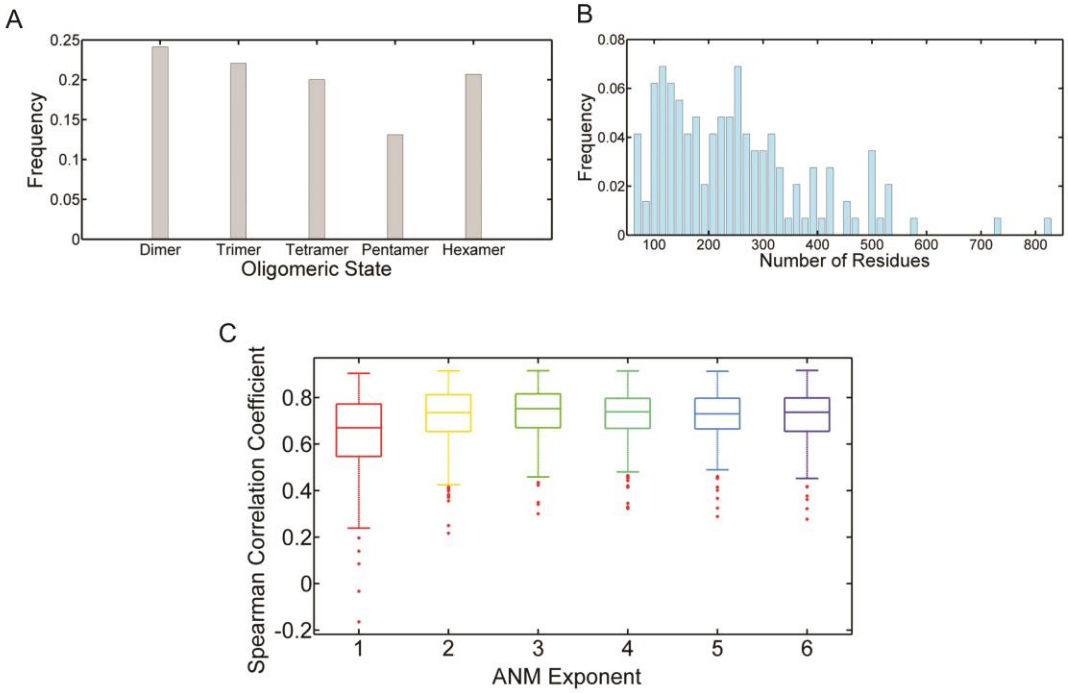

Figure A.1. Dataset characterization and ANM exponent. (A) The dataset includes a similar number of proteins for each type of oligomeric group except for pentamers, which are less abundant. (B) Distribution of the number of residues per protein for the proteins included in the dataset. (C) Correlation of Anisotropic Network Model (ANM) predicted Mean Square Fluctuations (MSF) with experimental B factors for different ANM exponents (*a*). The exponent *a* = 3 yields the highest average correlation with the experimental B factors.

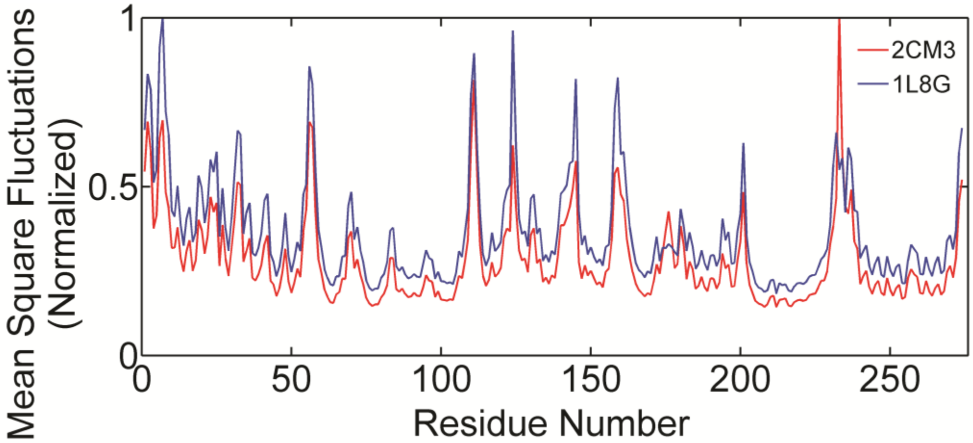

Figure A.2. Correlation of mean square fluctuations (MSF) of the monomeric and oligomeric forms of tyrosine phosphatase. The MSF calculations are performed on the crystallized monomeric form of the enzyme (pdb 1L8G) and a single monomer obtained from the enzyme oligomer (pdb 2CM3). We observe a very high correlation in the dynamics of the two structures (spearman correlation coefficient 0.98).

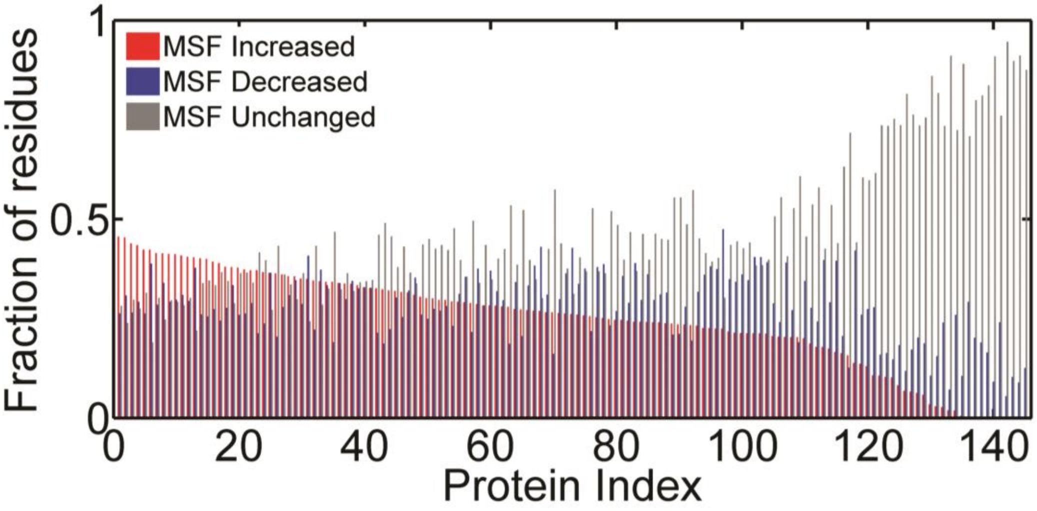

Figure A.3. Fractions of residues having increased, decreased and no significant change in MSF upon oligomerization for each protein (at fold change cutoff 1.5) in the dataset of 145 proteins. The plot is sorted by the fraction of residues with increased fluctuations.

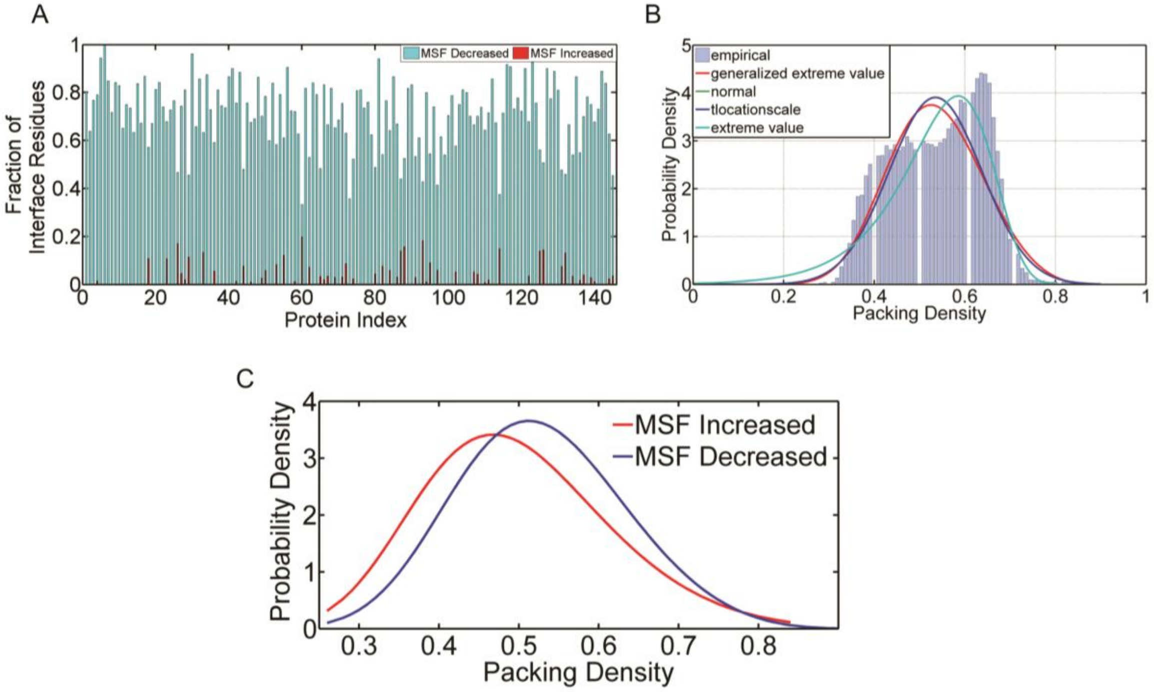

Figure A.4. Fluctuation change and packing density distribution for interface residues. (A) Interface Residues with Increased MSF. For some proteins, a small fraction of the interface residues (red bars) show increased fluctuations upon oligomerization. (B) Probability density fit for packing densities. The distribution of packing densities of residues computed from Voronoia shows best fit with the Generalized Extreme Value (GEV) distribution with a negative log likelihood value of 3.12 e^4. Other distributions to which the packing density data is fit are tlocationscale and extreme value, both of which are used to model data with heavy tails. (C) Packing density distribution for interface residues. The interface residues with increased MSF have lower packing densities than the residues with reduced MSF.

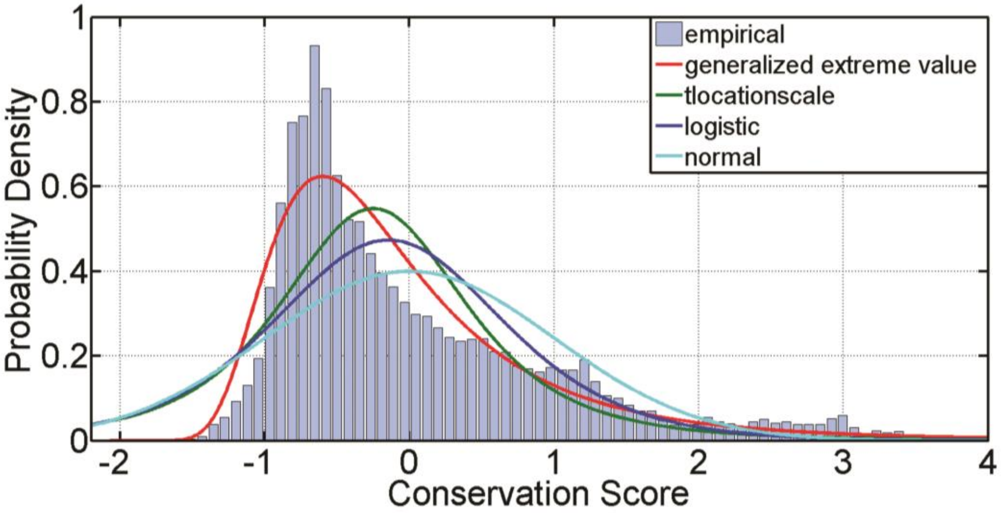

Figure A.5. Probability density fits for residue conservation scores. The conservation profile of residues shows a best fit with the Generalized Extreme Value (GEV) distribution (negative log likelihood 4.29 e^4) in comparison to several other distributions. Both the tlocationscale and logistic distributions are used to model data distributions which have heavier tails compared to the normal distribution.

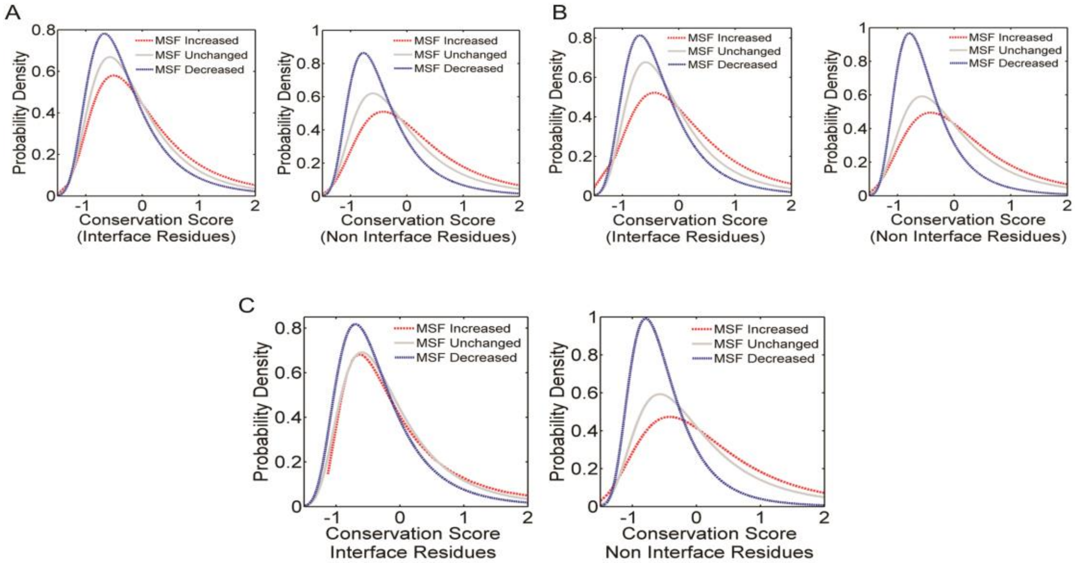

Figure A.6. Distribution of conservation scores of interface and non-interface residues that have different Fold Change Ratio (FCR) cutoffs. Results are shown for FCR 1.25 (A), 1.75 (B) and 2 (C); we observe the same pattern i.e. residues having reduced fluctuations are more conserved than the others, both in interface and

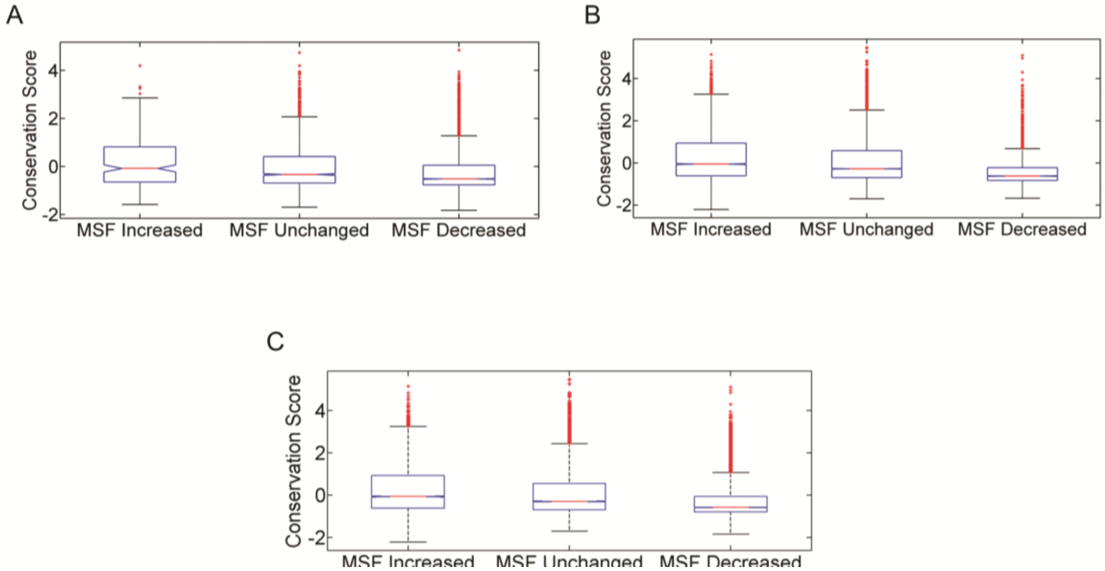

Figure A.7. Non-parametric test of significance for residue conservation with MSF change. Kruskal-Wallis test of significance for the conservation scores of three MSF change categories (increased, unchanged and decreased) for only interface (A), only non-interface (B) and all residues (C). The p-values for the respective tests were: interface residues p-value = 2.694 e^{-29}, non-interface residues p-value = 1.86 e^{-266} and all residues p-value = 0.

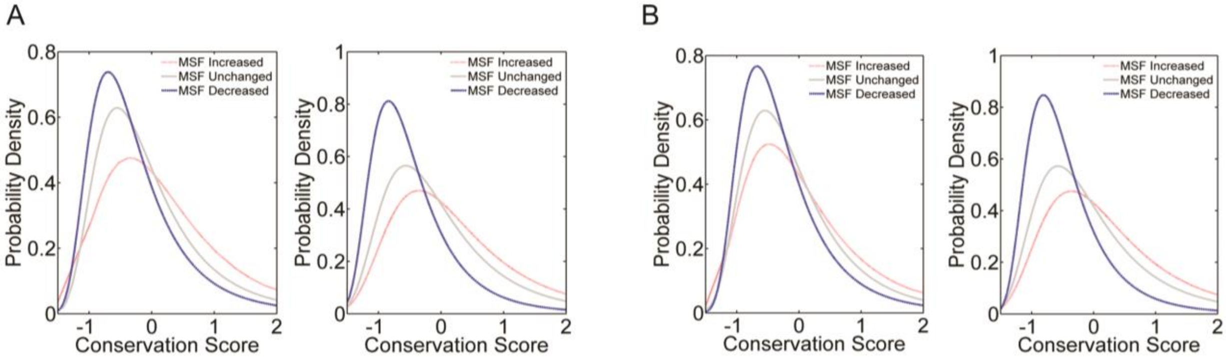

Figure A.8. Distribution of conservation scores for two smaller datasets. (A) Dataset with 40 proteins, and (B) dataset with 80 proteins. For each figure, the distribution for interface residues is shown on *left* and non-interface residues on *right*.

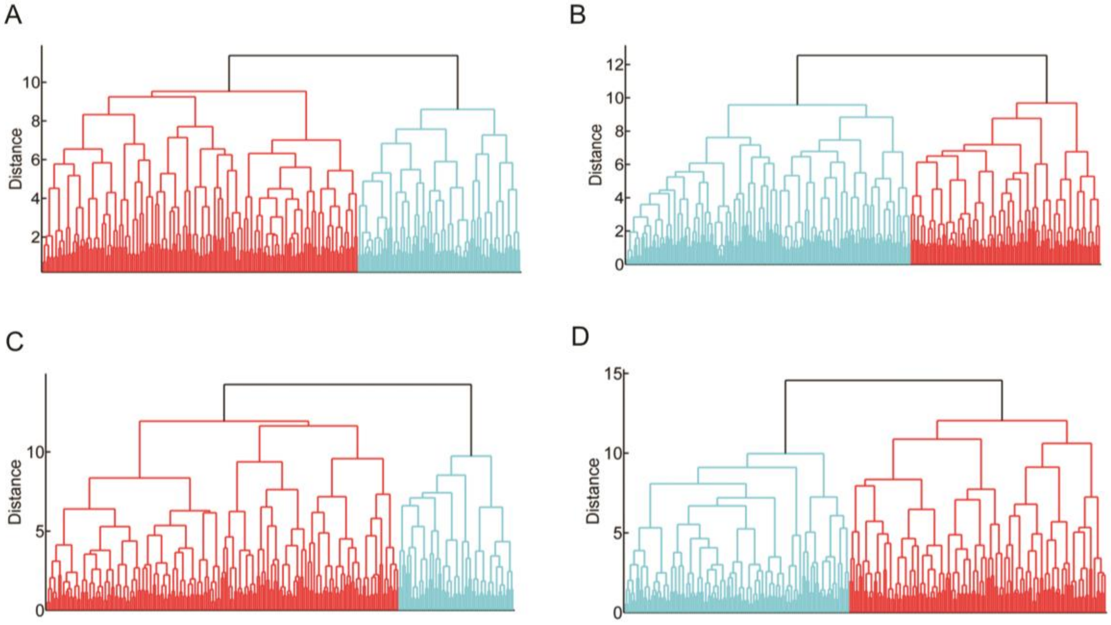

Figure A.9. Dendrograms from hierarchical clustering of residue fluctuation correlations. All trees are truncated at 90 percent of their maximum heights to generate 2 clusters. Clusters obtained for the isolated monomer without substrate (A) and monomer with substrate (B). Corresponding PDB files used were 8tim and 1tph. Clusters obtained for the monomer computed as part of oligomer without substrate (C) and with substrate (D).

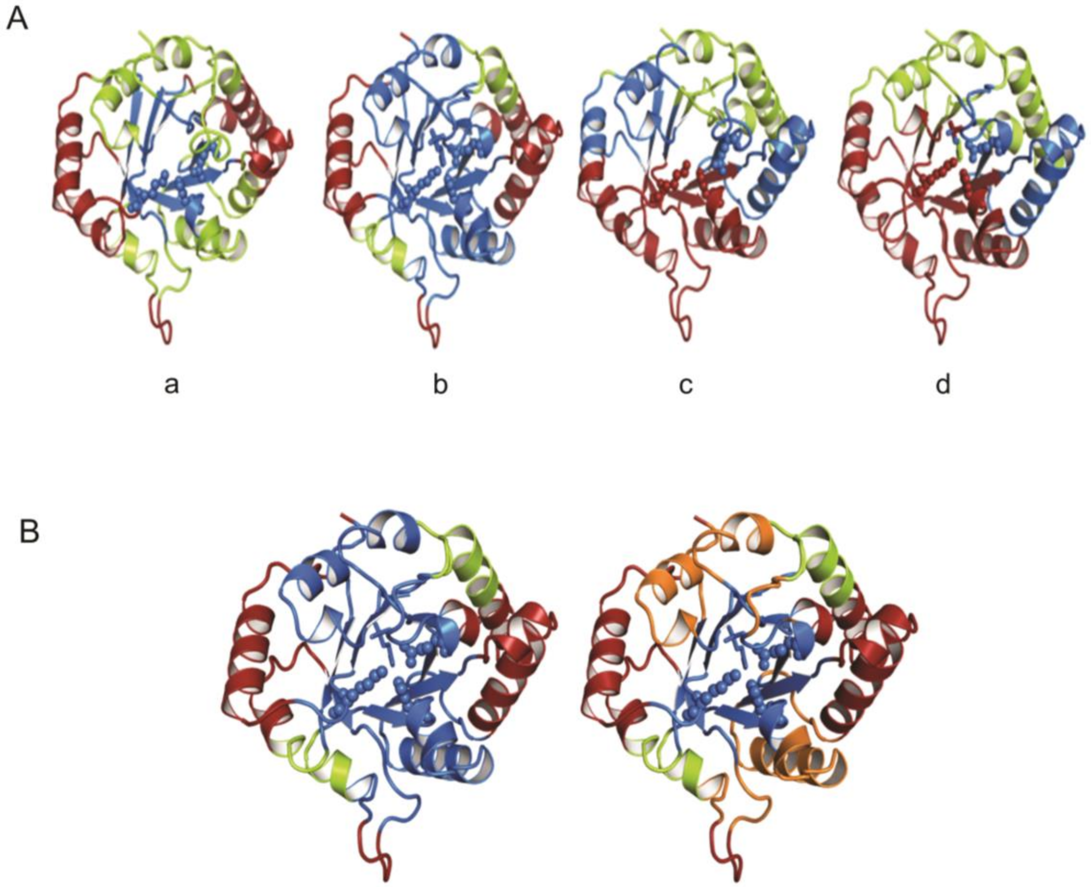

Figure A.10. Hierarchical clustering for 3 clusters. A**.** Mapping of 3 clusters from hierarchical clustering. The dendrogram for 8tim was truncated at 83 percent of its maximum height (both for the isolated monomer and monomer in the context of the oligomer) and the 1tph correlation cluster was truncated at 77 percent to yield 3 clusters. Regions of the structure that map to the same community are colored the same. a**.** Monomer without substrate (8tim) in isolation, b. Monomer with substrate (1tph) in isolation, c. Monomer in context of the oligomer without substrate (8tim) and d. Monomer in context of the oligomer with substrate (1tph). B. Close up view of the monomer with substrate (1tph) in isolation with 3 and 4 clusters. The dendrogram for 1tph in isolation was cut at 74 percent of its maximum height to produce 4 clusters. It can be seen that even with 4 clusters the active site of the monomer remains rigid unlike the oligomer, where the E165 residue moves in coordination with loop 6 that closes over the active site.

APPENDIX B. SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Table B.1. Dataset of proteins used in the study. The MD trajectories were downloaded from the MOlecular Dynamics Extended Library (MODEL) database. We retained proteins having at least 50 residues with a minimum trajectory time scale of 100 ns. The table is sorted by the number of residues.

PDB ID	Simulation Program	Duration	Protein Name	Number of residues
2gb1	Amber 8	1000 ns	Protein G	56
1bpi	Amber 8	100 ns	Bovine Pancreatic Trypsin Inhibitor	58
1g6x	Amber 8	100 ns	Pancreatic trypsin inhibitor	58
1ark	Amber 8	108.93 ns	Nebulin	60
1i6f	Amber 8	100 ns	Neurotoxin V5	60
1fas	Amber 8	100 ns	Fasciculin 1	61
3ci2	Amber 8	100 ns	Chymotrypsin inhibitor 2	64
1csp	Amber 8	100 ns	Cold Shock protein	67
1sdf	Amber 8	100 ns	Stromal Cell Derived factor	67
1tba	Amber 8	134.2 ns	Transcription initiation factor IID	67
1fvq	Amber 8	100 ns	Copper transporting ATPase	72
1jw2	Amber 8	100 ns	Hemolysin Expression modulating protein	72
1txa	Amber 8	100 ns	Toxin B	73
4icb	Amber 8	100 ns	Calbindin D9K	76
1sro	Amber 8	100 ns	PNPase	76
1ubq	Amber 8	811.5 ns	Ubiquitin	76
1pht	Amber 8	100 ns	Phosphatidylinositol kinase	83
1cei	Amber 8	107.06 ns	Colicin E7 Immunity Protein	85
1ls9	Amber 8	100 ns	Cytochrome C6	91
1j5d	Amber 8	100 ns	Plastocyanin	98
1opc	Amber 8	586.22 ns	OMPR	99
1kte	Amber 8	1001.0 ns	Thioltransferase	105
1fkb	Amber 8	100 ns	Fk506 Binding Protein	107
1nso	Amber 8	100.22 ns	Retroviral Protease	107
1jli	Amber 8	100 ns	Interleukin 3	112
1ooi	Amber 8	100 ns	Odorant binding protein (LUSH)	124
1agi	Amber 8	100 ns	Angiogenin	125
1k40	Amber 8	100.22 ns	Adhesin kinase	126
1bfg	Amber 8	105.832 ns	Basic fibroblast growth factor	126
1chn	Amber 8	100 ns	CHEY	126
1idr	Amber 8	159 ns	Hemoglobin Hbn	126
1lys	Amber 8	329.5 ns	Hen Egg White Lysozyme	129
1pdo	Amber 8	100 ns	Mannose Permease	129
1lit	Amber 8	100 ns	Lithostathine	131
1cbs	Amber 8	100 ns	Cellular retinoic acid binding protein	137
1kxa	Amber 8	100 ns	Sindbis virus capsid protein	158
1emr	Amber 8	100 ns	Leukemia Inhibitory Factor	159
1czt	Amber 8	100 ns	Protein (Coagulation Factor V)	160
1il6	Amber 8	100 ns	Interleukin 6	166
1sur	Amber 8	100 ns	PAPS Reductase	215
1acb	Amber 8	100 ns	Alpha Chymotrypsin	241
1cgi	Amber 8	100 ns	Alpha Chymotrypsinogen	245
2hvm	Amber 8	100 ns	Hevamine	273
1gnd	Amber 8	100 ns	Guanine nucleotide dissociation inhibitor	430

Table B.2. Distribution of $Kappa_{max}$ for the dataset. For each protein, we identified the distance cutoff r_c and community level N_c for which we obtained the maximum value for Kappa coefficient. We show the values for $Kappa_{max}$ (K_{-}max) for a subset of 5, 10, 20, 30 and 50 low frequency modes. The median $Kappa_{max}$ (K_{-}max) for 20 modes is 0.61 and mode for $r_c$is 7.5.

	5 modes			10 modes			20 modes			30 modes			50 modes		
PDB ID	K_{max}	r_c	N_c	K_{max}	r_c	N_c	K_{max}	r_c	N_c	K_{max}	r_c	N_c	K_{max}	r_c	N_c
2gb1	0.54	7	10	0.61	8	8	0.63	7.5	7	0.53	6.5	10	0.64	7	10
1bpi	0.56	6.5	4	0.63	8	2	0.65	7.5	2	0.59	7.5	2	0.55	6.5	2
1g6x	0.60	6.5	8	0.60	7	7	0.60	6.5	9	0.60	7	7	0.60	6.5	9
1ark	0.74	7.5	4	0.70	7	4	0.72	7.5	4	0.69	7.5	4	0.72	7.5	5
1i6f	0.62	7.5	6	0.58	6	6	0.62	6	6	0.62	7.5	6	0.60	6.5	6
1fas	0.50	8	10	0.50	8	10	0.50	6.5	9	0.48	7	9	0.47	7	9
3ci2	0.66	6.5	10	0.65	6.5	10	0.63	6.5	10	0.68	7	10	0.63	7.5	7
1csp	0.55	8	9	0.67	7	2	0.64	7	2	0.61	7	2	0.61	7	2
1sdf	0.74	7.5	2	0.74	7.5	2	0.74	7.5	2	0.74	7.5	2	0.74	7.5	2
1tba	0.81	7.5	2	0.78	8	2	0.64	7.5	7	0.78	8	3	0.83	8	2
1fvq	0.57	7.5	8	0.75	7.5	2	0.79	8	2	0.54	7.5	10	0.65	7.5	8
1jw2	0.55	8	6	0.51	7.5	8	0.51	7.5	9	0.57	7.5	8	0.52	8	9
1txa	0.64	8	6	0.64	8	3	0.65	7	7	0.64	8	3	0.57	8	10
1sro	0.76	8	2	0.73	7.5	2	0.75	6	3	0.68	7	2	0.66	6	6
1ubq	0.77	6	4	0.82	6	4	0.82	6	4	0.82	6	4	0.82	6	4
4icb	0.61	7.5	9	0.62	8	10	0.63	7.5	7	0.58	6.5	10	0.54	8	8
1pht	0.66	8	6	0.60	7	8	0.65	7	6	0.65	6	6	0.62	6	6
1cei	0.51	7.5	8	0.49	7	10	0.52	8	9	0.46	6.5	7	0.45	7	10
1ls9	0.56	6.5	9	0.54	6.5	10	0.60	7	10	0.53	6.5	10	0.58	6.5	9
1j5d	0.57	7	10	0.56	7	8	0.57	7	7	0.55	7.5	5	0.52	6	5
1opc	0.49	6.5	6	0.58	6	7	0.55	6	10	0.53	6.5	10	0.57	6	10
1kte	0.58	6.5	6	0.58	7.5	5	0.62	6.5	6	0.68	7.5	4	0.59	7.5	8
1fkb	0.58	7	9	0.64	7	10	0.62	7.5	9	0.59	6	5	0.63	6.5	9
1nso	0.72	7.5	2	0.76	8	2	0.72	8	2	0.72	7.5	2	0.72	8	2
1jli	0.58	6	10	0.57	6	10	0.58	7.5	10	0.64	7	9	0.59	6	10
1ooi	0.51	6	10	0.54	6.5	6	0.59	6.5	9	0.54	7	9	0.55	6.5	6
1agi	0.74	6.5	2	0.74	8	2	0.74	6.5	2	0.69	7.5	3	0.66	6	3
1bfg	0.60	8	10	0.59	7	10	0.56	7	10	0.55	6	10	0.59	6.5	10
1chn	0.65	6.5	2	0.70	6.5	2	0.70	6.5	2	0.70	8	2	0.73	6	2
1idr	0.57	8	6	0.59	7.5	8	0.60	7.5	8	0.62	8	6	0.61	7.5	8
1k40	1.00	7	2	1.00	6.5	2	1.00	6.5	2	0.98	6	2	1.00	6.5	2
1lys	0.48	6.5	10	0.54	6	5	0.55	6	9	0.57	7	10	0.50	6.5	9
1pdo	0.49	6	9	0.45	8	10	0.47	6	10	0.52	7.5	7	0.47	6	8
1lit	0.57	7.5	9	0.58	7.5	7	0.54	6	10	0.56	6.5	9	0.56	7.5	7
1cbs	0.59	7.5	2	0.55	6.5	5	0.59	7.5	5	0.67	8	2	0.65	7.5	4
1kxa	0.43	6.5	8	0.43	6	10	0.44	8	8	0.39	6.5	10	0.46	6.5	9
1emr	0.58	7	10	0.61	7.5	2	0.58	6	2	0.58	6.5	2	0.58	6.5	2
1czt	0.38	8	8	0.35	6	7	0.37	8	8	0.42	6	9	0.40	7.5	8
1il6	0.51	7	9	0.60	7	10	0.57	7	10	0.55	7	10	0.49	6	10
1sur	0.64	7.5	2	0.50	8	10	0.75	8	2	0.54	7.5	2	0.63	7	2
1acb	0.55	6	5	0.57	6.5	4	0.53	7.5	5	0.54	7.5	3	0.59	7	5
1cgi	0.53	7	5	0.52	8	5	0.51	8	5	0.56	7.5	6	0.51	6	6
2hvm	0.43	7.5	8	0.44	7.5	9	0.41	8	8	0.40	7.5	9	0.44	7.5	8
1gnd	0.83	7	2	0.80	6	2	0.83	7.5	2	0.81	7.5	2	0.80	7	2

Table B.3. Distribution of median Kappa coefficient over all community levels with a subset of 20 modes. We choose a $\boldsymbol{r_c}$ which maximizes the Kappa coefficient over all community levels. The median Kappa when considering all communities and choosing a different $\boldsymbol{r_c}$ for each protein is higher (0.49) than using a generalized $\boldsymbol{r_c}$ (7.5 **Å**) for all proteins (Kappa coefficient = 0.41)**.**

PDB ID	Median Kappa (r_c = 6Å)	Median Kappa (r_c = 6.5Å)	Median Kappa (r_c = 7Å)	Median Kappa (r_c = 7.5Å)	Median Kappa (r_c = 8Å)	Max Kappa	r_c for Max Kappa
1acb	0.36	0.31	0.40	0.40	0.35	0.40	7.5
1agi	0.52	0.55	0.51	0.50	0.52	0.55	6.5
1ark	0.50	0.52	0.54	0.60	0.48	0.60	7.5
1bfg	0.28	0.33	0.40	0.34	0.28	0.40	7
1bpi	0.41	0.44	0.41	0.47	0.49	0.49	8
1cbs	0.39	0.39	0.42	0.42	0.38	0.42	7.5
1cei	0.28	0.35	0.32	0.27	0.30	0.35	6.5
1cgi	0.38	0.40	0.42	0.43	0.38	0.43	7.5
1chn	0.54	0.47	0.44	0.45	0.41	0.54	6
1csp	0.45	0.40	0.43	0.42	0.39	0.45	6
1czt	0.31	0.26	0.22	0.25	0.27	0.31	6
1emr	0.35	0.43	0.41	0.39	0.33	0.43	6.5
1fas	0.32	0.29	0.37	0.35	0.29	0.37	7
1fkb	0.51	0.48	0.44	0.42	0.39	0.51	6
1fvq	0.42	0.52	0.41	0.42	0.49	0.52	6.5
1g6x	0.51	0.53	0.46	0.43	0.28	0.53	6.5
1gnd	0.47	0.46	0.48	0.45	0.48	0.48	8
1i6f	0.49	0.46	0.49	0.47	0.48	0.49	7
1idr	0.43	0.43	0.47	0.49	0.51	0.51	8
1il6	0.38	0.31	0.39	0.39	0.37	0.39	7
1j5d	0.46	0.39	0.51	0.43	0.46	0.51	7
1jli	0.37	0.41	0.33	0.36	0.32	0.41	6.5
1jw2	0.39	0.36	0.36	0.41	0.37	0.41	7.5
1k40	0.54	0.61	0.47	0.49	0.43	0.61	6.5
1kte	0.43	0.49	0.41	0.40	0.43	0.49	6.5
1kxa	0.38	0.36	0.25	0.29	0.38	0.38	8
1lit	0.32	0.36	0.35	0.35	0.36	0.36	8
1ls9	0.43	0.36	0.43	0.39	0.39	0.43	6
1lys	0.39	0.30	0.37	0.33	0.32	0.39	6
1nso	0.33	0.36	0.34	0.37	0.35	0.37	7.5
1ooi	0.36	0.52	0.40	0.40	0.35	0.52	6.5
1opc	0.43	0.42	0.35	0.43	0.42	0.43	6
1pdo	0.35	0.39	0.35	0.40	0.37	0.40	7.5
1pht	0.48	0.49	0.57	0.46	0.50	0.57	7
1sdf	0.50	0.52	0.54	0.54	0.52	0.54	7.5
1sro	0.58	0.48	0.49	0.41	0.48	0.58	6
1sur	0.41	0.38	0.42	0.36	0.40	0.42	7
1tba	0.13	0.51	0.58	0.55	0.61	0.61	8
1txa	0.37	0.34	0.57	0.48	0.49	0.57	7
1ubq	0.59	0.54	0.55	0.57	0.47	0.59	6
2gb1	0.43	0.48	0.47	0.48	0.55	0.55	8
2hvm	0.27	0.32	0.31	0.32	0.35	0.35	8
3ci2	0.49	0.41	0.42	0.35	0.39	0.49	6

Table B.4. Correlation for node betweenness. We identified the distance cutoff $\boldsymbol{r_c}$ which gives maximum correlation (ρ_{max})for node betweenness calculated from MD and GNM. For the subset of 5, 10, 20, 30 and 50 modes, we show the correlation for node betweenness for each protein and the corresponding value for $\boldsymbol{r_c}$.

	5 modes		10 modes		20 modes		30 modes		50 modes	
PDB ID	ρ_{max}	$\boldsymbol{r_c}$	ρ_{max}	$\boldsymbol{r_c}$	ρ_{max}	$\boldsymbol{r_c}$	ρ_{max}	$\boldsymbol{r_c}$	ρ_{max}	$\boldsymbol{r_c}$
2gb1	0.09	7.5	0.12	7	0.04	6	0.06	7.5	0.07	7.5
1bpi	0.28	6	0.38	7.5	0.51	8	0.42	7.5	0.46	8
1g6x	0.19	7.5	0.37	6	0.43	8	0.35	7.5	0.37	8
1ark	-0.28	8	0.40	7.5	0.56	7.5	0.54	8	0.38	6
1i6f	0.00	6	0.35	7.5	0.51	8	0.53	8	0.43	6
1fas	0.24	6	0.41	7.5	0.48	7.5	0.55	7.5	0.41	7.5
3ci2	-0.31	6	-0.09	7.5	-0.01	7	0.17	8	0.15	7
1csp	0.21	6	0.39	6.5	0.59	8	0.63	8	0.60	8
1sdf	0.40	6.5	0.40	6.5	0.50	7.5	0.55	7.5	0.51	7
1tba	0.23	6	0.42	6.5	0.37	7	0.40	7	0.36	7
1fvq	0.09	6	0.11	6.5	0.34	8	0.28	8	0.36	8
1jw2	0.18	6	0.24	6	0.08	8	0.20	8	0.14	8
1txa	0.29	6	0.12	6.5	0.46	6	0.42	7.5	0.32	8
1sro	0.32	6	0.40	6	0.50	7.5	0.53	7.5	0.56	7.5
1ubq	0.35	7	0.40	7	0.57	7	0.61	7.5	0.67	7
4icb	0.18	6	0.12	7	0.22	8	0.23	8	0.18	7.5
1pht	-0.06	6	0.07	8	0.24	7.5	0.29	8	0.31	8
1cei	0.14	7.5	0.19	6	0.12	6	0.07	7	0.15	6
1ls9	0.18	6.5	0.27	6.5	0.42	7.5	0.47	7.5	0.48	7
1j5d	0.30	6	0.35	6	0.40	8	0.39	7	0.29	7
1opc	0.11	7.5	0.10	8	0.31	7.5	0.29	7.5	0.28	7
1kte	-0.19	8	-0.22	8	0.00	7.5	0.07	7.5	0.10	7.5
1fkb	0.24	6.5	0.10	8	0.39	7.5	0.41	7.5	0.47	7.5
1nso	0.15	6	0.16	6.5	0.18	6	0.20	8	0.24	6.5
1jli	0.22	7	0.28	7.5	0.40	7.5	0.43	7	0.49	7.5
1ooi	0.11	6	0.12	7	0.40	8	0.36	8	0.44	8
1agi	0.35	6	0.29	7	0.39	7.5	0.34	7.5	0.30	7.5
1bfg	0.01	8	0.21	6.5	0.38	7	0.43	7	0.46	6
1chn	0.14	6	0.23	6	0.35	8	0.42	8	0.52	8
1idr	0.08	7.5	0.18	8	0.27	7	0.15	7	0.18	7.5
1k40	0.03	8	-0.02	8	0.00	6	0.07	8	0.24	7
1lys	0.16	6	0.22	6.5	0.36	7	0.37	7	0.49	7
1pdo	0.07	6.5	0.11	6.5	0.30	6.5	0.44	8	0.51	8
1lit	0.14	6	0.34	7	0.50	6.5	0.57	7	0.60	7
1cbs	0.13	6	0.19	6	0.26	8	0.32	8	0.32	6.5
1kxa	0.12	6	0.15	6	0.28	7.5	0.35	6.5	0.40	8
1emr	0.24	6.5	0.29	6.5	0.39	6.5	0.48	6.5	0.49	6.5
1czt	0.13	6.5	0.28	7.5	0.53	7.5	0.53	7.5	0.51	7.5
1il6	0.19	7	0.22	8	0.33	8	0.40	7	0.42	6.5
1sur	0.10	8	0.18	7.5	0.25	7.5	0.23	7.5	0.27	8
1acb	0.03	6	0.10	6.5	0.24	6	0.35	8	0.48	8
1cgi	-0.05	6.5	0.11	8	0.29	8	0.36	7.5	0.48	7.5
2hvm	0.08	6.5	0.09	6	0.11	8	0.14	8	0.22	7.5
1gnd	0.15	6	0.26	6.5	0.29	6.5	0.33	6	0.37	8

Table B.5. Correlation for node closeness. The table shows the correlation for node closeness between MD and GNM. Similar to the node betweenness, we consider, for each protein, the distance cutoff (r_c) which gives maximum correlation (ρ_max).

	5 modes		10 modes		20 modes		30 modes		50 modes	
PDB ID	ρ_{max}	r_c	ρ_{max}	r_c	ρ_{max}	r_c	ρ_{max}	r_c	ρ_{max}	r_c
2gb1	0.04	7.5	0.31	7.5	0.36	7.5	0.36	7.5	0.23	7.5
1bpi	0.38	6.5	0.80	7	0.80	7	0.84	8	0.86	8
1g6x	0.18	7.5	0.80	7.5	0.83	6	0.79	6	0.84	6.5
1ark	0.20	7.5	0.57	7.5	0.65	7.5	0.72	7	0.76	8
1i6f	0.51	6	0.72	7	0.75	7.5	0.72	8	0.66	7.5
1fas	0.65	6.5	0.76	7	0.84	7	0.82	7	0.75	7
3ci2	0.07	8	0.44	7.5	0.34	7.5	0.40	7.5	0.44	7.5
1csp	0.57	6	0.68	6.5	0.87	6	0.86	8	0.85	6
1sdf	0.80	6	0.92	7.5	0.86	7.5	0.89	7.5	0.91	6.5
1tba	0.53	6.5	0.48	6.5	0.52	6.5	0.50	8	0.65	8
1fvq	0.66	6	0.62	6	0.59	8	0.64	6	0.78	8
1jw2	0.74	6.5	0.82	6	0.82	6	0.79	6	0.80	6
1txa	0.32	6	0.56	7.5	0.54	7.5	0.57	7.5	0.44	7.5
1sro	0.65	6	0.77	6	0.84	7.5	0.90	7.5	0.88	7.5
1ubq	-0.38	6.5	0.29	7	0.15	7	0.47	7	0.16	7.5
4icb	0.66	8	0.82	7.5	0.72	8	0.75	8	0.81	7.5
1pht	0.30	7	0.36	7.5	0.71	7.5	0.67	7.5	0.70	7.5
1cei	0.33	6.5	0.49	6.5	0.53	6.5	0.25	6.5	0.20	6.5
1ls9	0.18	6	0.55	8	0.66	7.5	0.64	7.5	0.62	7.5
1j5d	0.42	6	0.61	6	0.51	8	0.51	8	0.46	7.5
1opc	0.42	7	0.61	6.5	0.69	8	0.66	7.5	0.64	7.5
1kte	0.15	6	0.21	8	0.44	7.5	0.42	7.5	0.42	8
1fkb	0.38	6	0.61	7.5	0.80	6.5	0.79	6.5	0.75	6
1nso	0.59	6	0.58	8	0.67	7.5	0.71	7.5	0.74	7.5
1jli	0.49	6.5	0.64	7	0.77	7	0.81	7	0.81	7.5
1ooi	0.12	6	0.28	6	0.51	8	0.51	7.5	0.43	7.5
1agi	0.76	6.5	0.68	8	0.59	7.5	0.61	8	0.62	7.5
1bfg	0.20	6.5	0.34	7	0.44	7	0.47	7.5	0.51	7
1chn	0.37	6	0.72	6	0.74	6	0.67	7.5	0.66	8
1idr	0.43	6	0.50	8	0.58	7.5	0.47	7.5	0.50	8
1k40	-0.23	7	-0.12	8	0.27	8	0.52	8	0.78	7.5
1lys	0.67	6.5	0.76	6.5	0.75	8	0.82	8	0.80	6.5
1pdo	0.36	7	0.55	7	0.74	7	0.72	7	0.53	7.5
1lit	0.32	6	0.37	6	0.73	6.5	0.78	6	0.81	6
1cbs	0.37	6.5	0.57	6	0.74	8	0.72	8	0.71	8
1kxa	0.28	8	0.28	6.5	0.42	6.5	0.57	6.5	0.61	6.5
1emr	0.05	6	0.35	7	0.66	8	0.66	8	0.68	8
1czt	0.45	7	0.59	7.5	0.68	6	0.62	7.5	0.59	6
1il6	0.55	7	0.32	7	0.77	7	0.78	7	0.72	7
1sur	0.54	6.5	0.25	7.5	0.54	7.5	0.62	6	0.56	6
1acb	0.18	7	0.12	6	0.60	8	0.60	6.5	0.68	6.5
1cgi	0.33	6.5	0.31	7	0.54	6	0.61	6	0.71	6
2hvm	0.40	7.5	0.47	6	0.59	6	0.62	7.5	0.59	6
1gnd	0.56	7	0.60	7	0.59	8	0.64	7.5	0.69	8

Table B.6. Distribution of root-mean square inner product (RMSIP) for the dataset. The principal eigenvectors are obtained with singular-value decomposition of the cross-correlation matrices from MD and GNM. They capture the major directions of variations from the matrix. We see a considerably good overlap (median RMSIP 0.82 over all subsets of modes) between the principal eigenvectors from MD and GNM which suggests a close agreement between the two.

	5 modes		10 modes		20 modes		30 modes		50 modes	
PDB ID	Max RMSIP	r_c	Max RMSIP	r_c	Max RMSIP	r_c	Max RMSIP	r_c	Max RMSIP	r_c
2gb1	0.82	7	0.86	7.5	0.83	6	0.82	6	0.95	6
1bpi	0.78	8	0.86	6	0.82	6	0.83	6	0.94	6
1g6x	0.78	7.5	0.83	6.5	0.82	6	0.83	6	0.94	7.5
1ark	0.82	7	0.86	7.5	0.84	6	0.85	6	0.92	7
1i6f	0.76	6	0.82	7.5	0.82	6	0.83	6	0.93	6.5
1fas	0.80	7.5	0.84	7.5	0.85	6	0.83	6	0.92	7.5
3ci2	0.77	8	0.82	7.5	0.82	7	0.80	7.5	0.90	6
1csp	0.77	7.5	0.81	6	0.84	6.5	0.84	6	0.90	6
1sdf	0.83	8	0.86	6.5	0.81	6	0.82	6	0.90	6
1tba	0.82	8	0.85	7	0.82	6.5	0.79	7	0.89	8
1fvq	0.70	7.5	0.82	7.5	0.80	6	0.80	6	0.86	6
1jw2	0.79	6	0.81	7.5	0.81	6	0.80	6	0.85	6
1txa	0.67	7.5	0.79	7.5	0.81	7.5	0.81	6	0.87	6
1sro	0.77	8	0.83	8	0.83	6	0.83	6	0.88	6
1ubq	0.78	8	0.88	8	0.87	6	0.83	6	0.88	6
4icb	0.71	7	0.83	7.5	0.82	6	0.80	6	0.84	6
1pht	0.81	7.5	0.86	7.5	0.84	7.5	0.82	6	0.84	6
1cei	0.71	6	0.81	7	0.83	6	0.81	6	0.82	6
1ls9	0.69	7	0.84	7.5	0.85	6	0.81	6	0.83	7
1j5d	0.68	7	0.76	7.5	0.81	7	0.81	6	0.83	6
1opc	0.72	7	0.84	8	0.86	7	0.83	7	0.82	7
1kte	0.70	7.5	0.81	6.5	0.83	7.5	0.82	6.5	0.82	6
1fkb	0.84	8	0.88	7.5	0.88	7.5	0.86	6	0.83	6
1nso	0.64	7	0.74	8	0.78	6	0.81	6	0.81	6
1jli	0.66	6.5	0.80	7	0.84	7	0.83	6	0.82	6
1ooi	0.72	7.5	0.82	7.5	0.87	7.5	0.84	6.5	0.81	6
1agi	0.71	6	0.80	7.5	0.82	6	0.82	6	0.81	6
1bfg	0.61	7.5	0.78	7.5	0.84	7	0.82	7	0.82	6
1chn	0.75	8	0.88	7	0.86	7.5	0.85	7	0.82	6
1idr	0.73	6.5	0.83	7.5	0.84	7.5	0.81	7	0.79	6
1k40	0.79	8	0.85	8	0.88	6	0.82	6	0.80	7
1lys	0.74	6	0.83	7.5	0.86	6.5	0.86	6.5	0.83	6
1pdo	0.78	6	0.82	7.5	0.83	7.5	0.84	6.5	0.81	6
1lit	0.73	8	0.82	6	0.87	7.5	0.86	6.5	0.84	6
1cbs	0.70	7.5	0.79	7.5	0.86	8	0.84	6	0.82	6.5
1kxa	0.67	7.5	0.81	8	0.82	7.5	0.81	7.5	0.81	6
1emr	0.63	8	0.78	7.5	0.83	7	0.84	6	0.81	6
1czt	0.68	8	0.79	8	0.82	7	0.83	7.5	0.83	6
1il6	0.66	8	0.76	8	0.85	7.5	0.85	7	0.81	6
1sur	0.78	7.5	0.78	8	0.84	7.5	0.84	7	0.81	7.5
1acb	0.61	6.5	0.78	7	0.83	7.5	0.86	7.5	0.85	6
1cgi	0.59	8	0.78	8	0.81	7	0.84	7.5	0.85	6
2hvm	0.72	7	0.77	8	0.83	7.5	0.85	8	0.85	7
1gnd	0.73	7.5	0.77	8	0.81	8	0.83	8	0.84	7.5

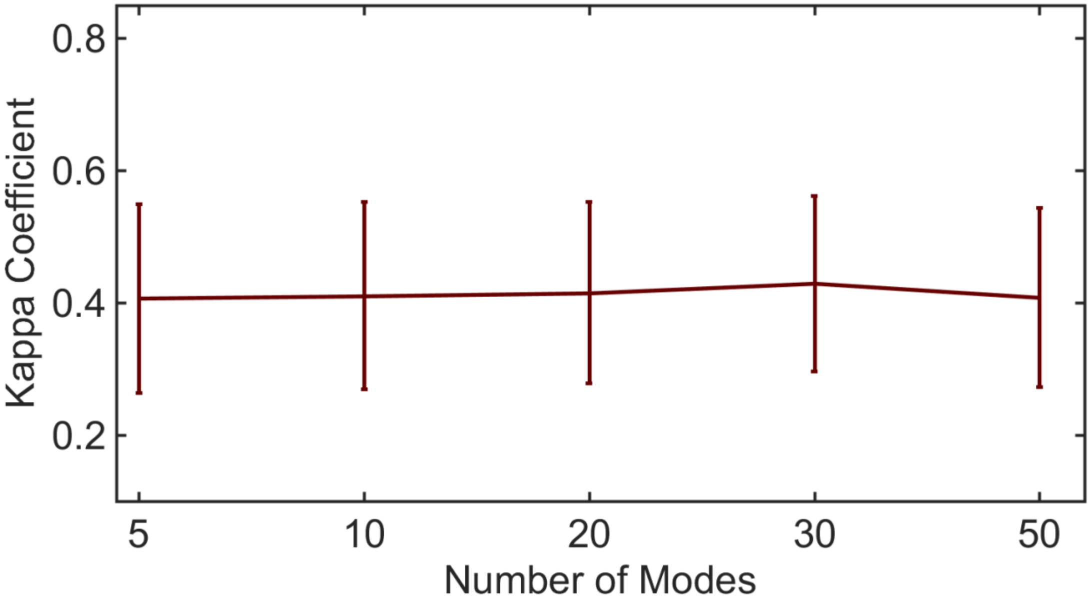

Figure B.1. Distribution of Kappa coefficient for all community levels when using a generalized distance cutoff ($\boldsymbol{r_c}$ = 7.5). We verified the variation of Kappa coefficient upon choosing a generalized $\boldsymbol{r_c}$ = 7.5 for all proteins. The figure shows the median Kappa over all protein for all community levels for each subset of modes. It is interesting to see the median remaining almost the same across all modes, expect for 50 modes, where it slightly decreases. The error bars indicate standard error for the Kappa coefficient for a given subset of modes.

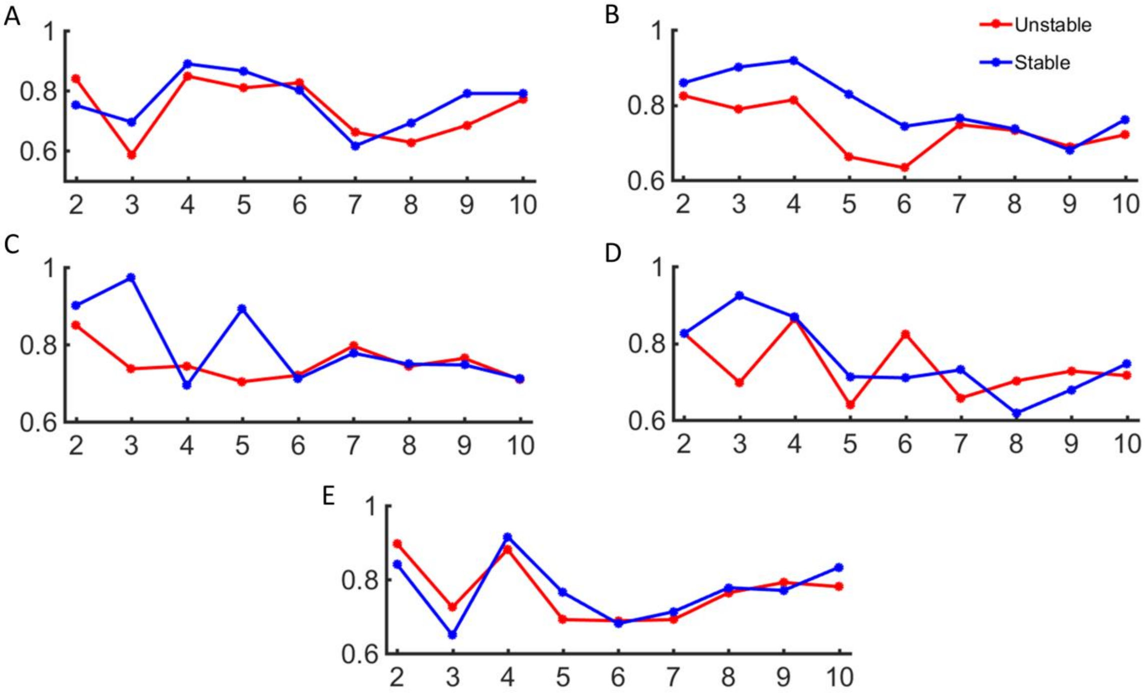

Figure B.2. Agreement of communities from unstable and stable mutant forms of T4 Lysozyme with the wild-type using (A) subset of 5 modes, (B) subset of 10 modes, (C) subset of 20 modes, (D) subset of 30 modes , (E) subset of 50 modes. For each plot the abscissa is the number of communities and the ordinate is the Kappa coefficient. We observe that the stable forms more closely resemble the community structure of the wild-type protein (higher Kappa coefficient) than the unstable forms. The agreement/disagreement of the stable/unstable forms is more distinct for community levels 2-6 and also using when using a subset of 10 and 20 modes.

APPENDIX C. SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Table C.1. Performance of AR-Pred against other active site prediction methods. The percentage of proteins for which AR-Pred predicts the same or more number of true positive active site residues relative to the other methods is tabulated. The calculations are performed at each percent threshold that considers the top 10, 20, 30, 40 and 50 percent of the predictions.

Method / Threshold Percent	Concavity	AADS	POOL	FOD
10	57.89	89.47	61.11	84.21
20	36.84	73.68	72.22	89.47
30	57.89	78.95	77.78	89.47
40	63.16	78.95	61.11	63.16
50	68.42	84.21	66.67	78.95
Median	57.89	78.95	66.67	84.21

Table C.2. Performance of AR-Pred against other allosteric site prediction methods. The percentage of proteins for which AR-Pred predicts the same or more number of true positive allosteric site residues relative to the other methods is tabulated. The calculations are performed at each percent threshold that considers the top 10, 20, 30, 40 and 50 percent of the predictions.

Method / Threshold Percent	AlloPred	AlloSitePro	SPACER
10	80.00	93.33	86.67
20	80.00	93.33	86.67
30	73.33	93.33	86.67
40	86.67	93.33	80.00
50	86.67	93.33	66.67
Median	80.00	93.33	86.67

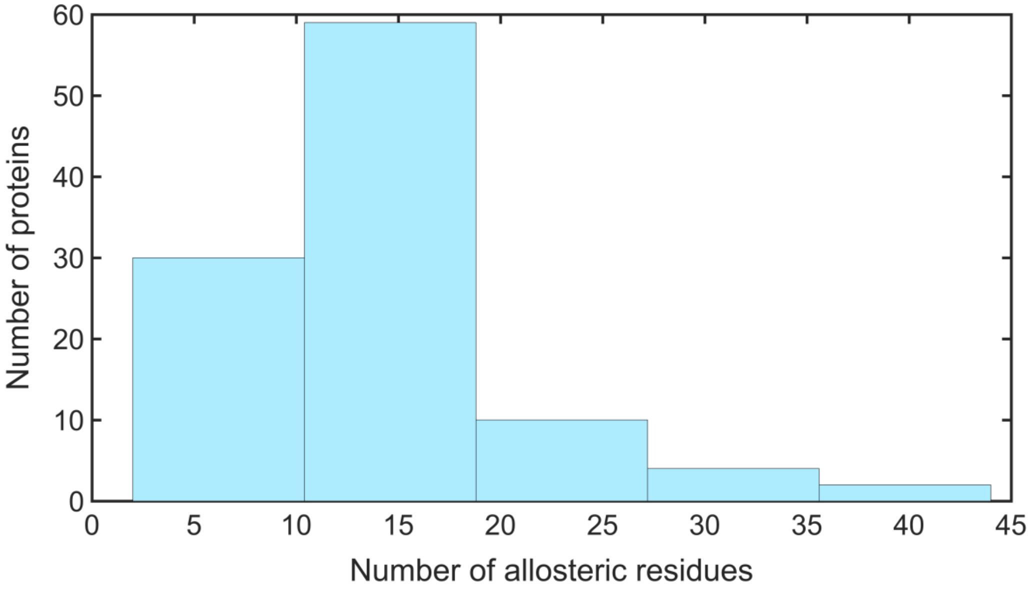

Figure C.1. Distribution of allosteric residues. The number of proteins having a certain number of allosteric residues is shown. Each bin considers a certain range of allosteric residues and the number of proteins that have the same range of allosteric residues. It is seen that a majority of proteins in the dataset have 10-18 allosteric residues.

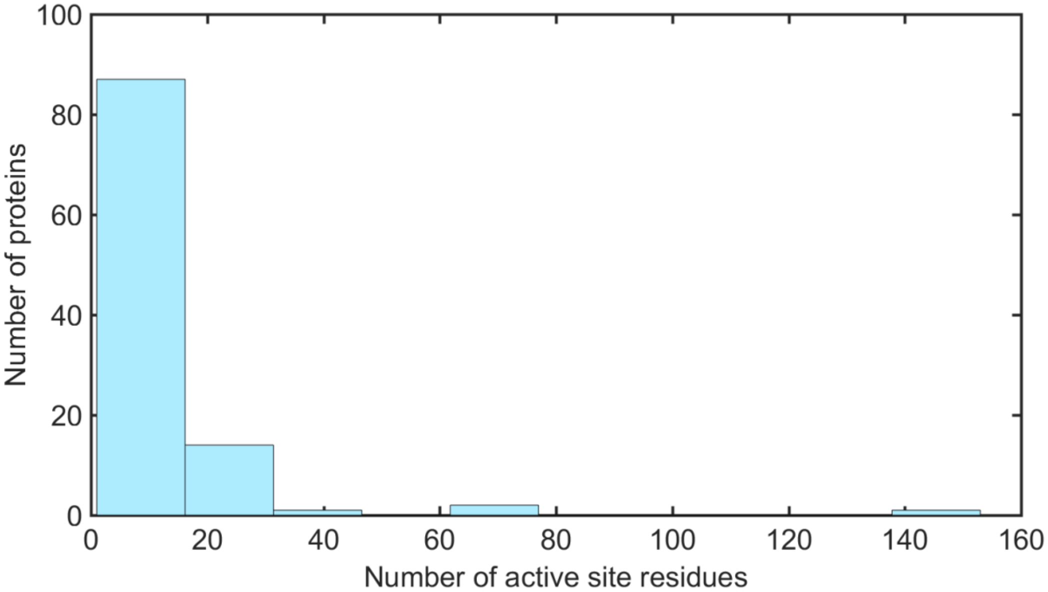

Figure C.2. Distribution of active site residues. The number proteins having a given number of active site residues is shown. Only two groups are prominent unlike the allosteric residue distribution which has more bins.

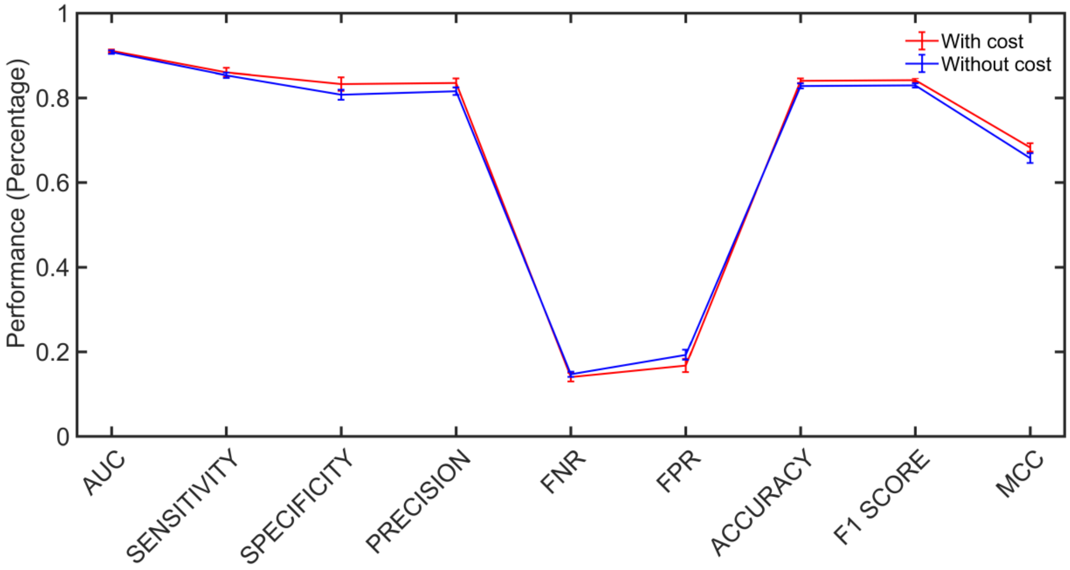

Figure C.3. Effect of cost on active site model performance. Median performance of models for active site prediction with and without including misclassification costs (cost for false positives and false negatives).

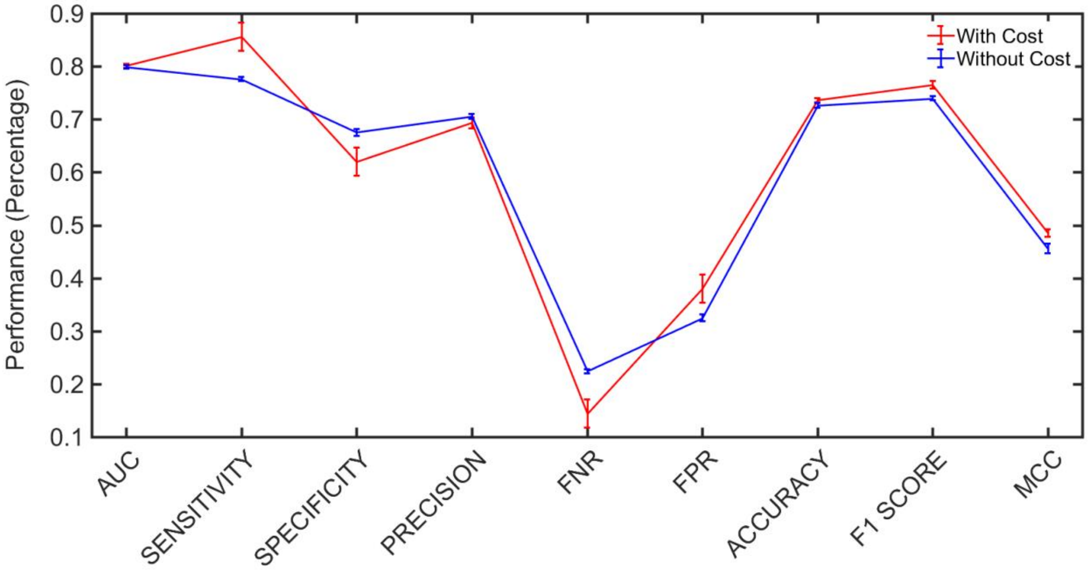

Figure C.4. Effect of cost on allosteric site model performance. Median performance of models for allosteric site prediction with and without including misclassification costs (cost for false positives and false negatives).

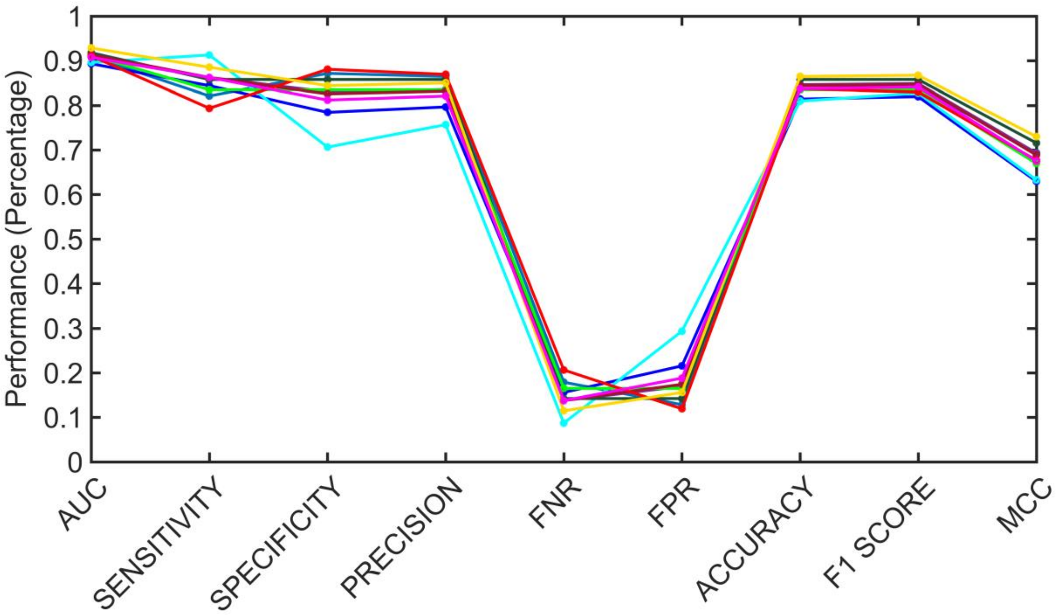

Figure C.5. Performance of individual models for active site prediction. Different metrics show the performance of AR-Pred's active site prediction models on their respective validation datasets.

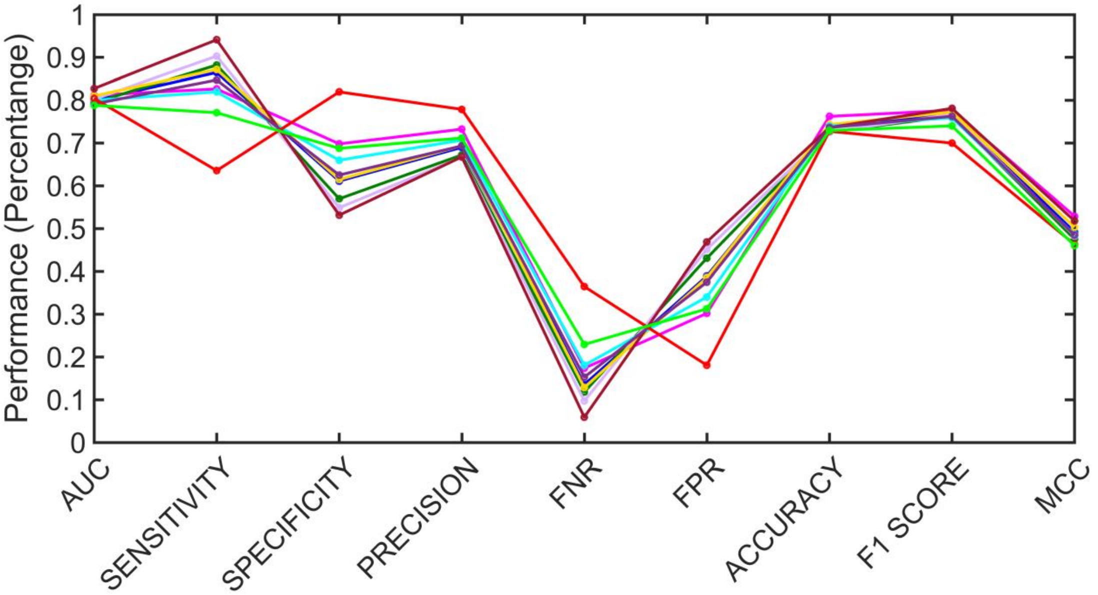

Figure C.6. Performance of individual models for allosteric site prediction. Metrics for AR-Pred's allostery prediction models on their respective validation datasets are shown. The performance is considerably less than that of the active site prediction models and so are the inter-model variabilities.

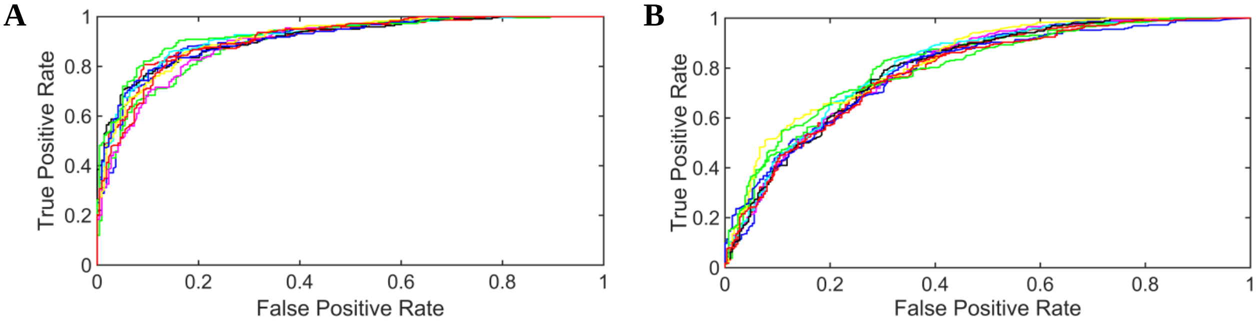

Figure C.7. Receiver Operating Characteristics (ROC) Area Under Curve (AUC). AUCs for active site models (A) vs allosteric models (B).

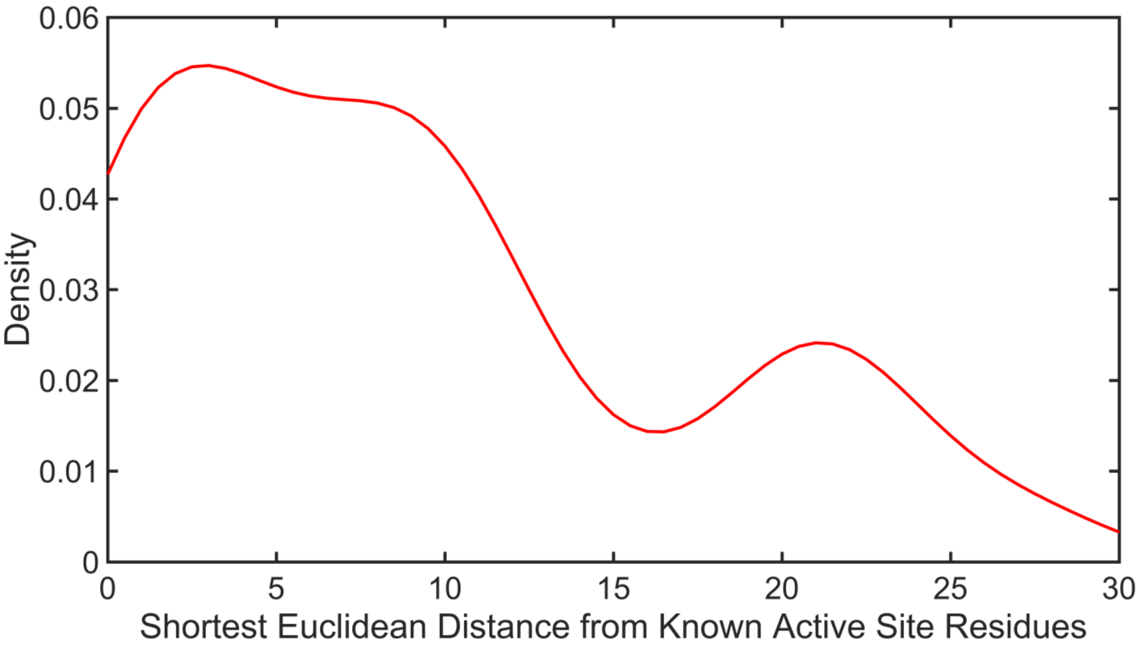

Figure C.8. Distribution of shortest distances for predicted active sites. The shortest distances of the top 15 predicted active sites from the reported sites is fit into a distribution and plotted. The shortest distance is defined as the minimum distance between the heavy atoms of the predicted active sites and any of the reported active site residues. All distances are reported in Angstrom.

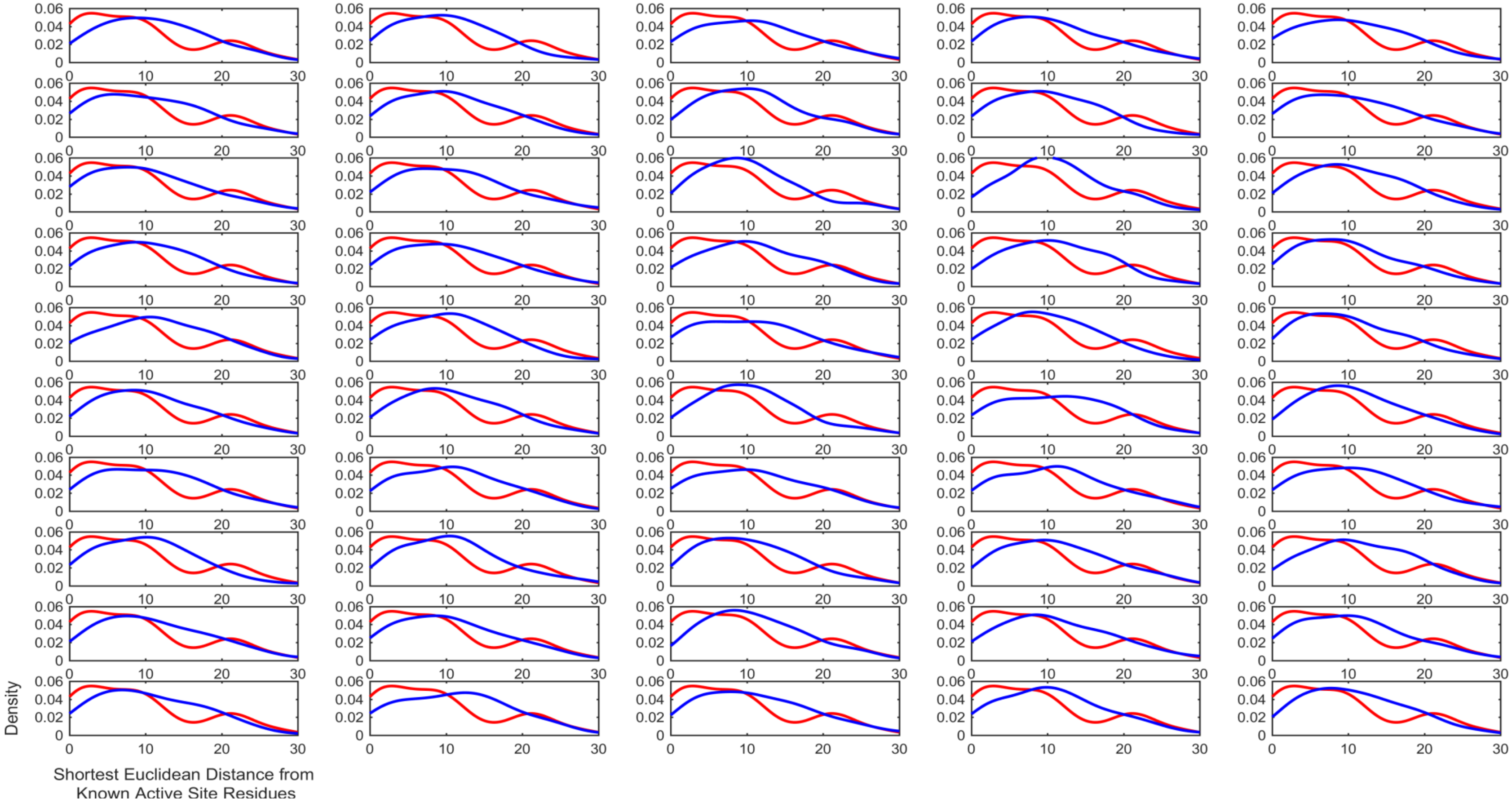

Figure C.9. Comparisons of shortest distance distributions for predicted active site and random residues. The distributions of the shortest distances from the reported active sites for the top 15 predicted residues (red) from all the proteins are compared with the distributions of 15 randomly picked residues (blue). The comparisons are made for 50 iterations. The distances are measured in Angstrom. The distributions are sharper at shorter distances for the predicted residues than the randomly selected residues.

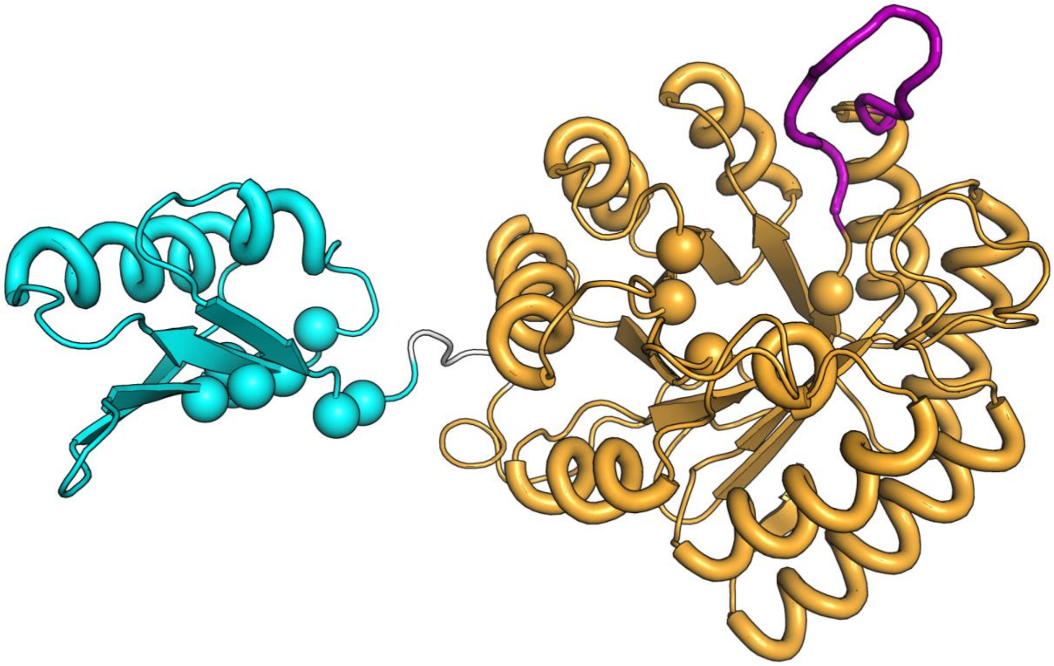

Figure C.10. *Thermotoga maritima* (PDB 3PG9) DAH7PS regulatory and catalytic domains. The regulatory (cyan) and catalytic (orange) domains for the protein are shown. The reported allosteric and active site residues are shown as cyan and orange spheres, respectively. The β2-α2 loop is colored in in violet.

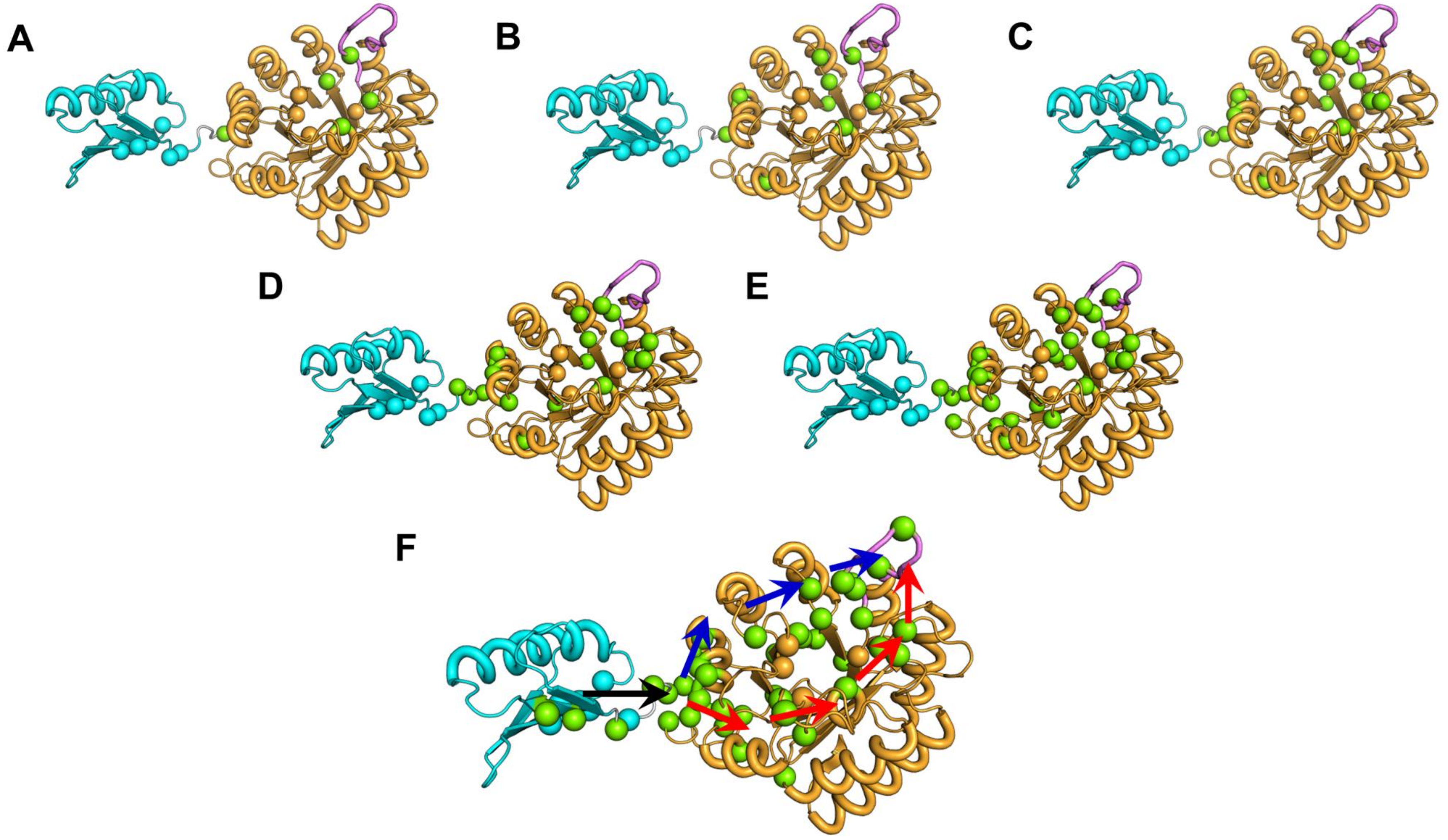

Figure C.11. Alloteric residue predictions for DAH7S. The predictions for allosteric residues made on DAH7S considering top 5 (A), top 10 (B), top 15 (C), top 20 (D), top 30 (E), and top 40 (F) predicted allosteric residues are shown. Fig F shows the two possibly allosteric routes that may be involved in transmitting the signal from the regulatory the active site domain, regulating the conformational transition of the β2-α2 loop.

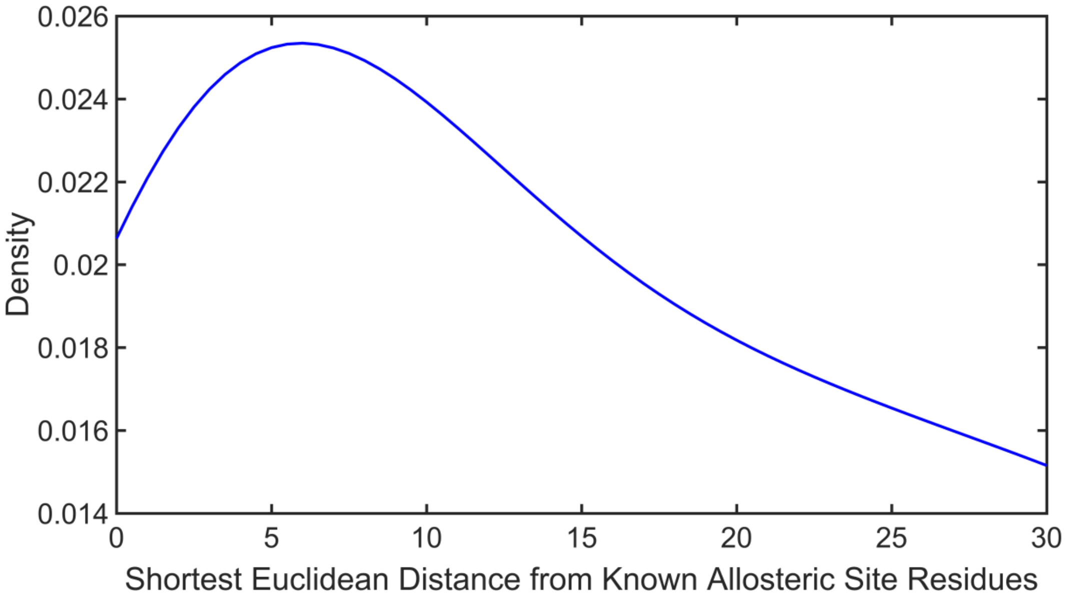

Figure C.12. Distribution of shortest distances for predicted allosteric sites. The shortest distances of the top 15 predicted allosteric sites from the reported sites is fit into a distribution and plotted. The shortest distance is defined in the same way as in Fig. C8. All distances are reported in Angstrom.

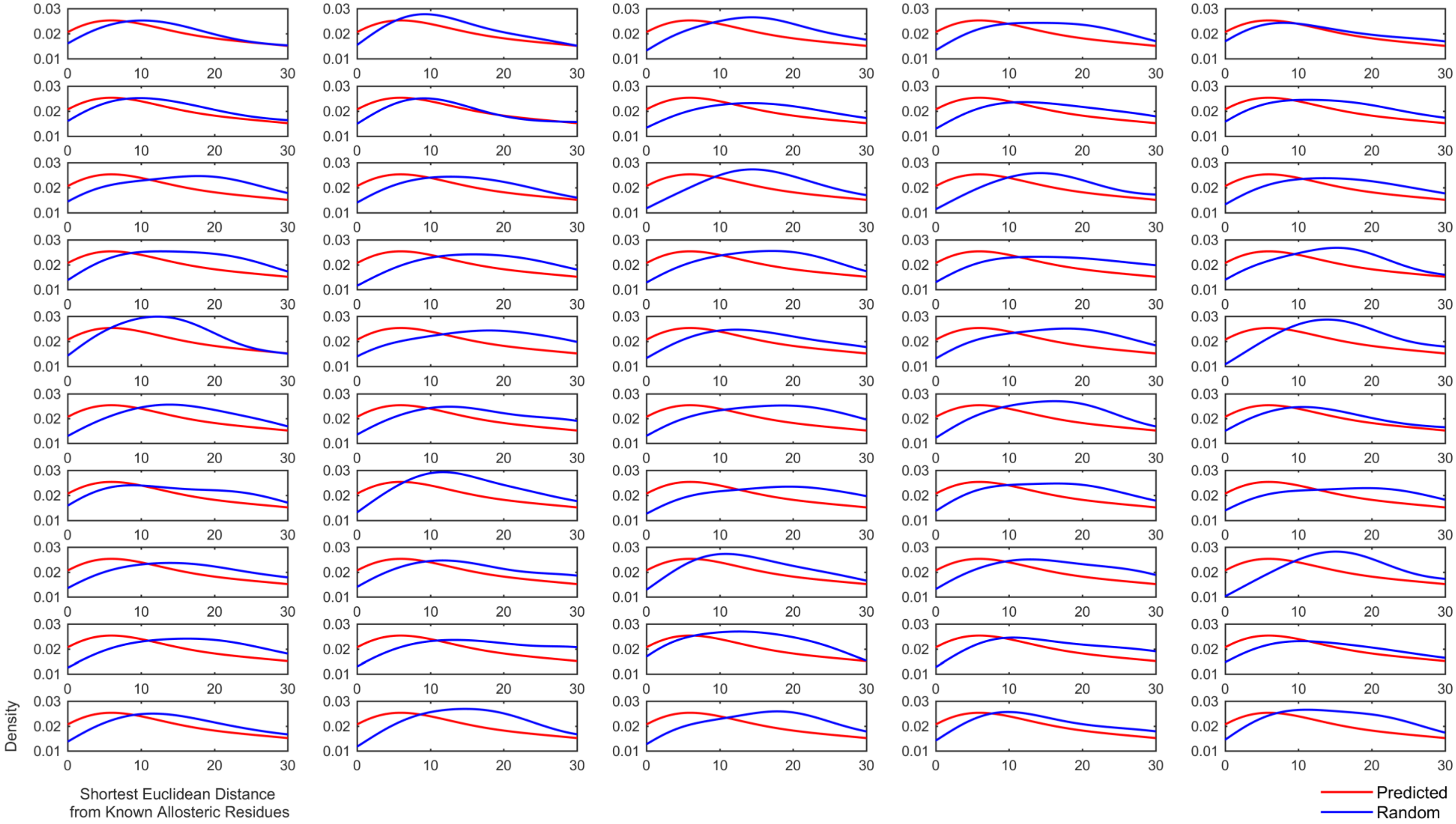

Figure C.13. Comparisons of shortest distance distributions for predicted allosteric site and random residues. The distributions of the shortest distances from the reported allosteric sites for the top 15 predicted residues (red) from all the proteins are compared with the distributions of 15 randomly picked residues (blue). The comparisons are made for 50 iterations. The distances are measured in Angstrom. The distributions are sharper at shorter distances for the predicted residues than for the randomly selected residues.

APPENDIX D. SUPPLEMENTARY MATERIAL FOR CHAPTER 5

Table D.1. List of proteins in the experimental ensemble dataset.

Set #	Protein Name	#Residues	#Structures	Organism	Representative Structure
1	Sarcoplasmic/endoplasmic reticulum calcium	995	63	*Oryctolagus*	3NAL_A
2	Peptidyl-prolyl cis-trans isomerase A	159	136	*Homo sapiens*	3ODL_A
3	Human Lysozyme C	131	218	*Homo sapiens*	1B5U_A
4	B. anthracis Dihydrofolate reductase (DHFR)	162	76	*Bacillus anthracis*	3FL8_F
5	Cytochrome c peroxidase, mitochondrial	292	165	*Saccharomyces*	2AQD_A
6	HLA class II histocompatibility antigen, D-R alpha	172	108	*Homo sapiens*	1T5W_A
7	Thaumatin I	202	80	*Thaumatococcus*	3AOK_A
8	FK506-binding protein	108	59	*Homo sapiens*	1D6O_A
9	Human serum albumin (HSA)	555	99	*Homo sapiens*	2BXB_B
10	Phi6 RNA-directed RNA polymerase	665	55	*Pseudomonas*	1UVJ_A
11	Squalene synthase	332	61	*Homo sapiens*	3WCF_F
12	Camphor 5-monooxygenase	402	134	*Pseudomonas*	1UYU_B
13	Azurin	129	202	*Pseudomonas*	1E5Y_C
14	Proteinase K	280	61	*Engyodontium*	3DVR_X
15	Beta-lactamase	359	143	*Escherichia coli*	4KZ5_B
16	Hepatitis C RNA-directed RNA polymerase	548	162	*Hepatitis C virus*	2XHU_B
17	Tankyrase-2	186	64	*Homo sapiens*	4PNN_B
18	Heparin-binding growth factor 1	122	61	*Homo sapiens*	2HW9_A
19	Casein kinase II subunit alpha	326	78	*Homo sapiens*	3NGA_A
20	Thioredoxin 1	104	80	*Escherichia coli*	2H73_A
21	H-2 class I histocompatibility antigen, alpha chain	272	89	*Mus musculus*	1S7U_A
22	T4 lysozyme	163	183	*Enterobacteria*	1G0J_A
23	GTPase HRas	165	100	*Homo sapiens*	4L9W_A
24	Heparin-binding growth factor 1	121	130	*Homo sapiens*	1JQZ_A
25	Aldose reductase	309	120	*Homo sapiens*	2IKH_A
26	Phosphopentomutase	390	60	*Bacillus cereus*	3M8Z_B
27	MHC class I antigen	274	64	*Homo sapiens*	1ZSD_A
28	Carboxypeptidase B	304	58	*Sus scrofa*	2PJ5_B
29	HLA class I histocompatibility antigen, A-2 alpha	276	256	*Homo sapiens*	3KLA_A
30	Chemotaxis protein CheY	115	109	*Escherichia coli*	3F7N_B
31	DNA polymerase beta	326	154	*Homo sapiens*	8ICZ_A
32	Human Dihydrofolate reductase	183	74	*Homo sapiens*	1BOZ_A
33	Glucosylceramidase	488	64	*Homo sapiens*	1OGS_B
34	D-alanyl-D-alanine Carboxypeptidase	461	72	*Actinomadura sp.*	4BEN_C
35	WD repeat-containing protein 5	294	80	*Homo sapiens*	2H6Q_B
36	LeuT Transporter	503	45	*Aquifex aeolicus*	3F3D_A

37	Cathepsin S	217	58	*Homo sapiens*	2FRA_B
38	Thermolysin	317	122	*Bacillus*	1KEI_A
39	Polymerase	458	55	*Human poliovirus*	3OL6_A
40	Hen egg white lysozyme	130	586	*Gallus gallus*	194L_A
41	Beta-2-microglobulin	100	242	*Mus musculus*	1RJY_E
42	Phospholipase A2	122	80	*Daboia russellii*	1SV9_A
43	Beta-lactamase TEM	260	59	*Escherichia coli*	1NYY_A
44	Guanyl-specific ribonuclease T1	105	89	*Aspergillus oryzae*	1BU4_A
45	E-coli Dihydrofolate reductase	160	80	*Escherichia coli*	1DHI_B
46	Insulin-degrading enzyme	942	61	*Homo sapiens*	3OFI_A
47	Cationic trypsin	224	421	*Bos taurus*	1S0Q_A
48	Elastase 1	241	116	*Sus scrofa*	2BD3_A
49	Endothiapepsin	331	52	*Endothia*	3PI0_A
50	Macrophage metalloelastase	153	83	*Homo sapiens*	3F17_A

Table D.2. List of proteins with MD trajectory data

Set #	**Protein Name**	**Organism**	**Representative PDB with MD data**	**Simulation Details**
1	Beta-2-microglobulin	*Mus musculus*	1HSA	Amber 8, 20
2	Camphor 5-monooxygenase	*Pseudomonas putida*	1AKD	Amber 8, 10.5
3	H-2 class I histocompatibility antigen, D-B	*Mus musculus*	1HSA	Amber 8, 20
4	Thermolysin	*Bacillus*	1FJ3	Amber 8 v1,
5	Cytochrome c peroxidase, mitochondrial	*Saccharomyces*	1JDR	Amber 8, 10
6	HLA class I histocompatibility antigen, A-	*Homo sapiens*	2BVO	Amber 8, 20
7	MHC class I antigen	*Homo sapiens*	2AXG	Amber 8, 10
8	Elastase 1	*Sus scorfa*	1ESA	Amber 9, 80
9	Thaumatin I	*Thaumatococcus*	1THV	Amber 9, 80
10	HLA class II histocompatibility antigen,	*Homo sapiens*	1DLH	Amber 8, 10
11	Peptidyl-prolyl cis-trans isomerase A	*Homo sapiens*	2CPL	Amber 8, 80.5
12	Heparin-binding growth factor 1	*Homo sapiens*	1FMM	Amber 8, 10
13	Hen Egg White Lysozyme C	*Gus gallus*	1DPX	Amber 8, 20
14	Heparin-binding growth factor 1	*Gallus gallus*	1FMM	Amber 8, 10
15	Human Lysozyme C	*Homo sapiens*	1JSF	Amber 8, 10
16	Phospholipase A2	*Daboia russellii*	1BBC	Amber 8, 10
17	FK506-binding protein	*Homo sapiens*	1FKB	Amber 8v1,

Table D.3. Comparison of dcANMs based on experimental and MD datasets.

Test / Train	O_1^{max}	O_2^{max}	O_3^{max}	CO_1^{20}	CO_2^{20}	CO_3^{20}	$RMSIP_3^{20}$	$RMSIP_6^{20}$	$RMSIP_{10}^{20}$	$RMSIP_{20}^{20}$
Exp / MD	0.32	0.33	0.37	0.56	0.57	0.64	0.60	0.57	0.54	0.49
MD / Exp	0.34	0.32	0.30	0.60	0.56	0.55	0.58	0.58	0.56	0.50

Values for each metric are averaged over the 17 proteins.

The 'Train' set refers to the ensemble from which the internal distance changes were extracted to train the dcANM. The dcANM is built on the representative structure in each dataset. The 'Test' set refers to the ensemble from which the PCs were extracted. The modes from the dcANM generated using the 'Train' set are tested against the PCs from the 'Test' set using each of the 10 different metrics.

Table D.4. Comparison of performance metrics between short and long MD simulations.

Representative PDB	Simulation Type	Simulation Time	O_1^{max}	O_2^{max}	O_3^{max}	CO_1^{20}	CO_2^{20}	CO_3^{20}	$RMSIP$	$RMSIP$	$RMSIP$	$RMSIP$
1HSA	Short	20	0.28	0.24	0.30	0.53	0.53	0.66	0.58	0.63	0.62	0.57
1AKD	Short	10.5	0.38	0.31	0.25	0.73	0.61	0.62	0.66	0.61	0.58	0.53
1HSA	Short	20	0.55	0.46	0.43	0.90	0.81	0.78	0.83	0.76	0.68	0.57
1FJ3	Short	10	0.55	0.32	0.43	0.78	0.58	0.82	0.73	0.67	0.59	0.51
1JDR	Short	10	0.20	0.12	0.32	0.43	0.27	0.56	0.44	0.46	0.45	0.44
2BVO	Short	20	0.76	0.62	0.43	0.89	0.91	0.71	0.84	0.78	0.70	0.62
2AXG	Short	10	0.67	0.29	0.44	0.90	0.54	0.84	0.77	0.68	0.65	0.57
1ESA	Long	80	0.22	0.36	0.23	0.48	0.63	0.65	0.59	0.62	0.62	0.55
1THV	Long	80	0.17	0.41	0.53	0.38	0.66	0.82	0.64	0.60	0.59	0.52
1DLH	Short	10	0.34	0.52	0.36	0.71	0.89	0.75	0.78	0.70	0.67	0.57
2CPL	Long	80.5	0.13	0.21	0.21	0.31	0.42	0.54	0.43	0.52	0.58	0.56
1FMM	Short	10	0.35	0.30	0.34	0.58	0.58	0.75	0.65	0.64	0.60	0.54
1DPX	Short	20	0.37	0.33	0.31	0.56	0.76	0.62	0.65	0.70	0.70	0.67
1FMM	Short	10	0.24	0.39	0.27	0.50	0.61	0.53	0.55	0.54	0.53	0.49
1JSF	Short	10	0.47	0.56	0.39	0.72	0.79	0.74	0.75	0.69	0.69	0.62
1BBC	Short	10	0.38	0.31	0.52	0.78	0.64	0.77	0.73	0.67	0.60	0.53
1FKB	Long	100	0.32	0.54	0.46	0.58	0.84	0.73	0.73	0.74	0.70	0.62
P-value (Wilcoxon Test)*			1.00	0.39	0.56	0.99	0.48	0.69	0.95	0.85	0.61	0.56
P-value (Welch's t-test)#			1.00	0.44	0.55	0.99	0.57	0.62	0.88	0.74	0.46	0.39

*Wilcoxon rank sum test with $H_o: \mu_S = \mu_L$ and with $H_A: \mu_S < \mu_L$
#Welch's t- test with $H_o: \mu_S = \mu_L$ and with $H_A: \mu_S < \mu_L$
(S = short simulations; L = long simulations)